Distributed MST Computation in the Sleeping Model: Awake-Optimal Algorithms and Lower Bounds

John Augustine∗ William K. Moses Jr.† Gopal Pandurangan ‡

Abstract

We study the distributed minimum spanning tree (MST) problem, a fundamental problem in distributed computing. It is well-known that distributed MST can be solved in $\tilde{O}(D + \sqrt{n})$ rounds in the standard CONGEST model (where $n$ is the network size and $D$ is the network diameter) and this is essentially the best possible round complexity (up to logarithmic factors). However, in resource-constrained networks such as ad hoc wireless and sensor networks, nodes spending so much time can lead to significant spending of resources such as energy.

Motivated by the above consideration, we study distributed algorithms for MST under the sleeping model [Chatterjee et al., PODC 2020], a model for design and analysis of resource-efficient distributed algorithms. In the sleeping model, a node can be in one of two modes in any round — sleeping or awake (unlike the traditional model where nodes are always awake). Only the rounds in which a node is awake are counted, while sleeping rounds are ignored. A node spends resources only in the awake rounds and hence the main goal is to minimize the awake complexity of a distributed algorithm, the worst-case number of rounds any node is awake.

We present distributed MST algorithms that have optimal awake complexity with a matching lower bound. We also show that our awake-optimal algorithms have essentially the best possible round complexity by presenting a lower bound on the product of the awake and round complexity of any distributed algorithm (including randomized). Specifically, we show the following results:

- **Awake-Optimal Algorithms.** First, we present a distributed randomized algorithm to find an MST that has $O(\log n)$ awake complexity (with high probability), where $n$ is the number of nodes in the network. Our randomized algorithm has $O(n \log n)$ round complexity. Second, we show that the $O(\log n)$ awake complexity bound can be achieved deterministically as well, by presenting a distributed deterministic algorithm that has $O(\log n)$ awake complexity. However, its round complexity $O(nN \log n)$, where $N$ is the value of the maximum ID.

- **Lower Bounds.** We show that both the above algorithms have optimal awake complexity by proving that $\Omega(\log n)$ is a lower bound on the awake complexity for computing an MST even for randomized algorithms. To better understand the relationship between awake and round complexities, we prove a lower bound$^1$ of $\tilde{\Omega}(n)$ on the product of round complexity.

---

∗Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India. augustine@cse.iitm.ac.in. Supported, in part, by an Extra-Mural Research Grant (file number EMR/2016/003016) and a MATRICS grant (file number MTR/2018/001198), both funded by the Science and Engineering Research Board, Department of Science and Technology, Government of India and by the VAJRA faculty program of the Government of India.

†Department of Computer Science, University of Houston, Houston, USA. wkmjr3@gmail.com. https://orcid.org/0000-0002-4533-7593. Supported, in part, by NSF grants CCF-1540512, IIS-1633720, CCF-1717075, and BSF grant 2016419.

‡Department of Computer Science, University of Houston, Houston, USA. gopalpandurangan@gmail.com. https://orcid.org/0000-0001-5833-6592. Supported, in part, by NSF grants CCF-1527867, CCF-1540512, IIS-1633720, CCF-1717075, and BSF grant 2016419 and by the VAJRA faculty program of the Government of India.

$^1\tilde{\Omega}$ hides a $1/(\text{polylog } n)$ factor.
and awake complexity for any distributed algorithm (even randomized) that outputs an MST. This lower bound shows that our randomized algorithm that has the optimal awake complexity of $O(\log n)$ also has essentially the best possible round complexity of $O(n \log n)$.

Keywords: Minimum Spanning Tree, Sleeping model, energy-efficient, awake complexity, round complexity, trade-offs

1 Introduction

We study the distributed minimum spanning tree (MST) problem, a central problem in distributed computing. This problem has been studied extensively for several decades starting with the seminal work of Gallagher, Humblet, and Spira (GHS) in the early 1980s [11]; for example, we refer to the survey of [23] that traces the history of the problem till the state of the art. The round (time) complexity of the GHS algorithm was $O(n \log n)$ rounds, where $n$ is the number of nodes in the network. The round complexity of the problem has been continuously improved since then and now tight optimal bounds are known. It is now well-established that $\Theta(D + \sqrt{n})$ is essentially (up to logarithmic factor) a tight bound for the round complexity of distributed MST [25, 9, 19, 8]. The lower bound applies even to randomized Monte Carlo algorithms [8], while deterministic algorithms that match this bound (up to logarithmic factor) are now well-known (see e.g., [24, 21, 19, 22, 10]). Thus, the round complexity of the problem in the traditional CONGEST distributed model is settled (see also the recent works of [15, 13]). We note that in this model, any node can send, receive, or do local computation in any round and only $O(\log n)$-sized messages can be sent through any edge per round.

An MST serves as a basic primitive in many network applications including efficient broadcast [23]. For example, MST is useful for energy-efficient broadcast in wireless networks and has been extensively studied in this context, see e.g., [1, 16]. In resource-constrained networks such as wireless ad hoc and sensor networks, where nodes spend a lot of energy or other resources over the course of an algorithm, a round complexity of $\tilde{O}(D + \sqrt{n})$ to construct a MST can be large. It is worth studying whether MST can be constructed in a distributed way using a small amount of resources, i.e., resources much smaller than that taken over the (worst-case) number of rounds. We note that MST is a global problem, i.e., one needs to traverse the entire graph to compute it and hence takes at least $\Omega(D)$ rounds [18]. In the traditional model, since a node is active over the entire course of the algorithm, it can spend a lot of resources such as energy. A natural question is whether an MST can be constructed in a more efficient fashion where any node is active only for a small number of rounds.

Motivated by such considerations, this paper studies the distributed MST problem in the sleeping model [6]. In the sleeping model (see Section 1.1), nodes can operate in two modes: awake and sleeping. Each node can choose to enter the awake or asleep state at the start of any specified round. In the sleeping mode, a node cannot send, receive, or do any local computation; messages sent to it are also lost. However, the resources utilized in sleeping rounds is negligible and hence only awake rounds are counted. The goal in the sleeping model is to design distributed algorithms that solve problems in a small number of awake rounds, i.e., have small awake complexity (also called awake time), which is the (worst-case) number of awake rounds needed by any node until it terminates. This is motivated by the fact that, if the awake complexity is small, then every node takes only a small number of rounds during which it uses a significant amount of resources. For example, in ad hoc and wireless or sensor networks, a node’s energy consumption depends on the amount of

---

2 Message complexity has also been well-studied, see e.g., [23], but this is not the focus of this paper.

3 Throughout, the $\tilde{O}$ notation hides a polylog $n$ factor and $\tilde{\Omega}$ hides a $1/(\text{polylog } n)$ factor.
time it is actively communicating with nodes. In fact, significant amount of energy is spent by a node even when it is just waiting to hear from a neighbor [6]. On the other hand, when a node is sleeping — when all of its radio devices are switched off — it spends little or no energy. While the main goal is to minimize awake complexity, we would also like to minimize the (traditional) time complexity (also called run time) of the algorithm, which counts the (worst-case) total number of rounds taken by any node, including both awake and sleeping rounds.

The work of Barenboim and Maimon [2] shows that global problems such as broadcast and constructing a (arbitrary) spanning tree (but not an MST) can be accomplished in $O(\log n)$ awake rounds in the sleeping model. (A similar result can be shown for the related energy complexity model [3] — see Section A.) This is significant because it shows that even such global problems can be accomplished in a very small number of awake rounds, bypassing the $\Omega(D)$ lower bound on the round complexity (in the traditional model). In this work, we focus on another fundamental global problem, namely MST, and show that it can be solved efficiently with $O(\log n)$ awake complexity in the sleeping model, bypassing the $\tilde{\Omega}(D + \sqrt{n})$ lower bound on the round complexity (in the traditional model). We also give trade-off bounds between the awake complexity and round complexity of MST.

### 1.1 Distributed Computing Model and Complexity Measures

We consider the standard synchronous CONGEST model [24], where nodes are always “awake” from the start of the algorithm (i.e., round 1). We are given a distributed network modeled as an arbitrary undirected connected weighted graph $G(V, E, w)$, where the node set $V (|V| = n)$ represent the processors, the edge set $E (|E| = m)$ represents the communication links between them, and $w(e)$ is the weight of edge $e \in E$.

The network diameter is denoted by $D$, also called as the hop-diameter (that is, the unweighted diameter) of $G$, and in this paper by diameter we always mean hop-diameter. We assume that the weights of the edges of the graph are all distinct. This implies that the MST of the graph is unique. (The definitions and the results generalize readily to the case where the weights are not necessarily distinct.)

Each node hosts a processor with limited initial knowledge. We assume that nodes have unique IDs, and at the beginning of the computation each node is provided its ID as input and the weights of the edges incident to it. We assume node IDs are of size $O(\log n)$ bits. We assume that each node has ports (each port having a unique port number); each incident edge is connected to one distinct port. We also assume that nodes know $n$, the number of nodes in the network. For the deterministic algorithm, we make the additional assumption that nodes know the value of $N$, the largest ID of any node. Nodes initially don’t have any other global knowledge and have knowledge of only themselves.

Nodes are allowed to communicate through the edges of the graph $G$ and it is assumed that communication is synchronous and occurs in rounds. In particular, we assume that each node knows the current round number, starting from round 1. In each round, each node can perform some local computation (which happens instantaneously) including accessing a private source of randomness, and can exchange (possibly distinct) $O(\log n)$-bit messages with each of its neighboring nodes. As is standard, the goal, at the end of the distributed MST computation, is for every node to know which of its incident edges belong to the MST.

In the sleeping model [6], a node can be in either of the two states — sleeping or awake — before it finishes executing the algorithm (locally). (Initially, we assume that all nodes are awake.) That is, any node $v$, can decide to sleep starting at any (specified) round of its choice; we assume all nodes know the correct round number whenever they are awake. It can wake up again later
at any specified round and enter the awake state. In the sleeping state, a node does not send or receive messages, nor does it do any local computation. Messages sent to it by other nodes when it was sleeping are lost. We note that the model allows a node to enter a sleeping mode and awake at a later time any number of times.

Let $A_v$ denote the number of awake rounds for a node $v$ before termination. We define the \textit{(worst-case) awake complexity} as $\max_{v \in V} A_v$. For a randomized algorithm, $A_v$ will be a random variable and our goal is to obtain high probability bounds on the awake complexity. Apart from minimizing the awake complexity, we also strive to minimize the overall (traditional) round complexity (also called time complexity), where both, sleeping and awake rounds, are counted.

### 1.2 Our Contributions and Techniques

Our goal is to design distributed MST algorithms that have small awake complexity while also reducing round complexity. We present distributed MST algorithms that have \textit{optimal} awake complexity with a matching lower bound. We also show that our awake-optimal algorithms have essentially the best possible round complexity by presenting a lower bound on the product of the awake and round complexity of any distributed algorithm (including randomized). Please see Table 1.

| Algorithm     | Type            | Awake Time (AT) | Run Time (RT) | AT Lower Bound | AT $\times$ RT Lower Bound |
|---------------|-----------------|----------------|--------------|----------------|-----------------------------|
| *Randomized-MST* | Randomized      | $O(\log n)$   | $O(n \log n)$ | $\Omega(\log n)$ | $\Omega(n)$                |
| *Deterministic-MST* | Deterministic   | $O(\log n)$   | $O(nN \log n)$ | $\Omega(\log n)$ | $\Omega(n)$                |

*This algorithm outputs an MST with high probability.

Our lower bounds also apply to Monte Carlo randomized algorithms with constant success probability.

$n$ is the number of nodes in the network and $N$ is the largest node ID.

#### Awake-Optimal Algorithms and Techniques

Our first result is a distributed randomized algorithm in the sleeping model that has $O(\log n)$ awake complexity (see Section 2.2). In contrast to the round complexity lower bound of $\tilde{\Omega}(D + \sqrt{n})$, the awake complexity bound of the awake-optimal MST algorithm is significantly smaller. The round complexity of our algorithm is $O(n \log n)$.

We then show that the awake-optimal bound of $O(\log n)$ can be obtained deterministically by presenting a deterministic algorithm (see Section 2.3). However, the deterministic algorithm is more involved compared to the randomized algorithm and has a worse round complexity of $O(nN \log n)$, assuming that the node IDs are in the range $[1, N]$ and $N$ is known to all nodes.

Our algorithms use several techniques to implement the standard GHS algorithm for constructing an MST in an awake-efficient manner. The main challenge is to implement the two keys steps of GHS algorithm, namely finding the minimum outgoing edge and merging fragments efficiently in an awake-efficient manner. We define and use a tree called the Labeled Distance Tree (LDT) crucially in our construction which can be of independent interest. A LDT is a spanning tree such that each node is labelled by its distance from the root and every node knows the labels of its parent and children (if any). The main technical challenge is to construct a LDT in an awake-efficient way, in $O(\log n)$ awake time. This is non-trivial in the sleeping model since all neighboring nodes have to be awake simultaneously. In a weighted graph, a LDT can be constructed in an awake-efficient manner in a such a way that it is also an MST. Our approach is similar, yet different, to the construction of trees in [2, 3]. In particular, a similar, but somewhat different, tree structure called as Distributed Layered Tree (DLT) is used in Barenboim [2]. A DLT is a rooted tree where the

\footnote{For details on how this can be implemented, say, in wireless networks, see [6].}
vertices are labeled, such that each vertex has a greater label than that of its parent, according to a given order and each vertex knows its own label and the label of its parent. Another similar tree structure is used in [3]. Our LDT construction is different compared to the tree constructions in [2, 3], and somewhat simpler (especially, the randomized construction) compared to the DLT of [2]. More importantly, our LDT allows efficient construction of MST as opposed to any arbitrary spanning tree as in [2, 3].

**Lower Bounds.** We show that our algorithms achieve optimal awake time since we show that $\Omega(\log n)$ is a lower bound on the awake complexity of constructing a MST, even for randomized Monte Carlo algorithms (see Section 3.1). The round complexity of the awake-optimal randomized MST algorithm is $O(n \log n)$ which is significantly larger than the best possible bound of $\tilde{\Theta}(D + \sqrt{n})$. This motivates the natural question whether one can attain a small awake complexity, i.e., polylogarithmic in $n$, while also achieving optimal round complexity of $\tilde{\Theta}(D + \sqrt{n})$. Our next result shows that this is not possible in general by showing a lower bound on the product of awake and round complexities. Specifically, it shows that there exist graphs with an optimal round complexity of $c$, then the awake complexity has to be $\Omega(n/c)$ for any distributed algorithm, even for Monte-Carlo randomized algorithms with constant success probability. The precise result in stated in Theorem 4 (see Section 3.2). In other words, the product of the round complexity and awake complexity is $\tilde{\Omega}(n)$.

Our lower bound technique for showing a conditional lower bound on the awake complexity (conditional on upper bounding the round complexity) can be of independent interest. We use a lower bound graph family that is similar to that used in prior work (e.g., [8]), but our lower bound technique uses communication complexity to lower bound awake complexity by lower bounding the congestion caused in some node. This is different from the Simulation Theorem technique ([8]) used to show lower bounds for round complexity in the traditional setting.

For lack of space, all omitted proofs and additional related work are placed in the Appendix.

### 2 Optimal Awake Time MST Algorithms

In this section, we present our algorithms to create an MST that take optimal awake complexity. We first present a toolbox of procedures which are used repeatedly in the subsequent algorithms. We then develop a randomized algorithm that creates an MST in optimal awake time. We then show how to construct a deterministic algorithm that also is optimal in awake time, however, we pay a cost in terms of slower run time.

#### 2.1 Main Ideas & Useful Procedures

Both of the algorithms we develop in this section can be seen as variations of the classical GHS algorithm to find the MST, adapted to optimize the awake time of the algorithm. Recall that each phase of the traditional GHS algorithm consists of two steps. Step (i) corresponds to finding the minimum outgoing edges (MOEs) for the current fragments and step (ii) involves merging these fragments.

In the course of both our algorithms, we maintain a certain structure over the graph. Specifically, our algorithms work in phases where at the end of each phase, we ensure that the original graph has been partitioned into a forest of disjoint trees that satisfy the following property. For each such tree, all nodes within the tree know the ID of the root of the tree (called fragment ID), their parents and children, if any, in the tree, and their distance from the root (note that it is the hop distance, ignoring the weights) of that tree. We call each such a tree a Labeled Distance Tree (LDT) and a forest of such trees a Forest of Labeled Distance Trees (FLDT). By the end of the algorithms...
we design, our goal is to have the FLDT reduce to just one LDT which corresponds to the MST of the original graph. The challenge is to construct an LDT (which will also be an MST) in an awake-optimal manner.

The purpose of maintaining such a structure is that we know how to design fast awake procedures to propagate information within an LDT. Specifically, we show how to set up a schedule of rounds for nodes to wake up in so that information can be passed from the root to all children in an LDT in $O(1)$ awake rounds, information can be passed from a child to the root in $O(1)$ awake rounds, and information can be spread from one LDT to its neighbors in $O(1)$ awake rounds. There are several well-known procedures typically associated with GHS such as broadcast, upcast-min, etc. that make use of such information propagation. Once we have an LDT, it is easy to implement these procedures in constant awake time; we place the detailed descriptions of these procedures in Appendix B and give a brief description below. For the processes described below, it is assumed that the initial graph has already been divided into an FLDT where each node $u$ knows the ID of the root, $\text{root}$, of the tree it belongs to as well as $u$’s parent and children in that tree. First of all, we define a transmission schedule that is used in each of the procedures and will be directly utilized in the algorithms. Then we briefly mention the procedures typically associated with GHS.

Transmission schedule of nodes in a tree. We describe the function $\text{Transmission-Schedule}(\text{root}, u, n)$, which takes a given node $u$ and a tree rooted at $\text{root}$ (to which $u$ belongs) and maps them to a set of rounds in a block of $2n + 1$ rounds. $\text{Transmission-Schedule}(\text{root}, u, n)$ may be used by node $u$ to determine which of the $2n + 1$ rounds to be awake in. Consider a tree rooted at $\text{root}$ and a node $u$ in that tree at distance $i$ from the root. For ease of explanation, we assign names to each of these rounds as well. For all non-root nodes $u$, the set of rounds that $\text{Transmission-Schedule}(\text{root}, u)$ maps to includes rounds $i$, $i + 1$, $n + 1$, $2n - i + 1$, and $2n - i + 2$ with corresponding names $\text{Down-Receive}$, $\text{Down-Send}$, $\text{Side-Send-Receive}$, $\text{Up-Receive}$, and $\text{Up-Send}$, respectively. $\text{Transmission-Schedule}(\text{root}, \text{root}, n)$ only maps to the set containing rounds 1, $n + 1$, and $2n + 1$ with names $\text{Down-Send}$, $\text{Side-Send-Receive}$, and $\text{Up-Receive}$, respectively.\(^5\)

This transmission schedule plays an important role in ensuring that the following procedures related to GHS take only $O(1)$ awake time. During one instance of $\text{Transmission-Schedule}(\text{root}, u, n)$, by having each node wake up in a carefully selected non-empty subset of its 5 named rounds, we guarantee that all nodes have woken up at least once and that information is propagated in the correct “direction” in the fragment as needed.

Broadcasting a message in a tree. Procedure $\text{Fragment-Broadcast}(n)$, run by all nodes in a tree, allows the root of that tree to transmit a message to each node in its tree in $O(n)$ running time and $O(1)$ awake time.

Upcasting the minimum value in a tree (also called convergecast). Procedure $\text{Upcast-Min}(n)$, run by all nodes in a tree, allows the smallest value among all values held by the nodes, if any, to be propagated to the root of the tree in $O(n)$ running time and $O(1)$ awake time.

Transmitting a message between nodes of adjacent trees. Procedure $\text{Transmit-Adjacent}(n)$, run by all nodes in a tree, allows each node in a tree to transfer a message, if any, to neighboring nodes belonging to other trees in $O(n)$ running time and $O(1)$ awake time.

2.2 Awake-Optimal Randomized Algorithm

Brief Overview. We describe Algorithm $\text{Randomized-MST}$. We use a synchronous variant of the classical GHS algorithm (see e.g., [23, 24]) to find the MST of the initial graph, where we

\(^5\)In this description, we assumed that $\text{Transmission-Schedule}(\cdot, \cdot, n)$ was started in round 1. However, if $\text{Transmission-Schedule}(\cdot, \cdot, n)$ is started in round $r$, then just add $r - 1$ to the values mentioned here and in the previous sentence to get the correct rounds.
modify certain parts of the algorithm in order to ensure the awake time is optimal. Each phase of the traditional GHS algorithm consists of two steps. Step (i) corresponds to finding the minimum outgoing edges (MOEs) for the current fragments and step (ii) involves merging these fragments. In the current algorithm, we focus on modifying step (i). Specifically, we utilize randomness to restrict which MOEs are considered “valid” for the current phase by having each fragment leader flip a coin and only considering MOEs from fragments whose leaders flip tails to those that flip heads. This restriction ensures that the diameter of any subgraph formed by fragments and valid MOEs is constant. This restriction, coupled with careful analysis that includes showing that the number of fragments decreases by a constant factor in each phase on expectation, guarantees that the MST of the original graph is obtained after \(O(\log n)\) phases with high probability.

**Detailed Algorithm.** Algorithm RANDOMIZED-MST consists of nodes participating in \(4\lceil \log_{4/3} n \rceil + 1\) phases of the following computations. Recall that between phases we want to maintain a FLDT that is eventually converted into a single LDT. Also recall that in each phase, there are two steps. We describe the two steps in detail below.

**Step (i): Finding MOE of each fragment.** In Step (i), all fragments perform the following series of actions. Each node in a fragment participates in the following sequence of procedures: first \textsc{Fragment-Broadcast}(n) to transmit the message to find the MOE to all nodes in the fragment (i.e., to find the “local” MOEs from each node in the fragment to outside), second \textsc{Upcast-Min}(n) to convergecast the smallest among the (local) MOEs to the root of the fragment to find the overall MOE of the fragment, then \textsc{Fragment-Broadcast}(n) to transmit the MOE to all nodes in the fragment, and finally \textsc{Transmit-Adjacent}(n) to inform adjacent fragments of the current fragment’s MOE.

We note that because of the properties of the LDT, \textsc{Fragment-Broadcast}(n), \textsc{Upcast-Min}(n) and \textsc{Transmit-Adjacent}(n) can be accomplished in \(O(1)\) awake time and hence the finding MOE of each fragment can be accomplished in \(O(1)\) awake time.

Consider the supergraph formed by taking fragments as the node set and their MOEs as edges. In this supergraph, there may be trees of large diameter, say some \(k\). If we were to try merging such large diameter subgraphs, it may take \(O(k)\) awake time as a new fragment ID can be propagated to each adjacent node in the supergraph in \(O(1)\) time and such info may have to be propagated \(O(k)\) times. Hence, we shorten the diameter of subgraphs. Intuitively, one might consider this as cutting down a forest of trees to a forest of stars, done as follows. However, this means we only merge some fragments (the ones that have valid MOEs). We show later in the analysis of merging that this still reduces the number of fragments by a constant factor in every phase, so we finish in \(O(\log n)\) phases.

We have each fragment’s root flip a coin and only allow MOEs from fragments whose roots flipped tails to those that flipped heads.\(^6\) Henceforth, we say that a fragment flipped either a heads (tails) or is considered a heads (tails) fragment if that fragment’s root flipped a heads (tails). Each fragment root flips an unbiased coin and uses \textsc{Fragment-Broadcast}(n) to inform the nodes in the fragment if the root flipped a tails or a heads. All nodes participate in a \textsc{Transmit-Adjacent}(n) to inform adjacent fragments of the current fragment’s coin flip. Now, we only consider an MOE as “valid” if it originated from a fragment that flipped a tails and is to a fragment that flipped a heads. Now, each node adjacent to an MOE knows if that MOE is a valid one or not. By using \textsc{Upcast-Min}(n) and \textsc{Fragment-Broadcast}(n) to transmit the info of “valid” or “invalid”, all nodes in the fragment will know if the outgoing MOE is valid or not.

**Step (ii): Merging fragments.** In Step (ii), we merge each subgraph formed by fragments

\(^6\)Intuitively, the center of the star is a fragment whose root flipped heads and the leaves, if any, are fragments whose roots flipped tails.
and valid MOEs into a single fragment. Consider a subgraph consisting of several tails fragments whose MOEs lead to a single heads fragment. The heads fragment retains its fragment ID while the remaining fragments take on the ID of the heads fragment. Furthermore, these other fragments also re-orient themselves such that they form subtrees of the heads fragment. Specifically, consider a tails fragment \( T \) with root \( root_T \) and an MOE to a heads fragment \( H \) where nodes \( u_T \) and \( u_H \) are the nodes of the MOE belonging to \( T \) and \( H \), respectively. The nodes in fragment \( T \) re-orient themselves such that \( u_T \) is the new root of the tree. Additionally, \( u_T \) considers \( u_H \) its parent in the merged graph. This is described below in Procedure \textsc{Merging-Fragments}(n), a process similar to that in [2]. The process is also illustrated in Figures 2, 3, 4, and 5 found in Appendix C. Recall that at the end of the previous step, each node in the fragment knows whether the fragment root flipped heads or tails and whether the MOE leading out of the fragment is a valid MOE or not.

As mentioned earlier in Section 2, each node in an LDT \( T \) maintains the ID of the fragment it belongs to as well as the distance of the node from the root of the fragment. We show how to correctly maintain this information when we merge LDTs with the help of two temporary variables, \textsc{NEW-FRAGMENT-ID} and \textsc{NEW-LEVEL-NUM}, respectively. Additionally, each node of a fragment that updates these values must also re-orient itself by updating its parent and child pointers so that the overall merged fragment is also a tree. In the course of describing how to update fragment ID and distance to the root, we also show how to identify a node’s new parent and children, if any. This information is recorded by the node and we mention when the node internally updates this information.\(^7\)

First, all nodes participate in \textsc{Transmit-Adjacent}(n) to transmit their fragment ID and level number. Now \( u_T \) sets its \textsc{NEW-LEVEL-NUM} to that of \( u_H \) plus one and stores the fragment ID of \( u_H \) in \textsc{NEW-FRAGMENT-ID}. Additionally, \( u_T \) records the info that \( u_H \) is its new parent and its neighbors in \( T \) are its new children, to be updated internally later on. The remaining nodes of \( T \) initialize \textsc{NEW-FRAGMENT-ID} and \textsc{NEW-LEVEL-NUM} to \( \perp \), which we also refer to as those values being empty. Now, all nodes \( v \) in the fragment participate in two instances of \textsc{Transmission – Schedule}(\( root_T, v, n \)). At a high level, the first instance is used to update the \textsc{NEW-FRAGMENT-ID} and and \textsc{NEW-LEVEL-NUM} of nodes on the path from \( u_T \) to \( root_T \). The second instance is used to update the required variables for all the remaining nodes.

In the first instance, each node \( v \) during its \textsc{Up-Send} round sends up the value in its \textsc{NEW-LEVEL-NUM} if it is not empty and \( v \) also sends up the value of \textsc{NEW-FRAGMENT-ID} if it is not empty. During an \textsc{Up-Receive} round, if \( v \) receives a non-empty \textsc{NEW-LEVEL-NUM} from its child, \( v \) sets its own \textsc{NEW-LEVEL-NUM} to the received value plus one. Similarly, if \( v \) receives a non-empty \textsc{NEW-FRAGMENT-ID}, \( v \) sets its own \textsc{NEW-FRAGMENT-ID} to that value. Additionally, if \( v \) receives a non-empty \textsc{NEW-LEVEL-NUM}, from its child, it records internally that its child will be its new parent and its neighbors in \( T \) will be its children.

In the second instance of \textsc{Transmission – Schedule}(\( root_T, v, n \)), each node \( v \) during its \textsc{Down-Send} round sends down the value of its \textsc{NEW-LEVEL-NUM} and also the value of its \textsc{NEW-FRAGMENT-ID}. During a \textsc{Down-Receive} round, if \( v \)’s \textsc{NEW-LEVEL-NUM} is non-empty and it receives a non-empty value from its parent, \( v \) updates its \textsc{NEW-LEVEL-NUM} to that value plus one. Similarly, if \( v \)’s \textsc{NEW-FRAGMENT-ID} is non-empty and it receives a non-empty value from its parent, \( v \) updates its \textsc{NEW-FRAGMENT-ID} to that value.

At the end of Step (ii), each node \( v \) updates its fragment ID to \textsc{NEW-FRAGMENT-ID} and updates its level number to \textsc{NEW-LEVEL-NUM}, assuming they are non-empty, and subsequently clears those variables, i.e., sets their values to \( \perp \), in preparation for the next phase.\(^8\) Additionally,\(^7\)

\[^7\]This re-orientation information is not immediately updated to avoid complications when we call a subprocedure that is to be run on the old tree.

\[^8\]\textsc{NEW-FRAGMENT-ID} and \textsc{NEW-LEVEL-NUM} are variables that serve a temporary purpose each phase. As such,
any information about re-orientation, i.e., updating parent and/or children, is now updated locally within the tree.

**Analysis.** We now prove that the algorithm correctly outputs the MST of the original graph with the desired running time and awake time.

Recall that the number of fragments can never increase from one phase to the next. Let phase $P$ correspond to the last phase in which there is more than one fragment at the beginning of the phase. We will show that $P = 4\lceil\log_{4/3} n\rceil$. The following lemma shows that for the first $P$ phases of the algorithm, the number of fragments reduces by a constant factor in each phase with high probability.

**Lemma 1.** For each phase of Algorithm Randomized-MST where there are initially at least two fragments at the start of that phase, the number of fragments is reduced by at least a factor of $4/3$ in that phase on expectation. Furthermore, by phase $4\lceil\log_{4/3} n\rceil + 1$, there is at most one fragment in the graph.

We are now ready to argue that the algorithm is correct.

**Lemma 2.** Algorithm Randomized-MST results in each node of the initial graph knowing which of its edges are in the MST with high probability.

We also bound the running time and awake time of the algorithm below.

**Lemma 3.** Algorithm Randomized-MST takes $O(n \log n)$ running time and $O(\log n)$ awake time.

**Theorem 1.** Algorithm Randomized-MST is a randomized algorithm to find the MST of a graph with high probability in $O(n \log n)$ running time and $O(\log n)$ awake time.

### 2.3 Awake-Optimal Deterministic Algorithm

**Brief Overview.** We describe Algorithm Deterministic-MST, which is similar to Algorithm Randomized-MST (described in Section 2.2). Unlike Algorithm Randomized-MST, where where we use random coin flips in step (i) to limit the diameter of subgraphs in the fragment graph (where each fragment is a (super)node and the edges between the fragments are the MOEs), the main technical challenge is to deterministically keep the diameter of these merging subgraphs (in the fragment graph) to be constant; this is crucial for implementing one phase of merging in $O(1)$ awake time.\(^9\)

We use the following combination of techniques. In step (i), we first sparsify such graphs by allowing any fragment to be adjacent to at most 4 other fragments. Then, in step (ii), we use a fast awake time coloring algorithm and selective merging to ensure that subgraphs do not have a large diameter. We leverage this property to ensure fast awake time merging of these fragments occurs. We also ensure that a sufficient number of fragments participate in merging so that the number of fragments decreases from one phase to the next. The above modifications result in an algorithm that outputs the MST of the original graph in $\lceil\log_{240000/239999} n\rceil + 240000$ phases. Note that we require nodes to know the value of $N$, the range from which node IDs are taken.

**Detailed Algorithm.** We now give a detailed break up of each phase of the algorithm. Recall that there are $\lceil\log_{240000/239999} n\rceil + 240000$ phases and in each phase, there are two steps. We describe each in detail separately.

\(^9\)One can accomplish this deterministically in $O(\log^* n)$ time even in the traditional model using coloring or maximal matching in the fragment graph which is a tree (see e.g., [22]), but this will lead to an overhead of $O(\log^* n)$ factor in the awake time.
Step (i): Finding MOE of each fragment. Initially, we find each fragment’s MOE in a manner similar to Step (i) of Algorithm Randomized-MST.

Once that is over, each node in a given fragment knows, for each of its edges adjacent to it, whether that edge is an MOE from some other fragment to the given fragment. Let us differentiate these MOEs to the fragment from the MOE from the fragment by calling the former INCOMING-MOE. Now, we have each fragment select up to 3 “valid” MOEs from its INCOMING-MOE, chosen arbitrarily. This is in contrast to how the valid MOEs were chosen during Algorithm Randomized-MST, where we used coin flips to determine valid MOEs. Define an incoming MOE node \( v \) of fragment \( f \) as a node \( v \) belonging to fragment \( f \) such that \( v \) is adjacent to an edge that is an MOE from some other fragment to \( f \). In the context of a given fragment \( f \), define a valid MOE child node \( v \) of a node \( u \) as a child node of \( u \) such that in the subtree rooted at \( v \), there exists an incoming MOE node of \( f \). At a high level, the total number of INCOMING-MOE is communicated to the root of the fragment. The root then allots up to 3 virtual “tokens” (i) to its valid MOE child nodes to be used to select INCOMING-MOE and (ii) to itself if the root is an incoming MOE node. Any node that receives one or more such tokens distributes them among its valid MOE child nodes and itself if it is an incoming MOE node. This process is repeated until all tokens are transmitted to incoming MOE nodes of the fragment.

In more detail, consider a given fragment \( f \) with root \( v \). Each node \( u \) that belongs to \( f \) runs Transmission – Schedule(root, \( u, n \)). In \( u \)'s Up-Receive round, \( u \) receives info on the number of nodes in its children’s subtrees (if any) that are incoming MOE nodes. In \( u \)'s Up-Send round, \( u \) aggregates the total number of such incoming MOE nodes. It adds 1 to that number if \( u \) is itself an incoming MOE node. Then \( u \) sends this number up to its parent in the fragment. Now, at the end of Transmission – Schedule(root, \( u, n \)), the root of the fragment is aware of how many nodes in the fragment (including the root itself) are incoming MOE nodes. If that number is \( \leq 3 \), then all of them are accepted as valid, otherwise at most 3 of them are selected as valid as follows. All nodes \( v \) in the fragment participate in Transmission – Schedule(root, \( u, n \)). If the root is an incoming MOE node, it selects itself as a valid incoming MOE node. The root then decides in an arbitrary manner how many nodes from each of its children’s subtrees can become valid MOE nodes. This information is transmitted during the root’s Down-Send round in the form a number indicating how many MOE nodes can be in the subtree rooted at that child. As for an arbitrary node \( u \) in the fragment, in \( u \)'s Down-Receive round, it receives this number from its parent. Subsequently, during \( u \)'s Down-Send round, \( u \) does the following. If \( u \) is an incoming MOE node, then \( u \) selects itself as a valid incoming MOE node and reduces the number by one. If the number \( u \) has is non-zero, then \( u \) distributes the number among its children that have INCOMING-MOE in their trees.

For every fragment \( f \), the information of whether an incoming MOE from a fragment \( f' \) is valid or not is communicated to each \( f' \) as follows. All nodes in \( f \) and every other fragment run Transmit-Adjacent(n) and MOE nodes communicate whether they are valid or not to their neighbors. Let \( u \) be an incoming MOE node from fragment \( f \) whose neighbor in \( f' \) (and the other end of the MOE) is node \( v \). Now, \( v \) from \( f' \) knows if \( (v, u) \) is a valid MOE from \( f' \) to \( f \) or not. Now all nodes participate in Upcast-Min with the following conditions: (i) nodes other than \( v \) send up \( \infty \), (ii) if \( (v, u) \) is a valid MOE, then \( v \) sends up the edge weight of \( (v, u) \) else it sends up \( -\infty \). As a result, once Upcast-Min is finished, the root of fragment \( f' \) will see that \( (v, u) \) is the MOE with either its original weight or \( -\infty \) as the edge weight and can thus conclude if \( (v, u) \) is a valid MOE or not.

Step (ii): Merging fragments. In Step (ii), we first make each fragment \( f \)'s nodes aware of the fragment IDs from and to which it has valid MOEs. We first collect information about valid MOEs (both incoming & outgoing) at the root of the fragment \( f \). Then we have all nodes in the fragment participate in a Fragment-Broadcast(n) to send this information (which takes only
$O(\log n)$ bits to encode information about all such MOEs). More detail follows below.

Recall that at the end of step (i), each node in the fragment graph $G'$ (formed by the fragments as (super)nodes and valid MOEs as edges) has at most 3 valid incoming MOEs and at most one valid outgoing MOE. Thus, the maximum degree of any node in $G'$ is 4. We first make each fragment aware of all its valid MOEs. At a high level, we have all valid MOE nodes in the fragment send up info on their MOEs to the root of the fragment. Thus, we run Procedure $\text{NEIGHBOR-AWARENESS}(n)$, a variant of $\text{UPCAST-MIN}(n)$, where (i) each valid MOE node $u$ (both incoming and outgoing) to some fragment $f'$ sends up a tuple of $\langle$ $u$'s ID, weight of $u$'s edge to $f'$, fragment ID of $f'$, color of $f'$ $\rangle$, (ii) other nodes send up a tuple of $\langle\infty, \infty, \infty, -1\rangle$ if they do not have another value to send up, and (iii) each node, instead of only sending up the one tuple it knows of, sends up a concatenation of the at most 4 tuples that have non-$\infty$ values. Note that the fourth value of the tuple, color of $f'$, is set to $-1$ throughout this procedure as the fragments and their nodes are as yet uncolored. However, this procedure is used elsewhere where color is needed. Now, all nodes participate in Procedure $\text{FRAGMENT-BROADCAST}(n)$ so that all nodes of $f$ are aware of tuples corresponding to the at most 4 neighboring fragments of with valid MOEs from/to $f$. Each node stores this info in the variable $\text{NBR-INFO}$.

Subsequently, we color the fragments and then selectively merge them. Consider a color palette consisting of colors Blue, Red, Orange, Black, and Green. Furthermore, let there exist a total ordering on this palette based on the relation of priority, where we say that a color $A$ has a higher priority than color $B$ and denote the relation by $A > B$, such that Blue $>$ Red $>$ Orange $>$ Black $>$ Green. Let us consider the supergraph $G'$ where the nodes are the set of fragments present at the beginning of the phase and the edges are the valid MOEs, as computed from step (i). Recall that the maximum degree of any node in $G'$ is 4, so 5 colors are sufficient for coloring (there always exists a $\Delta + 1$ coloring of a graph with maximum degree $\Delta$). At a high level, we wake fragments up in order of the fragment IDs and color them with the first available color, i.e., the highest priority color not chosen by any of its neighbors.

We now describe the coloring in more detail. Intuitively, we have fragments wake up in the order of their fragment IDs. Once awake, the fragment selects the highest priority color not yet taken by any of its neighboring fragments. However, the above description does not mention how neighboring fragments become aware of a given fragments chosen color. This is done by having a fragment $F$ AND its neighboring fragments wake up in the rounds that $F$ was assigned. As a given fragment has $O(1)$ neighbors, it does not have to stay awake for too long. Now, we describe the procedure in more detail below.

All nodes perform the Procedure $\text{FAST-AWAKE-COLORING}(n, N)$, described below which requires each node to know both the total number of nodes $n$ and the range from which every ID is taken $[1, N]$. Recall that all nodes in each fragment $f$ are aware of any neighboring fragment $f'$ (including its fragment ID) with a valid MOE from/to $f$. Furthermore, each node in $f$ is aware of the fragment ID of $f$. All fragment IDs are numbers in the range $[1, N]$. We have all nodes participate in $N$ stages of the following. In stage $i$, only nodes belonging to a fragment with fragment ID $i$ or belonging to a fragment with a valid MOE to/from a fragment with fragment ID $i$ participate in the following process while other nodes are asleep. Abusing notation, let node $u$ belong to fragment $f$ with fragment ID $f$. If $f = i$, then node $u$ chooses the highest priority color not selected by fragments in $\text{NBR-INFO}$ to be $f$'s color. This information is communicated to the root via $\text{UPCAST-MIN}(n)$ and subsequently communicated to all nodes in $f$ via $\text{FRAGMENT-BROADCAST}(n)$. Finally, $\text{NEIGHBOR-AWARENESS}(n)$ is run so that all nodes in neighboring fragments know the color of $f$.

Now, we describe the selective merging in more detail. We first identify the set of fragments that will merge into other fragments and will thus “not survive” the phase. These are all fragments
that were colored Blue during Procedure FAST-AWAKE-COLORING\((\text{n} , \text{N})\). Recall that there are two types of fragments that are colored blue. Those with neighbors in \(G'\) and those without, which we call singleton fragments.

Those Blue fragments with neighbors pick one of their neighbors in \(G'\) arbitrarily (which is of course a non-Blue fragment) and then merge into them. This can be achieved by running Procedure MERGING-FRAGMENTS\(\big(n\big)\) from Section 2.2 where we consider Blue fragments as Tails fragments and all non-Blue fragments as Heads fragments. The merged fragment takes on the fragment ID of the fragment that acted as the Heads fragment.

We now look at how to merge the singleton fragments into the remaining fragments. During Procedure FAST-AWAKE-COLORING\((\text{n} , \text{N})\), these singleton fragments are colored Blue but they did not merge into other fragments during the previous instance of Procedure MERGING-FRAGMENTS\(\big(n\big)\). We have each of these Blue fragments merge into the fragment at the end of its MOE. But before doing that, we need each node in these Blue fragments to become aware of any changes to fragment IDs and level numbers of neighboring nodes. This information is conveyed by having all nodes in the original graph participate in one instance of TRANSMIT-ADJACENT\(\big(n\big)\) to inform nodes in neighboring fragments about such updates. Subsequently, all nodes participate in Procedure MERGING-FRAGMENTS\(\big(n\big)\) where we consider these Blue singleton fragments as Tails fragments and all remaining fragments as Heads fragments. The merged fragment takes on the fragment ID of the fragment that acted as the Heads fragment. At the end of all this, each of the previous singleton fragments is merged into some other fragment and does not survive the phase.

**Analysis.** We now prove that the algorithm correctly outputs the MST of the original graph with the desired running time and awake time.

We show that the number of phases needed to reduce the number of fragments to one is at most \(\lceil \log_{240000/239999} \text{n} \rceil + 240000\). Correctness immediately follows as we implement GHS. We first argue that in each phase of the algorithm where there is initially a sufficient number of fragments \(\text{N} = 240000\), the number of fragments is reduced by a constant factor of \(240000/239999\). We then show that in an additional \(\text{N}\) phases, we can reduce the number of fragments to one. As we only add MST edges and all nodes will be in this one fragment, the final fragment represents the MST of original graph.

Let \(P\) represent the phase by which the number of fragments at the beginning of the phase is less than \(\text{N}\). We eventually show that \(P = \lceil \log_{240000/239999} \text{n} \rceil\). We first argue that in every phase up to \(P\), the number of fragments is reduced by a constant factor. We do this by considering an arbitrary phase \(i\) and identifying a set of fragments in that phase that are guaranteed to merge into other fragments, thus “being lost” or “not surviving” in that phase. We show that this set is at least a constant fraction of the total set of fragments that existed at the start of the phase.

Consider an arbitrary phase \(i\) such that at the beginning of the phase there exists a set \(F_i\) of fragments and define \(F_i = |F_i|\). Furthermore, define the supergraph \(H\) as the undirected graph where the nodes are the set \(F_i\) and the edges are the valid MOEs between the different fragments, i.e., the graph obtained after pruning MOEs in step (i) of the phase. In the subsequent analysis we use nodes and fragments interchangeably in the context of graph \(H\). We now show that the number of blue fragments (which by the algorithm are all merged into other fragments) constitute a sufficiently large constant fraction of \(F_i\).

**Lemma 4.** Let \(H'\) be a connected subgraph of \(H\). If \(|H'| \geq 342\), then at least \(\lceil |H'|/342 \rceil\) of the fragments are Blue.

The above lemma by itself is insufficient to show that the required number of fragments are removed in each phase. The reason is that \(H\) may consist of a set of disjoint connected subgraphs.
Let us assume that $|H| \geq 240000$. Let $S$ denote the set of all disjoint connected subgraphs (i.e., connected components) in $H$. Now, either $|S| \geq |H|/342^2$ or $|S| < |H|/342^2$.

If $|S| \geq |H|/342^2$, then since each subgraph in $S$ contains at least one Blue fragment which disappears in the phase, the total number of fragments that survive the phase is at most $|H| - |H|/342^2 \leq 239999|H|/240000$.\(^{10}\)

Now let us look at the situation where $|S| < |H|/342^2$. Divide $S$ into the sets $S_1$ and $S_2$ which contain the disjoint connected subgraphs of $H$ which have $< 342$ fragments and $\geq 342$ fragments, respectively. Observe that $S = S_1 \cup S_2$. (It is easy to see that $|S_2| \geq 1$ since otherwise if all subgraphs belonged to $S_1$, there would be less than $|H|$ total fragments.) We now show that a sufficient number of fragments in the subgraphs in set $S_2$ are Blue fragments, thus resulting in a sufficient number of fragments being removed in the phase. Let us lower bound how many fragments are present in the subgraphs in set $S_2$. Recall that $|S_1| \leq |S|$, we are considering the situation where $|S| < |H|/342^2$, and each subgraph in $|S_1|$ can have $< 342$ fragments. We lower bound the fragments in $S_2$ by pessimistically ignoring the less than 342 fragments from each of the $|S_1|$ subgraphs (recall that $|S_1| \leq |S|$), i.e., the number of fragments in $S_2$ is

\[
\geq |H| - \left(\frac{|H|}{342^2}\right) \cdot 342 \quad \text{(since the total number of subgraphs is at most $|H|/342^2$)}
\]

\[
= 341|H|/342.
\]

We now lower bound the number of Blue fragments in subgraphs in $S_2$. Let the subgraphs in $S_2$ be denoted by $H_1, H_2, \ldots, H_{|S_2|}$. Since each subgraph in $S_2$ is of size at least 342, we can use Lemma 4. Now, the number of Blue fragments in $S_2$ is

\[
= \sum_{i=1}^{S_2} \left(\frac{|H_i|}{342}\right) - |S_2|
\]

\[
\geq 1/342 \cdot \left(\sum_{i=1}^{S_2} |H_i|\right) - |S_2|
\]

\[
\geq 1/342 \cdot (341|H|/342) - |S_2| \quad \text{(since $|S_2| \leq |S| < |H|/342^2$)}
\]

\[
= 340|H|/342^2.
\]

Recall that all Blue fragments do not survive a phase. Thus, the number of fragments that survive the current phase is

\[
\leq |H| - 340|H|/342^2
\]

\[
\leq 239999|H|/240000.
\]

Thus, in both situations, we see that the number of fragments that survives the present phase is upper bounded as desired.

\(^{10}\)It is easy to see that each subgraph contains a Blue fragment because Blue is the highest priority color and so the first fragment that colors itself in any given subgraph colors itself Blue.
Lemma 5. After $\lceil \log_{240000}/239999 \cdot n \rceil$ phases, there are at most 240000 fragments at the beginning of the phase.

An easy observation is that whenever $F_i \geq 2$, at least one fragment is Blue and will not survive the phase. Thus, if there are at most 240000 fragments at the beginning of the phase, then running an additional 240000 phases guarantees that only one fragment will remain. Initially, each node is a fragment by itself and over the course of the algorithm, only possible MST edges are added to any fragment. Thus, we have the following lemma.

Lemma 6. Algorithm Deterministic-MST correctly outputs an MST after $\lceil \log_{240000}/239999 \cdot n \rceil + 240000$ phases.

We analyze the running time and awake time for each node in each phase. Since there are $O(\log n)$ phases, it is easy to then get the overall running and awake times.

Let us look at each step of a phase individually. In Step (i), all nodes participate in one instance of Fragment-Broadcast($n$), one instance of Upcast-Min($n$), two instances of Transmit-Adjacent($n$), and two instances of Transmission - Schedule($\cdot$,$\cdot$,$n$). Each of these procedures takes $O(1)$ awake time and $O(n)$ run time.

In Step (ii), all nodes participate in one instance of Fragment-Broadcast($n$), one instance of Neighbor-Awareness($n$), one instance of Fast-Awake-Coloring($n$, $N$), two instances of Merging-Fragments($n$), and one instance of Transmit-Adjacent($n$). We know that Fragment-Broadcast($n$) and Transmit-Adjacent($n$) each take $O(1)$ awake time and $O(n)$ running time. Neighbor-Awareness($n$) is a variant of Upcast-Min($n$) and from its description, it is easy to see that it too takes $O(1)$ awake time and $O(n)$ run time. Merging-Fragments($n$) was described and analyzed in the previous section and took $O(1)$ awake time and $O(n)$ running time. We see that Fast-Awake-Coloring($n$, $N$) consists of $N$ stages. However, each node is awake for and participates in at most 5 such stages. In each stage, nodes that are awake participate in one instance of Upcast-Min($n$), one instance of Fragment-Broadcast($n$), and one instance of Neighbor-Awareness($n$). Thus, each stage contributes $O(1)$ awake time for participating nodes and $O(n)$ running time. Thus, Fast-Awake-Coloring($n$, $N$) has $O(1)$ awake time and $O(nN)$ running time. Thus, we have the following lemma.

Lemma 7. Each phase of Algorithm Deterministic-MST takes $O(1)$ awake time and $O(nN)$ running time.

Thus, combining Lemma 6 and Lemma 7, we get the following theorem.

Theorem 2. Algorithm Deterministic-MST is a deterministic algorithm to find the MST of a given graph in $O(\log n)$ awake time and $O(nN \log n)$ running time.

Remark. The coloring procedure, Fast-Awake-Coloring, is the main reason for the large run time. As we noted near the beginning of this section, we can replace this procedure with one that can accomplish this deterministically in $O(\log^* n)$ run time even in the traditional model (see e.g., [22]). However, we suffer an overhead of $O(\log^* n)$ factor in the awake time. As a result, using this modified procedure would allow us to get the following corollary.

Corollary 1. There exists a deterministic algorithm to find the MST of a given graph in $O(\log n \log^* n)$ awake time and $O(n \log n \log^* n)$ run time.
3 Lower Bounds

We now turn our attention to lower bounds. We begin by showing that $\Omega(\log n)$ is indeed an unconditional lower bound on the awake time for MST. This shows that our algorithms presented in Section 2 achieve optimal awake complexity. We then show a lower bound of $\tilde{\Omega}(n)$ on the product of the awake and round complexities. This can be considered as a conditional lower bound on awake complexity, conditioned on an upper bound on the round complexity. This conditional lower bound shows that our randomized awake optimal algorithm (see Section 2.2) has essentially the best possible round complexity (up to a polylog $n$ factor).

3.1 Unconditional Lower bound on Awake Complexity of MST

We leverage the proof technique used in [3] to show that solving MST with probability at least 1/4 requires at least $\Omega(\log n)$ time. They consider a path and show that under any algorithm, every segment of length $13^i$, with probability at least 1/2, has a vertex $v^*$ that has not heard any message from outside that segment even after being awake for $i$ rounds. Their model is based on radio networks, but their claim holds more generally in our model and problem as well provided some subtleties are addressed. Therefore, for the sake of completeness, we are providing the proof in full.

For our purpose, we consider a ring of $\Theta(n)$ nodes with random weights and edges. The two largest weighted edges will be apart by a hop distance of at least $\Omega(n)$ (with constant probability) and any MST algorithm must detect which one has the lower weight. Clearly, this will require communication over either one of the $\Omega(n)$ length paths between the two edges. Under this setting, we get the following theorem.

**Theorem 3.** Any algorithm to solve MST with probability exceeding 1/8 on a ring network requires $\Omega(\log n)$ awake time even when message sizes are unbounded.

3.2 Lower Bound on the Product of Awake and Round Complexity

We adapt the lower bound technique from [7] to the sleeping model and show a lower bound on the product of the awake and round complexities, thereby exposing an inherent trade off between them.

If an edge $e = (u, v)$ must transmit $B$ bits when executing an algorithm in the CONGEST model, then, both $u$ and $v$ must be awake for at least $\Omega(B/\log n)$ rounds. Thus congestion increases awake time. As one might expect, fast algorithms for MST in the CONGEST model are likely to encounter more congestion that translates to higher awake time. Our goal is to exploit this intuition to prove the lower bound.

We use a communication complexity based reduction to reduce set disjointness (SD) in the classical communication complexity model to the distributed set disjointness (DSD) problem in the sleeping model and then extend DSD to the minimum spanning tree problem via an intermediate connected spanning subgraph (CSS), both in the sleeping model.

The SD problem is defined in the classical communication complexity model wherein two players, Alice and Bob, possess two $k$-bit strings $x = (x_i)_{1 \leq i \leq k}$ and $y = (y_i)_{1 \leq i \leq k}$, respectively. They are required to compute an output bit $d(x, y)$ that is 1 iff there is no $i \in [k]$ such that $x_i = y_i = 1$ (i.e., the inner product $\langle x, y \rangle = 0$), and 0 otherwise. Alice and Bob must compute $d(x, y)$ while exchanging the least number of bits. It is well-known that any protocol that solves SD requires $\Omega(k)$ bits (on expectation) to be exchanged between Alice and Bob even if they employ a randomized protocol [26] that can fail with a small fixed probability $\varepsilon > 0$. 

15
The DSD problem is defined in the sleeping model on a graph $G_{rc}$ that is shown in Fig. 1. Let $r$ and $c$ be two positive integers such that $rc + \Theta(\log n) = n$ (the network size) and we are particularly interested in the regime where $c \in \omega(\sqrt{n \log^2 n})$ and $r \in o(\sqrt{n / \log^2 n})$. The graph comprises $r$ rows (or parallel paths) $p_\ell$, $1 \leq \ell \leq r$, with $p_1$ referring to the parallel path at the bottom. Each parallel path comprises $c$ nodes arranged from left to right with the first node referring to the leftmost node and the last node referring to the rightmost node. The first node in $p_1$ and the last node in $p_1$ are designated Alice and Bob because they are controlled by the players Alice and Bob in our reduction. Alice (resp., Bob) is connected to first node (resp., last node) of each $p_\ell$, $2 \leq \ell \leq r$. Additionally, we pick $\Theta(\log n)$ equally spaced nodes $X$ (of cardinality that is a power of two) from $p_1$ such that the first and last nodes in $p_1$ are included in $X$. For each $x \in X$, say at position $j$ in $p_1$, we add edges from $x$ to the $j$th node in each $p_\ell$, $2 \leq \ell \leq r$. Using $X$ as leaves, we construct a balanced binary tree. We will use $I$ to denote the internal nodes of this tree. Alice is in possession of $x$ and Bob is in possession of $y$ and, to solve DSD, they must compute $d(x, y)$ in the sleeping model over the network $G_{rc}$.

In the CSS problem defined in the sleeping model on $G_{rc}$, some edges in $G_{rc}$ are marked and at least one node in the network must determine whether the marked edges form a connected spanning subgraph of $G_{rc}$. For the MST problem, we require edges in $G_{rc}$ to be weighted and the goal is to construct a minimum spanning tree of $G_{rc}$ such that the endpoints of each MST edge $e$ are aware that $e$ is an MST edge.

$G_{rc}$ is constructed such that any node can reach some $x \in X$ within $O(c / \log n)$ steps and any pair of nodes in $X$ are within $O(\log n)$ steps. Recall that $c \in \omega(\sqrt{n \log^2 n})$. Thus,

**Observation 1.** The network graph $G_{rc}$ has diameter $D \in \Theta(c / \log n)$. Moreover, $D \in \omega(\sqrt{n \log^2 n})$. Therefore, DSD, CSS, and MST (if edges in $G_{rc}$ are assigned weights) can be computed in $O(D) =$
Lemma 8. Suppose there is a protocol $P$ in the sleeping model that solves DSD on $G_{rc}$ with $c \in \omega(\sqrt{n} \log^2 n)$ and $r \in o(\sqrt{n}/\log^2 n)$ in $T$ (worst-case) rounds such that $T \in o(c)$ (and we know such an algorithm exists from Observation 1, in particular because $D \in O(c/\log n)$). Then, the node awake time of $P$ must be at least $\Omega(r/\log^2 n)$. This holds even if $P$ is randomized and has an error probability that is bounded by a small constant $\varepsilon > 0$.

DSD $\rightarrow$ CSS We now show a reduction from DSD $\rightarrow$ CSS by encoding a given DSD problem instance as a CSS instance in the following manner. Recall that in DSD, Alice and Bob have bit strings $x$ and $y$, respectively, of length $r$ each. Furthermore, recall that Alice (resp., Bob) is connected to first node (resp., last node) of each $p_{\ell}, 2 \leq \ell \leq r$.

Lemma 9. Suppose there is a protocol $Q$ in the sleeping model that solves CSS on $G_{rc}$ with $c \in \omega(\sqrt{n} \log^2 n)$ and $r \in o(\sqrt{n}/\log^2 n)$ in $T$ rounds such that $T \in o(c)$. Then, the node awake time of $Q$ must be at least $\Omega(r/\log^2 n)$. This holds even if $Q$ is randomized and has an error probability that is bounded by a small constant $\varepsilon > 0$.

CSS $\rightarrow$ MST Recall that CSS requires a subset of the edges in the network graph $G_{rc}$ to be marked; we are to report whether the marked edges form a spanning subgraph of $G_{rc}$. MST on the other hand takes a weighted network graph and computes the minimum spanning tree. A reduction from CSS to MST can be constructed by assigning a weight of 1 for marked edges in the CSS instance and $n$ for all other edges and asking if any edge of weight $n$ is included in the MST. This leads us to the following lemma.

Lemma 10. Suppose there is a protocol $M$ in the sleeping model that solves MST on $G_{rc}$ with $c \in \omega(\sqrt{n} \log^2 n)$ and $r \in o(\sqrt{n}/\log^2 n)$ in $T$ rounds such that $T \in o(c)$. Then, the node awake time of $M$ must be at least $\Omega(r/\log^2 n)$. This holds even if $M$ is randomized and has an error probability that is bounded by a small constant $\varepsilon > 0$.

Thus, putting it all together and generalizing to arbitrary graphs, we can conclude with the following theorem.

Theorem 4. Consider positive integers $r$ and $c$ such that $rc + \Theta(\log n) = n$ (the network size), $c \in \omega(\sqrt{n} \log^2 n)$ and and $r \in o(\sqrt{n}/\log^2 n)$. Suppose there exists a randomized algorithm $M$ for MST in the sleeping model that runs in time $T \in o(c)$ rounds and guaranteed to compute the MST with probability at least $1 - \varepsilon$ for any small fixed $\varepsilon > 0$. Then, the worst case awake complexity of $M$ must be at least $\Omega(r/\log^2 n)$.

4 Conclusion

We presented distributed algorithms for the fundamental MST problem that are optimal with respect to awake complexity. We also showed that there is an inherent trade-off bottleneck between awake and round complexities of MST. In other words, one cannot attain optimal complexities simultaneously under both measures. Several questions arise from our work. First, while our awake-optimal randomized algorithm has the best possible round complexity, our awake-optimal
deterministic algorithm does not. Furthermore, its round complexity depends on the size of the ID range. It will be interesting to design an awake-optimal deterministic algorithm that also has $O(n)$ round complexity which is the best possible due to our trade-off lower bound. Second, it will be interesting to design an MST algorithm that trades-off awake complexity with round complexity. Finally, an important general question is studying other fundamental global problems such as shortest paths and minimum cuts in the sleeping model.

References

[1] Christoph Ambühl. An optimal bound for the MST algorithm to compute energy efficient broadcast trees in wireless networks. In Automata, Languages and Programming, 32nd International Colloquium, ICALP 2005, Lisbon, Portugal, July 11-15, 2005, Proceedings, volume 3580 of Lecture Notes in Computer Science, pages 1139–1150. Springer, 2005.

[2] Leonid Barenboim and Tzalik Maimon. Deterministic logarithmic completeness in the distributed sleeping model. In Seth Gilbert, editor, 35th International Symposium on Distributed Computing, DISC 2021, October 4-8, 2021, Freiburg, Germany (Virtual Conference), volume 209 of LIPIcs, pages 10:1–10:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPIcs.DISC.2021.10.

[3] Yi-Jun Chang, Varsha Dani, Thomas P. Hayes, Qizheng He, Wenzheng Li, and Seth Pettie. The energy complexity of broadcast. In Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing, pages 95–104, 2018.

[4] Yi-Jun Chang, Varsha Dani, Thomas P. Hayes, and Seth Pettie. The energy complexity of bfs in radio networks. In Proceedings of the 39th Symposium on Principles of Distributed Computing, pages 273–282, 2020.

[5] Yi-Jun Chang, Tsvi Kopelowitz, Seth Pettie, Ruosong Wang, and Wei Zhan. Exponential separations in the energy complexity of leader election. ACM Trans. Algorithms, 15(4):49:1–49:31, 2019. Conference version: STOC 2017.

[6] Soumyottam Chatterjee, Robert Gmyr, and Gopal Pandurangan. Sleeping is efficient: MIS in $O(1)$-rounds node-averaged awake complexity. In Yuval Emek and Christian Cachin, editors, PODC ’20: ACM Symposium on Principles of Distributed Computing, pages 99–108, 2020.

[7] Atish Das Sarma, Stephan Holzer, Liah Kor, Amos Korman, Danupon Nanongkai, Gopal Pandurangan, David Peleg, and Roger Wattenhofer. Distributed verification and hardness of distributed approximation. In Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing, STOC ’11, page 363–372, New York, NY, USA, 2011. Association for Computing Machinery.

[8] Atish Das Sarma, Stephan Holzer, Liah Kor, Amos Korman, Danupon Nanongkai, Gopal Pandurangan, David Peleg, and Roger Wattenhofer. Distributed verification and hardness of distributed approximation. SIAM J. Comput., 41(5):1235–1265, 2012.

[9] Michael Elkin. Unconditional lower bounds on the time-approximation tradeoffs for the distributed minimum spanning tree problem. In Proc. of the ACM Symposium on Theory of Computing, pages 331 – 340, 2004.
[10] Michael Elkin. A simple deterministic distributed MST algorithm, with near-optimal time and message complexities. In Proceedings of the 2017 ACM Symposium on Principles of Distributed Computing (PODC), pages 157–163, 2017.

[11] R. Gallager, P. Humblet, and P. Spira. A distributed algorithm for minimum-weight spanning trees. ACM Transactions on Programming Languages and Systems, 5(1):66–77, January 1983.

[12] Juan A. Garay, Shay Kutten, and David Peleg. A sublinear time distributed algorithm for minimum-weight spanning trees. SIAM Journal on Computing, 27(1):302–316, 1998.

[13] Mohsen Ghaffari and Bernhard Haeupler. Distributed algorithms for planar networks II: low-congestion shortcuts, mst, and min-cut. In Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, pages 202–219. SIAM, 2016.

[14] Mohsen Ghaffari and Julian Portmann. Average awake complexity of mis and matching, 2021.

[15] Bernhard Haeupler, David Wajc, and Goran Zuzic. Universally-optimal distributed algorithms for known topologies. In STOC ’21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, pages 1166–1179. ACM, 2021.

[16] Maleq Khan, Gopal Pandurangan, and V. S. Anil Kumar. Distributed algorithms for constructing approximate minimum spanning trees in wireless sensor networks. IEEE Trans. Parallel Distributed Syst., 20(1):124–139, 2009.

[17] Valerie King, Cynthia A. Phillips, Jared Saia, and Maxwell Young. Sleeping on the job: Energy-efficient and robust broadcast for radio networks. Algorithmica, 61(3):518–554, 2011. doi:10.1007/s00453-010-9422-0.

[18] Shay Kutten, Gopal Pandurangan, David Peleg, Peter Robinson, and Amitabh Trehan. On the complexity of universal leader election. J. ACM, 62(1), 2015.

[19] Shay Kutten and David Peleg. Fast distributed construction of small $k$-dominating sets and applications. J. Algorithms, 28(1):40–66, 1998.

[20] Tzalik Maimon. Sleeping model: Local and dynamic algorithms, 2021. arXiv:2112.05344.

[21] Gopal Pandurangan. Distributed network algorithms. In Distributed Network Algorithms, 2021. URL: https://sites.google.com/site/gopalpandurangan/home/distributed-network-algorithms.

[22] Gopal Pandurangan, Peter Robinson, and Michele Scquizzato. A time- and message-optimal distributed algorithm for minimum spanning trees. In Proceedings of the 49th Annual ACM Symposium on the Theory of Computing (STOC), pages 743–756, 2017.

[23] Gopal Pandurangan, Peter Robinson, and Michele Scquizzato. The distributed minimum spanning tree problem. Bull. EATCS, 125, 2018.

[24] David Peleg. Distributed Computing: A Locality Sensitive Approach. SIAM, 2000.

[25] David Peleg and Vitaly Rubinovich. A near-tight lower bound on the time complexity of distributed minimum-weight spanning tree construction. SIAM J. Comput., 30(5):1427–1442, 2000.

[26] Alexander A. Razborov. On the distributional complexity of disjointness. Theor. Comput. Sci., 106(2):385–390, 1992.
A Additional Related Work

The sleeping model and the awake complexity measure was introduced by Chatterjee, Gmyr and Pandurangan [6] who showed that MIS in general graphs can be solved in $O(1)$ rounds node-averaged awake complexity. Node-averaged awake complexity is measured by the average number of rounds a node is awake. The (worst-case) awake complexity of their MIS algorithm is $O(\log n)$, while the worst-case complexity (that includes all rounds, sleeping and awake) is $O(\log^{3.41} n)$ rounds. Subsequently, Ghaffari and Portmann[14] developed a randomized MIS algorithm that has worst-case complexity of $O(\log n)$, while having $O(1)$ node-averaged awake complexity (both bounds hold with high probability).

Barenboim and Maimon [2] showed that many problems, including broadcast, construction of a spanning tree, and leader election can be solved deterministically in $O(\log n)$ awake complexity in the sleeping model. They also showed that fundamental symmetry breaking problems such as MIS and $(\Delta + 1)$-coloring can be solved deterministically in $O(\log \Delta + \log^* n)$ awake rounds in the $\textsf{LOCAL}$ model, where $\Delta$ is the maximum degree. More generally, they also define the class of $O$-$\textsf{LOCAL}$ problems (that includes MIS and coloring) and showed that problems in this class admit a deterministic algorithm that runs in $O(\log \Delta)$ awake time and $O(\Delta^2)$ round complexity. Maimon [20] presents trade-offs between awake and round complexity for O-$\textsf{LOCAL}$ problems.

The sleeping model is very similar to another model called the energy complexity model of Chang, Kopelowitz, Pettie, Wang, and Zhan [5] proposed for radio networks. The energy complexity model is similar to the awake complexity in the sleeping model. In the energy complexity model, a node can be in one of the three modes in a round: idle, sending, or listening. If it is not in the idle mode, then the node is spending energy. Hence rounds spent in the other modes count towards energy complexity. However, energy complexity model has some additional restrictions that pertain to wireless networks. The most important being that nodes can only broadcast messages (hence the same message is sent to all neighbors) and a when a node transmits no other neighboring node can. (Also a node cannot transmit and listen in the same round.) The energy model has a few variants depending on how collisions are handled. There is a version of the energy model called the “Local” where collisions are ignored and nodes can transmit messages at the same time; this is essentially similar to the sleeping model (apart from the notion of broadcast only in the energy complexity model). In particular, upper bounds in the local version apply directly to the sleeping model. Also, our algorithms in the sleeping model can be made to work in the energy complexity model yielding similar bounds (with possibly a $O(\text{polylog}(n))$ multiplicative factor) to the energy complexity.

Also lower bounds shown in the local version of the energy model applies in many instances. For example, Chang et al. [3] show a lower bound $\Omega(\log n)$ on the energy complexity of broadcast which applies also to randomized algorithms. This lower bound is shown for the local version of their model, and this result holds also for the awake complexity in the sleeping model. We adapt this lower bound result to show a $\Omega(\log n)$ lower bound on the awake complexity of MST even for randomized algorithms.

Several global problems such as broadcast, leader election, and breadth-first search have been studied in the energy model [5, 3, 4].

Motivated also by energy-considerations in radio networks, King, Phillips, Saia, and Young [17] present an energy-efficient algorithm in a model similar to the sleeping model to solve a reliable broadcast problem.
B Toolbox of Procedures - In Detail

In the course of developing our algorithms, we make use of the following tools. Notice that for several of the processes described below, it is assumed that the initial graph has already been divided into a FLDT. For each such tree, all nodes within the tree know the ID of the root of the tree (called fragment ID), their parents and children, if any, in the tree, and their distance from the root of that tree.

Transmission schedule of nodes in a tree. We describe the function \(Transmission - Schedule(root, u, n)\), which takes a given node \(u\) and a tree rooted at \(root\) (to which \(u\) belongs) and maps them to a set of rounds in a block of \(2^n + 1\) rounds. \(Transmission - Schedule(root, u, n)\) may be used by node \(u\) to determine which of the \(2^n + 1\) rounds to be awake in. Consider a tree rooted at \(root\) and a node \(u\) in that tree at distance \(i\) from the root. For ease of explanation, we assign names to each of these rounds as well. For all non-root nodes \(u\), the set of rounds that \(Transmission - Schedule(root, u, n)\) maps to includes rounds \(i, i+1, n+1, 2n - i + 1, \text{ and } 2n - i + 2\) with corresponding names \(\text{Down-Receive}, \text{Down-Send}, \text{Side-Send-Receive}, \text{Up-Receive}, \text{and } \text{Up-Send}\), respectively. \(Transmission - Schedule(root, root, n)\) only maps to the set containing rounds \(1, n + 1, \text{ and } 2n + 1\) with names \(\text{Down-Send}, \text{Side-Send-Receive}, \text{and } \text{Up-Receive}\), respectively.

Broadcasting a message in a tree. We describe the procedure \(\text{Fragment-Broadcast}(n)\) that may be run by all nodes in a tree in order for a message \(msg\) to be broadcast from the root of the tree to all nodes in the tree. Consider a block of \(2^n + 1\) rounds and let all nodes \(u\) in the tree rooted at \(root\) utilize \(Transmission - Schedule(root, u, n)\) to decide when to be awake. For each \(u\), in the round that corresponds to \(\text{Down-Receive}\), \(u\) listens for a new message from its parent. In the round that corresponds to \(\text{Down-Send}\), if \(u\) transmits the message it has to its children in the tree.

Observation 2. Procedure \(\text{Fragment-Broadcast}(n)\), run by all nodes in a tree, allows the root of that tree to transmit a message to each node in its tree in \(O(n)\) running time and \(O(1)\) awake time.

Upcasting the minimum value in a tree. We describe the procedure \(\text{Upcast-Min}(n)\) that may be run by all nodes in a tree rooted at \(root\) in order to propagate the minimum among all values held by the nodes of that tree up to \(root\). Consider a block of \(2^n + 1\) rounds and let all nodes \(u\) in the tree rooted at \(root\) utilize \(Transmission - Schedule(root, u, n)\) to decide when to be awake. If \(u\) is a leaf, then in the round that corresponds to \(\text{Up-Send}\), \(u\) transmits its message to its parent. For each \(u\) that is not a leaf, in the round that corresponds to \(\text{Up-Receive}\), \(u\) listens for messages from its children. In the round that corresponds to \(\text{Up-Send}\), \(u\) compares the messages it previously received in its \(\text{Up-Receive}\) round to its current message, if any, and stores the minimum value. It then transmits this minimum value to its parent in the tree.

Observation 3. Procedure \(\text{Upcast-Min}(n)\), run by all nodes in a tree, allows the smallest value among all values held by the nodes, if any, to be propagated to the root of the tree in \(O(n)\) running time and \(O(1)\) awake time.

\(^{11}\)In this description, we assumed that \(Transmission - Schedule(\cdot, \cdot, n)\) was started in round 1. However, if \(Transmission - Schedule(\cdot, \cdot, n)\) is started in round \(r\), then just add \(r - 1\) to the values mentioned here and in the previous sentence.
Transmitting a message between nodes of adjacent trees. We describe the procedure \textsc{Transmit-Adjacent}(n) that may be run by all nodes in a tree rooted at root in order to transmit a message held by nodes in the tree to their neighbors in adjacent trees.\footnote{We assume that all nodes in the graph belong to some tree and are running the algorithm in parallel. We ensure this assumption is valid whenever we call this procedure.} Consider a block of $2n+1$ rounds and let each node $u$ in the tree rooted at root utilize \textit{Transmission-Schedule(root, u, n)} to decide when to be awake. In the round that corresponds to \textit{Side-Send-Receive}, $u$ transmits its message to its neighbors not in the tree. It also listens for messages from those neighbors.

**Observation 4.** Procedure \textsc{Transmit-Adjacent}(n), run by all nodes in a tree, allows each node in a tree to transfer a message, if any, to neighboring nodes belonging to other trees in $O(n)$ running time and $O(1)$ awake time.
C Illustration of Fragments Merging during Procedure Merging-Fragments\(m\).

Figure 2: Initial configuration. Tails fragment has an MOE to a Heads fragment. Numbers in nodes refer to distance from root of that tree.
Figure 3: After \textsc{Transmit-Adjacent}(n) and first instance of \textit{Transmission-Schedule}(\cdot, \cdot, n), nodes from MOE node to root updated distance from root values internally. Fragment ID of Heads fragment propagated to these nodes as well.

Figure 4: After second instance of \textit{Transmission-Schedule}(\cdot, \cdot, n), remaining nodes updated their distance values internally. Fragment ID of Heads fragment propagated to these nodes as well.
Figure 5: Finally, all nodes update their distance from root values (which were stored in temporary variables). Furthermore, edges from MOE node to root are re-oriented to reflect new tree structure.
D Deferred Proofs

Proof of Lemma 1. Denote by $F_i$ the set of fragments that are present at the beginning of phase $i \in [1, P + 1]$. Define $F_i = |F_i|$. We want to show that for all $i \in [1, P]$, $E[F_{i+1}] = 3E[F_i]/4$. We can then leverage this to show a high probability bound on the number of fragments present in the system at the end of phase $P$. It is also easy to see that we only need to prove that the current lemma holds until phase $P$ ends as by definition phase $P$ is the last phase in which there are two or more fragments.

Consider some arbitrary phase $i$, $i \in [1, P]$. At the beginning of phase $i$, there are some $F_i$ fragments present. From the law of total expectation (since $F_i$ and $F_{i+1}$ are random variables from the same probability space), we know that $E[F_{i+1}] = E[E[F_{i+1}|F_i]]$. Thus, we now upper bound $E[F_{i+1}|F_i]$.

In order to count the number of fragments that are left after the end of a phase, we first define the notion of a given fragment “surviving” a given phase. Consider one of the following by Markov’s inequality:

$$E_i = \text{bound on phase if it does not flip tails and merge into a heads fragment. Thus, the probability that } f_i \text{ survives phase } i \text{ is } \leq \frac{3}{4}. \text{ Let } X_{f_i} \text{ be an indicator random variable that takes value 1 if fragment } f_i \text{ survives phase } i \text{ and 0 otherwise. It is easy to see that the expected number of fragments that survive phase } i \text{ acts as an upper bound on } E[F_{i+1}|F_i].$$

Now, we can calculate the value of $E[F_{i+1}]$ as follows:

$$E[F_{i+1}] = E[E[F_{i+1}|F_i]] \leq E \left[ \sum_{f \in F_i} X_{f} \right] \leq \sum_{f \in F_i} E \left[ X_{f} \right] \leq \sum_{f \in F_i} Pr(f \text{ survives phase } i) \leq 3F_i/4.$$

Recall that the number of fragments at the beginning of phase 1 is equal to $n$. Thus, we have $E[F_{P+1}] \leq E[F_P]/(4/3) \leq E[F_1]/(4/3)^P \leq n/(4/3)^P$. So we see that after $P$ phases, we have the following by Markov’s inequality:

$$Pr(F_{P+1} > 1) \leq E[F_{P+1}]/1 \leq n/(4/3)^P.$$

By setting $P$ to $4[\log_{4/3} n]$, we see that $Pr(F_{P+1} > 1) \leq 1/n^3$. Thus with high probability, the number of fragments at the beginning of phase $P + 1$ does not exceed 1.

Proof of Lemma 2. By Lemma 1, we see that in phases $i \in [1, P]$, the number of fragments reduce by a constant factor until only one is left with high probability. From phase $P$ onwards, there is at most one fragment with high probability and the number of fragments never increases in a given phase. Thus, since we run the algorithm for $P + 1$ phases and the initial graph is connected, we see that at the end of the algorithm, there exists only one fragment. Furthermore, it is easy to see from the algorithm that each node in the initial graph belongs to this fragment and every edge in this fragment was once the minimum outgoing edge from some fragment to another. In other words, the set of nodes and edges within the final fragment represent the MST of the initial graph.

Proof of Lemma 3. Recall that there are $O(\log n)$ phases of the algorithm. We show that in each phase, each node experiences awake time of $O(1)$ and running time of $O(n)$, thus giving us the desired bounds. In a given phase, a given node may run the following procedures a constant number of times: Fragment-Broadcast($n$), Upcast-Min($n$), and Transmit-Adjacent($n$). Each of these procedures takes $O(1)$ awake time and $O(n)$ running time. Additionally, to complete
Proof of Lemma 4. Let $B$, $R$, $O$, $Bl$, and $G$ represent the number of Blue, Red, Orange, Black, and Green fragments in $H'$, respectively. We can get a lower bound on the number Blue fragments in $H'$ by assuming that the adversary caused the maximum number of fragments to be colored other colors.

For a fragment to be colored Red, at least one of its neighbors must be colored Blue. Furthermore, any given Blue fragment can be a neighbor to at most 4 Red fragments. Thus, the number of Red fragments is upper bounded by $4B$. Similar arguments show that the number of Orange fragments, Black fragments, and Green fragments are upper bounded by $4^2B$, $4^3B$, and $4^4B$, respectively.

Thus, the number of fragments in $H'$, $|H'| = B + R + O + Bl + G \leq B + 4B + 4^2B + 4^3B + 4^4B \leq B(4^5 - 1)/3 \leq 342B$. Thus, we see that $B \geq |H'|/342$.

Proof of Theorem 3. Consider a weighted ring $R$ of length $4n + 4$ with each node $u$ having a random ID denoted $ID(u)$ and every edge $(u, v)$ having a random weight $w(u, v)$, with all IDs and weights being drawn uniformly and independently from a sufficiently large poly$(n)$ space. The IDs and weights will be distinct whp, so we condition on this for the rest of the proof.

Let $\mathcal{P}$ be any protocol executed by the nodes in the ring. For the purpose of this lower bound, we assume that $\mathcal{P}$ operates as follows. Initially, every node $u$ generates a sufficiently large random string $r_u$ and uses $r_u$ to make all its random choices throughout the execution of $\mathcal{P}$. Every pair of neighboring vertices $u$ and $v$ that are simultaneously awake in some round $r$ communicate their states to each other in their entireties. Thus, each node $u$ starts with an initial knowledge comprising just its own ID\(^{13}\), weights of incident edges and its own random string. More generally, at the end of its $a$th awake round, $u$ knows everything about a segment $S(u, a)$ (including weights of edges incident on end vertices of $S(u, a)$) but nothing outside it.

Let $I_k$ denote the family of connected segments of length $k$. For each $I \in I_k$, we define an event $U(I, a)$ as follows: there exists a special vertex $v^* \in I$ such that $S(v^*, a) \subseteq I$. We now show a crucial lemma that will help us complete the proof.

Lemma 11. Let $a_{\text{max}} = \lfloor \log_{13} n \rfloor$. For every $a \in \{0, 1, \ldots, a_{\text{max}}\} = \lfloor a_{\text{max}} \rfloor$ and $k = 13^a$ and $I \in I_k$, $\text{Pr}(U(I, a)) \geq 1/2$. Moreover, for any two non-overlapping segments $I_1$ and $I_2$ (i.e., with no common vertex), the events $U(I_1, a)$ and $U(I_2, a)$ are independent.

Proof. We prove this by induction on $a$. Clearly, $U(I, 0)$ holds for every $I \in I_1$, thus establishing the basis. Now consider any $a > 0$ and any $I \in I_k$ for $k = 13^a$. We split $I$ into 13 non-overlapping segments of length $13^{a-1}$ each. The probability that $U(*, a-1)$ will occur for five of the 13 segments is at least $5/6$, thanks to the independence of $U(*, a-1)$. Now let us focus on the five of the 13 segments (denoted $A$, $B$, $C$, $D$, and $E$ taken, say, in clockwise order) for which the event $U(*, a-1)$ occurred with $v(A)$, $v(B)$, $v(C)$, $v(D)$, and $v(E)$ denoting the special vertex. Again, due to independence, the time when each $v(*)$ wakes up for the $a$th time are independent. Thus, with probability at least $3/5$, $v(B)$, $v(C)$, or $v(D)$ will wake up for their respective $a$th time no later than $v(A)$ or $v(E)$. Therefore, with probability at least $(3/5)(5/6) = 1/2$, one of $v(B)$, $v(C)$, or $v(D)$ would become the special vertex pertaining to $U(I, a)$, thereby ensuring that $\text{Pr}(U(I, a) \geq 1/2)$.

\(^{13}\)Normally, we assume $KT_1$ knowledge (i.e., each node knows the IDs of its neighbors) at the beginning, but here, we assume $KT_0$ (i.e., no such knowledge) without loss of generality, because we can always add a single awake round for all nodes at the beginning to go from $KT_0$ to $KT_1$. 

27
Let \( e_1 = (u_1, v_1) \) and \( e_2 = (u_2, v_2) \) be the two edges with the largest weights. With probability at least \( 1/2 \), they are separated by a distance of \( n + 1 \). The MST for this ring comprises all edges other than either \( e_1 \) or \( e_2 \), whichever is weighted more. Thus, any MST algorithm for this ring must be able to decide between the two edges. This requires communication either between \( u_2 \) and \( v_1 \) or between \( v_2 \) and \( u_1 \). This will require (with probability at least \( 1/4 \)) \( \Omega(\log n) \) awake time for at least one node along either paths.

**Proof of Lemma 9.** To construct the required instance of DSD we mark the edges of all \( r \) parallel paths and all the tree edges. Furthermore, for each \( \ell \) such that \( x_\ell = 0 \) (resp., \( y_\ell = 0 \)), we mark the edge from Alice (resp., Bob) to the first node (resp., last node) in \( p_\ell \). Now, suppose there exists a \( Q \) that only requires \( o(r/\log^2 n) \) node awake time. Then, \( Q \) can be used to obtain a protocol \( P \) for DSD that runs in \( T \) time and also only requires \( o(r/\log^2 n) \) node awake time, which contradicts Lemma 8.

**Proof of Lemma 8.** Suppose for the sake of contradiction \( P \) runs in time \( T \) and has an awake time complexity of \( o(r/\log^2 n) \). Then, we can show that Alice and Bob can simulate \( P \) in the classical communication complexity model and solve SD on \( r \) bits by exchanging only \( o(r) \) bits which will yield a contradiction to the SD lower bound. We establish this by showing that Alice and Bob can simulate \( P \) to solve SD in the classical communication complexity model.

We show this simulation from Alice’s perspective. Bob’s perspective will be symmetric. Recall that \( p_\ell \) is the \( \ell \)th parallel path. Let \( p_\ell^j \), \( 1 \leq j \leq c \), denote the first \( j \) vertices of path \( p_\ell \). We define \( R_j \) to be the union of all \( p_\ell^j \) and \( I \) (recall that \( I \) is the set of the internal nodes of the binary tree), i.e.,

\[
R_j = \left( \bigcup_{\ell=1}^{r} p_\ell^j \right) \cup I.
\]

Note that \( R_j \) induces a cut \((R_j, \bar{R}_j)\) that is shown in Fig. 1. Alice begins by simulating \( R_{c-1} \) in round 1 as she knows the state of all nodes in \( R_{c-1} \). At each subsequent round \( t \), Alice simulates \( R_{c-t} \). Initially, all the information needed for the simulation is available for Alice because the structure of \( G_{rc} \) is fully known (except for Bob’s input).

As the simulation progresses, in each round \( t > 1 \), \( 1 \leq t \leq T \in o(c) \), all inputs will be available except for the new bits that may enter \( I \) through nodes in \( \bar{R}_{c-t} \). Alice will not need to ask for the bits needed by \( p_\ell^{c-t+1} \) because she simulated all nodes in \( p_\ell^{c-t+1} \), \( 1 \leq \ell \leq r \), in the previous round. Note that the portion simulated by Bob will encompass the portion from which Alice may need bits from Bob, so Bob will indeed have the bits requested by Alice. In order to continue the simulation, Alice borrows bits that \( P \) transmitted from \( R_{c-t} \) to \( I \cap R_{c-t} \) from Bob. Suppose during the course of the simulation in the communication complexity model, \( B \) bits are borrowed from Bob. Then nodes in \( I \) must have been awake for a collective total of at least \( \Omega(B/\log n) \) rounds (because each message of \( O(\log n) \) bits must be received by a node that is awake in \( P \)).\(^{14}\) This implies that at least one node in \( I \) must have been awake for \( \Omega(B/\log^2 n) \) rounds because \( |I| \in O(\log n) \) and the nodes in \( I \) are of constant degree.

Since the node awake time is \( o(r/\log^2 n) \) for \( P \), \( B \) must be \( o(r) \). But this contradicts the fact that SD requires \( \Omega(r) \) bits in the communication complexity model.

\(^{14}\)Note that all the \( B \) bits cannot solely come through a row path of length \( c \), since we are restricting \( T \in o(c) \). In other words, each of the bits has to go through at least one node in \( I \).