

Fault-tolerant Algorithms for Tick-Generation in Asynchronous Logic: Robust Pulse Generation

Danny Dolev, Matthias Függer, Christoph Lenzen, and Ulrich Schmid

Abstract

Today's hardware technology presents a new challenge in designing robust systems. Deep submicron VLSI technology introduced transient and permanent faults that were never considered in low-level system designs in the past. Still, robustness of that part of the system is crucial and needs to be guaranteed for any successful product. Distributed systems, on the other hand, have been dealing with similar issues for decades. However, neither the basic abstractions nor the complexity of contemporary fault-tolerant distributed algorithms match the peculiarities of hardware implementations.

This paper is intended to be part of an attempt striving to overcome this gap between theory and practice for the clock synchronization problem. Solving this task sufficiently well will allow to build a very robust high-precision clocking system for hardware designs like systems-on-chips in critical applications. As our first building block, we describe and prove correct a novel Byzantine fault-tolerant self-stabilizing pulse synchronization protocol, which can be implemented using standard asynchronous digital logic. Despite the strict limitations introduced by hardware designs, it offers optimal resilience and smaller complexity than all existing protocols.

1 Introduction & Related Work

With today’s deep submicron technology running at GHz clock speeds [20], disseminating the high-speed clock throughout a *very large scale integrated* (VLSI) circuit, with negligible skew, is difficult and costly [2, 3, 12, 24, 29]. Systems-on-chip are hence increasingly designed *globally asynchronous locally synchronous* (GALS) [4], where different parts of the chip use different local clock signals. Two main types of clocking schemes for GALS systems exist, namely, (i) those where the local clock signals are unrelated, and (ii) multi-synchronous ones that provide a certain degree of synchrony between local clock signals [30, 34].

GALS systems clocked by type (i) permanently bear the risk of *metastable upsets* when conveying information from one clock domain to another. To explain the issue, consider a physical implementation of a bistable storage element, like a register cell, which can be accessed by read and write operations concurrently. It can be shown that two operations (like two writes with different values) occurring very closely to each other can cause the storage cell to attain neither of its two stable states for an unbounded time [23], and thereby, during an unbounded time afterwards, successive reads may return none of the stable states. Although the probability of a single upset is very small, one has to take into account that every bit of transmitted information across clock domains is a candidate for an upset. Elaborate synchronizers [8, 21, 28] are the only means for achieving an acceptably low probability for metastable upsets here.

This problem can be circumvented in clocking schemes of type (ii): Common synchrony properties offered by multi-synchronous clocking systems are:

- *bounded skew*, i.e., bounded maximum time between the occurrence of any two matching clock transitions of any two local clock signals. Thereby, in classic clock synchronization, two clock transitions are matching iff they are both the k^{th} , $k \geq 1$, clock transition of a local clock.
- *bounded accuracy*, i.e., bounded minimum and maximum time between the occurrence of any two successive clock transitions of any local clock signal.

Type (ii) clocking schemes are particularly beneficial from a designer’s point of view, since they combine the convenient local synchrony of a GALS system with a global time base across the whole chip. It has been shown in [27] that these properties indeed facilitate metastability-free high-speed communication across clock domains.

The decreasing structure sizes of deep submicron technology also resulted in an increased likelihood of chip components failing during operation: Reduced voltage swing and smaller critical charges make circuits more susceptible to ionized particle hits, crosstalk, and electromagnetic interference [5, 18]. *Fault-tolerance* hence becomes an increasingly pressing issue in chip design. Unfortunately, faulty components may behave non-benign in many ways. They may perform signal transitions at arbitrary times and even convey inconsistent information to their successor components if their outgoing communication channels are affected by a failure. This forces to model faulty components as unrestricted, i.e., Byzantine, if a high fault coverage is to be guaranteed.

The DARTS fault-tolerant clock generation approach [15, 17] developed by some of the authors of this paper is a Byzantine fault-tolerant multi-synchronous clocking scheme. DARTS comprises a set of modules, each of which generates a local clock signal for a single clock domain. The DARTS modules (nodes) are synchronized to each other to within a few clock cycles. This is achieved by exchanging binary clock signals only, via single wires. The basic idea behind DARTS is to employ a simple fault-tolerant distributed algorithm [35]—based on Srikanth & Toueg’s consistent

broadcasting primitive [31]—implemented in asynchronous digital logic. An important property of the DARTS clocking scheme is that it guarantees that no metastable upsets occur during fault-free executions. For executions with faults, metastable upsets cannot be ruled out: Since Byzantine faulty components are allowed to issue unrelated read and write accesses by definition, the same arguments as for clocking schemes of type (i) apply. However, in [13], it was shown that by proper chip design the probability of a Byzantine component leading to a metastable upset of DARTS can be made arbitrarily small.

Although both theoretical analysis and experimental evaluation revealed many attractive additional features of DARTS, like guaranteed startup, automatic adaption to current operating conditions, etc., there is room for improvement. The most obvious drawback of DARTS is its inability to support late joining and restarting of nodes, and, more generally, its lack of self-stabilization properties. If, for some reasons, more than a third of the DARTS nodes ever become faulty, the system cannot be guaranteed to resume normal operation even if all failures cease. Even worse, simple transient faults such as radiation- or crosstalk-induced additional (or omitted) clock ticks accumulate over time to arbitrarily large skews in an otherwise benign execution.

Byzantine-tolerant self-stabilization, on the other hand, is the major strength of a number of protocols [1, 6, 9, 19, 22] primarily devised for distributed systems. Of particular interest in the above context is the work on self-stabilizing *pulse synchronization*, where the purpose is to generate well-separated anonymous pulses that are synchronized at all correct nodes. This facilitates self-stabilizing clock synchronization, as agreement on a time window permits to simulate a synchronous protocol in a bounded-delay system. Beyond optimal (i.e., $\lceil n/3 \rceil - 1$, c.f. [26]) resilience, an attractive feature of these protocols is a small stabilization time [1, 6, 19, 22], which is crucial for applications with stringent availability requirements. In particular, [1] synchronizes clocks in expected constant time in a synchronous system. Given any pulse synchronization protocol stabilizing in a bounded-delay system in expected time T , this implies an expected $(T + \mathcal{O}(1))$ -stabilizing clock synchronization protocol.

Nonetheless, it remains open whether a (with respect to the number of nodes n) sublinear convergence time can be achieved: While the classical consensus lower bound of $f + 1$ rounds for synchronous, deterministic algorithms in a system with $f < n/3$ faults [11] proves that *exact* agreement on a clock value requires at least $f + 1 \in \Omega(n)$ deterministic rounds, one has to face the fact that only approximate agreement on the current time is achievable in a bounded-delay system anyway. However, no non-trivial lower bounds on approximate deterministic synchronization or the exact problem with randomization are known by now.

Note that existing synchronization algorithms, in particular those that do not rely on pulse synchronization, have deficiencies rendering them unsuitable in our context. For example, they have exponential convergence time [9], require the relative drift of the nodes' local clocks to be very small [7, 22],¹ provide larger skew only [22] or make use of linear-sized messages [6]. Furthermore, standard models used by the distributed systems community do not account for metastability, resulting in the same to be true for the existing solutions.

It is hence natural to explore ways of combining and extending the above lines of research. The present paper is the first step towards this goal.

Detailed contributions. We describe and prove correct the novel FATAL pulse synchroniza-

¹Note that it is too costly and space consuming to equip each node with a quartz oscillator. Simple digital oscillators, like inverters with feedback, in turn exhibit drifts of at least several percent, which heavily vary with operating conditions.

tion protocol, which facilitates a direct implementation in standard asynchronous digital logic. It self-stabilizes within $\mathcal{O}(n)$ time with probability $1 - 2^{n-f}$,² in the presence of up to $\lceil n/3 \rceil - 1$ Byzantine faulty nodes, and is metastability-free by construction after stabilization in failure-free runs. While executing the protocol, non-faulty nodes broadcast a constant number of bits in constant time. In terms of distributed message complexity, this implies that stabilization is achieved after broadcasting $\mathcal{O}(n)$ messages of size $\mathcal{O}(1)$, improving by factor $\Omega(n)$ on the number of bits transmitted by previous algorithms.³ The protocol can sustain large relative clock drifts of more than 10%, which is crucial if the local clock sources are simple ring oscillators (uncompensated ring oscillators suffer from clock drifts of up to 9% [32]). If the number of faults is not overwhelming, i.e., a majority of at least $n - f$ nodes continues to execute the protocol in an orderly fashion, recovering nodes and late joiners (re)synchronize in constant time. This property is highly desirable in practical systems, in particular in combination with Byzantine fault-tolerance: Even if nodes randomly experience transient faults on a regular basis, quick recovery ensures that the mean time until failure of the system as a whole is substantially increased. All this is achieved against a powerful adversary that, at time t , knows the whole history of the system up to time $t + \varepsilon$ (where $\varepsilon > 0$ is infinitesimally small) and does not need to choose the set of faulty nodes in advance. Apart from bounded drifts and communication delays, our solution solely requires that receivers can unambiguously identify the sender of a message, which is a property that arises naturally in hardware designs.

We also describe how the pulse synchronization protocol can be implemented using asynchronous digital logic. Moreover, we sketch how the pulse synchronization protocol will be integrated with DARTS clocks to build a high-precision self-stabilizing clocking system for multi-synchronous GALS. The basic idea of our integration is to let the pulse synchronization protocol non-intrusively monitor the operation of DARTS clocks and to recover DARTS clocks that run abnormally. Like the original DARTS, the joint system is metastability-free in failure-free runs after stabilization. During stabilization, the fact that nodes merely undergo a constant number of state transitions in constant time ensures a very small probability of metastable upsets.

2 Model

Our formal framework will be tied to the peculiarities of hardware designs, which consist of modules that *continuously*⁴ compute their output signals based on their input signals. Following [14, 16], we define (the trace of) a *signal* to be a timed event trace over a finite alphabet \mathbb{S} of possible signal states: Formally, signal $\sigma \subseteq \mathbb{S} \times \mathbb{R}_0^+$. All times and time intervals refer to a global *reference time* taken from \mathbb{R}_0^+ , that is, signals describe the system's behaviour from time 0 on. The elements of σ are called *events*, and for each event (s, t) we call s the *state of event* (s, t) and t the *time of event* (s, t) . In general, a signal σ is required to fulfill the following conditions: (i) for each time interval $[t^-, t^+] \subseteq \mathbb{R}_0^+$ of finite length, the number of events in σ with times within $[t^-, t^+]$ is finite, (ii) from $(s, t) \in \sigma$ and $(s', t) \in \sigma$ follows that $s = s'$, and (iii) there exists an event at time 0 in σ .

²Note that the algorithm from [1] achieving an expected constant stabilization time in a synchronous model needs to run for $\Omega(n)$ rounds to ensure the same probability of stabilization.

³We remark that [22] achieves the same complexity, but considers a much simpler model. In particular, *all* communication is restricted to broadcasts, i.e., all nodes observe the same behaviour of a given other node, even if it is faulty.

⁴In sharp contrast to classic distributed computing models, there is no computationally complex discrete zero-time state-transition here.

Note that our definition allows for events (s, t) and $(s, t') \in \sigma$, where $t < t'$, without having an event $(s', t'') \in \sigma$ with $s' \neq s$ and $t < t'' < t'$. In this case, we call event (s, t') *idempotent*. Two signals σ and σ' are *equivalent*, iff they differ in idempotent events only. We identify all signals of an equivalence class, as they describe the same physical signal. Each equivalence class $[\sigma]$ of signals contains a unique signal σ_0 having no idempotent events. We say that *signal σ switches to s at time t* iff event $(s, t) \in \sigma_0$.

The *state of signal σ* at time $t \in \mathbb{R}_0^+$, denoted by $\sigma(t)$, is given by the state of the event with the maximum time not greater than t .⁵ Because of (i), (ii) and (iii), $\sigma(t)$ is well defined for each time $t \in \mathbb{R}_0^+$. Note that σ 's state function in fact depends on $[\sigma]$ only, i.e., we may add or remove idempotent events at will without changing the state function.

Distributed System On the topmost level of abstraction, we see the system as a set of $V = \{1, \dots, n\}$ physically remote *nodes* that communicate by means of *channels*. In the context of a VLSI circuit, “physically remote” actually refers to quite small distances (centimeters or even less). However, at gigahertz frequencies, a local state transition will not be observed remotely within a time that is negligible compared to clock speeds. We stress this point, since it is crucial that different clocks (and their attached logic) are not too close to each other, as otherwise they might fail due to the same event such as a particle hit. This would render it pointless to devise a system that is resilient to a certain fraction of the nodes failing.

Each node i comprises a number of *input ports*, namely $S_{i,j}$ for each node j , an *output port* S_i , and a set of *local ports*, introduced later on. An *execution* of the distributed system assigns to each port of each node a signal. For convenience of notation, for any port p , we refer to the signal assigned to port p simply by signal p . We say that *node i is in state s at time t* iff $S_i(t) = s$. We further say that *node i switches to state s at time t* iff signal S_i switches to s at time t .

Nodes exchange their states via the channels between them: for each pair of nodes i, j , output port S_i is connected to input port $S_{j,i}$ by a FIFO channel from i to j . Note that this includes a channel from i to i itself. Intuitively, S_i being connected to $S_{j,i}$ by a (non-faulty) channel means that $S_{j,i}(\cdot)$ should mimic $S_i(\cdot)$, however, with a slight delay accounting for the time it takes the signal to propagate. In contrast to an asynchronous system, this delay is bounded by the *maximum delay* $d > 0$.⁶

Formally we define: The *channel* from node i to j is said to be *correct* during $[t^-, t^+]$ iff there exists a function $\tau_{i,j} : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$, called the channel's *delay function*, such that: (i) $\tau_{i,j}$ is continuous and strictly increasing, (ii) $\forall t \in [t^-, t^+] : 0 \leq \tau_{i,j}(t) - t < d$, and (iii) for each $t \in [t^-, t^+]$, $(s, \tau_{i,j}(t)) \in S_{j,i} \Leftrightarrow (s, t) \in S_i$. We say that *node i observes node j in state s at time t* if $S_{i,j}(t) = s$.

Clocks and Timeouts Nodes are never aware of the current reference time and we also do not require the reference time to resemble Newtonian “real” time. Rather we allow for physical clocks that run arbitrarily fast or slow, as long as their speeds are close to each other in comparison. One may hence think of the reference time as progressing at the speed of the currently slowest correct clock. In this framework, nodes essentially make use of bounded clocks with bounded drift.

Formally, clock rates are within $[1, \vartheta]$ (with respect to reference time), where $\vartheta > 1$ is constant and $\vartheta - 1$ is the (*maximum*) *clock drift*. A *clock C* is a continuous, strictly increasing function

⁵To facilitate intuition, we here slightly abuse notation, as this way σ denotes both a function of time and the signal (trace), which is a subset of $\mathbb{S} \times \mathbb{R}_0^+$. Whenever referring to σ , we will talk of the signal, not the state function.

⁶With respect to \mathcal{O} -notation, we normalize $d \in \mathcal{O}(1)$, as all time bounds simply depend linearly on d .

$C : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ mapping reference time to some local time. Clock C is said to be *correct* during $[t^-, t^+] \subseteq \mathbb{R}_0^+$ iff we have for any $t, t' \in [t^-, t^+]$, $t < t'$, that $t' - t \leq C(t') - C(t) \leq \vartheta(t' - t)$. Each node comprises a set of clocks assigned to it, which allow the node to estimate the progress of reference time.

Instead of directly accessing the value of their clocks, nodes have access to so-called *timeout ports* of watchdog timers. A *timeout* is a triple (T, s, C) , where $T \in \mathbb{R}^+$ is a duration, $s \in \mathbb{S}$ is a state, and C is a clock, say of node i . Each timeout (T, s, C) has a corresponding timeout port $\text{Time}_{T,s,C}$, being part of node i 's local ports. Signal $\text{Time}_{T,s,C}$ is Boolean, that is, its possible states are from the set $\{0, 1\}$. We say that timeout (T, s, C) is *correct* during $[t^-, t^+] \subseteq \mathbb{R}_0^+$ iff clock C is correct during $[t^-, t^+]$ and the following holds:

1. For each time $t_s \in [t^-, t^+]$ when node i switches to state s , there is a time $t \in [t_s, \tau_{i,i}(t_s)]$ such that (T, s, C) is *reset*, i.e., $(0, t) \in \text{Time}_{T,s,C}$. This is a one-to-one correspondence, i.e., (T, s, C) is not reset at any other times.
2. For a time $t \in [t^-, t^+]$, denote by t_0 the supremum of all times from $[t^-, t]$ when (T, s, C) is reset. Then it holds that $(1, t) \in \text{Time}_{T,s,C}$ iff $C(t) - C(t_0) = T$. Again, this is a one-to-one correspondence.

We say that timeout (T, s, C) *expires* at time t iff $\text{Time}_{T,s,C}$ switches to 1 at time t , and it *is expired* at time t iff $\text{Time}_{T,s,C}(t) = 1$. For notational convenience, we will omit the clock C and simply write (T, s) for both the timeout and its signal.

A *randomized timeout* is a triple (\mathcal{D}, s, C) , where \mathcal{D} is a bounded random distribution on \mathbb{R}_0^+ , $s \in \mathbb{S}$ is a state, and C is a clock. Its corresponding timeout port $\text{Time}_{\mathcal{D},s,C}$ behaves very similar to the one of an ordinary timeout, except that whenever it is reset, the local time that passes until it expires next—provided that it is not reset again before that happens—follows the distribution \mathcal{D} . Formally, (\mathcal{D}, s, C) is correct during $[t^-, t^+] \subseteq \mathbb{R}_0^+$, if C is correct during $[t^-, t^+]$ and the following holds:

1. For each time $t_s \in [t^-, t^+]$ when node i switches to state s , there is a time $t \in [t_s, \tau_{i,i}(t_s)]$ such that (\mathcal{D}, s, C) is *reset*, i.e., $(0, t) \in \text{Time}_{\mathcal{D},s,C}$. This is a one-to-one correspondence, i.e., (\mathcal{D}, s, C) is not reset at any other times.
2. For a time $t \in [t^-, t^+]$, denote by t_0 the supremum of all times from $[t^-, t]$ when (\mathcal{D}, s, C) is reset. Let $\mu : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ denote the density of \mathcal{D} . Then $(1, t) \in \text{Time}_{\mathcal{D},s,C}$ “with probability $\mu(C(t) - C(t_0))$ ” and we require that the probability of $(1, t) \in \text{Time}_{\mathcal{D},s,C}$ —conditional to t_0 and C on $[t_0, t]$ being given—is independent of the system’s state at times smaller than t . More precisely, if superscript \mathcal{E} identifies variables in execution \mathcal{E} and t'_0 is the infimum of all times from $(t_0, t^+]$ when node i switches to state s , then we demand for any $[\tau^-, \tau^+] \subseteq [t_0, t'_0]$ that

$$P \left[\exists t' \in [\tau^-, \tau^+] : (1, t') \in \text{Time}_{\mathcal{D},s,C} \mid t_0^{\mathcal{E}} = t_0 \wedge C|_{[t_0, t']}^{\mathcal{E}} = C|_{[t_0, t']} \right] = \int_{\tau^-}^{\tau^+} \mu(C(\tau) - C(t_0)) d\tau,$$

independently of $\mathcal{E}|_{[0, \tau^-)}$.

We will apply the same notational conventions to randomized timeouts as we do for regular timeouts.

Note that, strictly speaking, this definition does not induce a random variable describing the time $t' \in [t_0, t'_0)$ satisfying that $(1, t') \in \text{Time}_{\mathcal{D}, s, C}$. However, for the state of the timeout port, we get the meaningful statement that for any $t' \in [t_0, t'_0)$,

$$P[\text{Time}_{\mathcal{D}, s, C} \text{ switches to 1 during } [t_0, t']] = \int_{t_0}^{t'} \mu(C(t') - C(t_0)) \, d\tau.$$

The reason for phrasing the definition in the above more cumbersome way is that we want to guarantee that an adversary knowing the full present state of the system and memorizing its whole history cannot reliably predict when the timeout will expire.⁷

We remark that these definitions allow for different timeouts to be driven by the same clock, implying that an adversary may derive some information on the state of a randomized timeout before it expires from the node's behaviour, even if it cannot directly access the values of the clock driving the timeout. This is crucial for implementability, as it might be very difficult to guarantee that the behaviour of a dedicated clock that drives a randomized timeout is indeed independent of the execution of the algorithm.

Memory Flags Besides timeout and randomized timeout ports, another kind of node i 's local ports are *memory flags*. For each state $s \in \mathbb{S}$ and each node $j \in V$, $\text{Mem}_{i,j,s}$ is a local port of node i . It is used to memorize whether node i has observed node j in state s since the last reset of the flag. We say that node i *memorizes node j in state s* at time t if $\text{Mem}_{i,j,s}(t) = 1$. Formally, we require that signal $\text{Mem}_{i,j,s}$ switches to 1 at time t iff node i observes node j in state s at time t and $\text{Mem}_{i,j,s}$ is not already in state 1. The times t when $\text{Mem}_{i,j,s}$ is *reset*, i.e., $(0, t) \in \text{Mem}_{i,j,s}$, are specified by node i 's state machine, which is introduced next.

State Machine It remains to specify how nodes switch states and when they reset memory flags. We do this by means of state machines that may attain states from the finite alphabet \mathbb{S} . A node's state machine is specified by (i) the set \mathbb{S} , (ii) a function tr , called the *transition function*, from $\mathcal{T} \subseteq \mathbb{S}^2$ to the set of Boolean predicates on the alphabet consisting of expressions " $p = s$ " (used for expressing guards), where p is from the node's input and local ports and s is from the set of possible states of signal p , and (iii) a function re , called the *reset function*, from \mathcal{T} to the power set of the node's memory flags.

Intuitively, the transition function specifies the conditions (guards) under which a node switches states, and the reset function determines which memory flags to reset upon the state change. Formally, let P be a predicate on node i 's input and local ports. We define P *holds at time t* by structural induction: If P is equal to $p = s$, where p is one of node i 's input and local ports and s is one of the states signal p can obtain, then P *holds at time t* iff $p(t) = s$. Otherwise, if P is of the form $\neg P_1$, $P_1 \wedge P_2$, or $P_1 \vee P_2$, we define P *holds at time t* in the straightforward manner.

We say node i *follows its state machine during $[t^-, t^+]$* iff the following holds: Assume node i observes itself in state $s \in \mathbb{S}$ at time $t \in [t^-, t^+]$, i.e., $S_{i,i}(t) = s$. Then, for each $(s, s') \in \mathcal{T}$, both:

1. Node i switches to state s' at time t iff $tr(s, s')$ holds at time t and i is not already in state s' .⁸

⁷This is a non-trivial property. For instance nodes could just determine, by drawing from the desired random distribution at time t_0 , at which local clock value the timeout shall expire next. This would, however, essentially give away early when the timeout will expire, greatly reducing the power of randomization!

⁸In case more than one guard $tr(s, s')$ can be true at the same time, we assume that an arbitrary tie-breaking ordering exists among the transition guards that specifies to which state to switch.

2. Node i resets memory flag m at some time in the interval $[t, \tau_{i,i}(t)]$ iff $m \in re(s, s')$ and i switches from state s to state s' at time t . This correspondence is one-to-one.

A node is defined to be *non-faulty* during $[t^-, t^+]$ iff during $[t^-, t^+]$ all its timeouts and randomized timeouts are correct and it follows its state machine. If it employs multiple state machines (see below), it needs to follow all of them.

In contrast, a faulty node may change states arbitrarily. Note that while a faulty node may be forced to send consistent output state signals to all other nodes if its channels remain correct, there is no way to guarantee that this still holds true if channels are faulty.⁹

Metastability In our discrete system model, the effect of metastability is captured by the lacking capability of state machines to instantaneously take on new states: Node i decides on state transitions based on the delayed status of port $S_{i,i}$ instead of its “true” current state S_i . This non-zero delay from S_i to $S_{i,i}$ bears the potential for metastability, as a successful state transition can only be guaranteed if after a transition guard from some state s to some state s' becomes true, all other transition guards from s to $s'' \neq s'$ remain false during this delay at least.

This is exemplified in the following scenario: Assume node i is in state s at some time t . However, since it switched to s only very recently, it still observes itself in state $s' \neq s$ at time t via $S_{i,i}$. Given that there is a transition (s', s'') in \mathcal{T} , $s'' \neq s$, whose condition is fulfilled at time t , it will switch to state s'' at time t (although state s has not even stabilized yet). That is, due to the discrepancy between $S_{i,i}$ and S_i , node i switches from state s to state s'' at time t even if (s, s'') is not in \mathcal{T} at all.¹⁰ In a physical chip design, this premature change of state might even result in inconsistent operations on the local memory, up to the point where it cannot be properly described in terms of \mathbb{S} , and thus in terms of our discrete model, anymore. Even worse, the state of i is part of the local memory and the node’s state signal may attain an undefined value that is propagated to other nodes and their memory. While avoiding the latter is the task of the input ports of a non-faulty node, our goal is to prevent this erroneous behaviour in situations where input ports attain legitimate values only.

Therefore, we define node i to be *metastability-free*, if the situation described above does not occur.

Definition 2.1 (Metastability-Freedom). *Node $i \in V$ is called metastability-free during $[t^-, t^+]$, iff for each time $t \in [t^-, t^+]$ when i switches to some state $s \in \mathbb{S}$, it holds that $\tau_{i,i}(t) < t'$, where t' is the infimum of all times in $(t, t^+]$ when i switches to some state $s' \in \mathbb{S}$.*

Multiple State Machines In some situations the previous definitions are too stringent, as there might be different “components” of a node’s state machine that act concurrently and independently, mostly relying on signals from disjoint input ports or orthogonal components of a signal. We model this by permitting that nodes run several state machines in parallel. All these state machines share the input and local ports of the respective node and are required to have disjoint state spaces. If node i runs state machines M_1, \dots, M_k , node i ’s output signal is the product of the output signals

⁹A single physical fault may cause this behaviour, as at some point a node’s output port must be connected to remote nodes’ input ports. Even if one places bifurcations at different physical locations striving to mitigate this effect, if the voltage at the output port drops below specifications, the values of corresponding input channels may deviate in unpredictable ways.

¹⁰Note that while the “internal” delay $\tau_{i,i}(t) - t$ can be made quite small, it cannot be reduced to zero if the model is meant to reflect physical implementations.

of the individual machines. Formally we define: Each of the state machines M_j , $1 \leq j \leq k$, has an additional own output port s_j . The state of node i 's output port S_i at any time t is given by $S_i(t) := (s_1(t), \dots, s_k(t))$, where the signals of ports s_1, \dots, s_k are defined analogously to the signals of the output ports of state machines in the single state machine case, each. Note that by this definition, the only (local) means for node i 's state machines to interact with each other is by reading the delayed state signal $S_{i,i}$.

We say that *node i 's state machine M_j is in state s at time t* iff $s_j(t) = s$, where $S_i(t) = (s_1(t), \dots, s_k(t))$, and that *node i 's state machine M_j switches to state s at time t* iff signal s_j switches to s at time t . Since the state spaces of the machines M_j are disjoint, we will omit the phrase “state machine M_j ” from the notation, i.e., we write “node i is in state s ” or “node i switched to state s ”, respectively.

Recall that the various state machines of node i are as loosely coupled as remote nodes, namely via the delayed status signal on channel $S_{i,i}$ only. Therefore, it makes sense to consider them independently also when it comes to metastability.

Definition 2.2 (Metastability-Freedom (Multiple State Machines)). *State machine M of node $i \in V$ is called metastability-free during $[t^-, t^+]$, iff for each time $t \in [t^-, t^+]$ when M switches to some state $s \in \mathbb{S}$, it holds that $\tau_{i,i}(t) < t'$, where t' is the infimum of all times in $(t, t^+]$ when M switches to some state $s' \in \mathbb{S}$.*

Note that by this definition the different state machines may switch states concurrently without suffering from metastability.¹¹ It is even possible that some state machine suffers metastability, while another is not affected by this at all.¹²

Problem Statement The purpose of the pulse synchronization protocol is that nodes generate synchronized, well-separated pulses by switching to a distinguished state *accept*. Self-stabilization requires that they start to do so within bounded time, for any possible initial state. However, as our protocol makes use of randomization, there are executions where this does not happen at all; instead, we will show that the protocol stabilizes with probability one in finite time. To give a precise meaning to this statement, we need to define appropriate probability spaces.

Definition 2.3 (Adversarial Spaces). *Denote for $i \in V$ by $\mathcal{C}_i = \{C_{i,k} \mid k \in \{1, \dots, c_i\}\}$ the set of clocks of node i . An adversarial space is a probabilistic space that is defined by subsets of nodes and channels $W \subseteq V$ and $E \subseteq V \times V$, a time interval $[t^-, t^+]$, a protocol \mathcal{P} (nodes' ports, state machines, etc.) as previously defined, sets of clock and delay functions $\mathcal{C} = \bigcup_{i \in V} \mathcal{C}_i$ and $\Theta = \{\tau_{i,j} : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+ \mid (i, j) \in V^2\}$, an initial state \mathcal{E}_0 of all ports, and an adversarial function \mathcal{A} . Here \mathcal{A} is a function that maps a partial execution $\mathcal{E}|_{[0,t]}$ until time t (i.e., all ports' values until time t), W , E , $[t^-, t^+]$, \mathcal{P} , \mathcal{C} , and Θ to the states of all faulty ports during the time interval $(t, t']$, where t' is the infimum of all times greater than t when a non-faulty node or channel switches states.*

¹¹However, care has to be taken when implementing the inter-node communication of the state components in a metastability-free manner, cf. Section 6.

¹²This is crucial for the algorithm we are going to present. For stabilization purposes, nodes comprise a state machine that is prone to metastability. However, the state machine generating pulses (i.e., having the state *accept*, cf. Definition 2.4) does not take its output signal into account once stabilization is achieved. Thus, the algorithm is metastability-free after stabilization in the sense that we guarantee a metastability-free signal indicating when pulses occur.

The adversarial space $\mathcal{AS}(W, E, [t^-, t^+], \mathcal{P}, \mathcal{C}, \Theta, \mathcal{E}_0, \mathcal{A})$ is now defined on the set of all executions \mathcal{E} satisfying that (i) the initial state of all ports is given by $\mathcal{E}|_{[0,0]} = \mathcal{E}_0$, (ii) for all $i \in V$ and $k \in \{1, \dots, c_i\} : C_{i,k}^\mathcal{E} = C_{i,k}$, (iii) for all $(i, j) \in V^2$, $\tau_{i,j}^\mathcal{E} = \tau_{i,j}$, (iv) nodes in W are non-faulty during $[t^-, t^+]$ with respect to the protocol \mathcal{P} , (v) all channels in E are correct during $[t^-, t^+]$, and (vi) given $\mathcal{E}|_{[0,t]}$ for any time t , $\mathcal{E}|_{(t,t']}$ is given by \mathcal{A} , where t' is the infimum of times greater than t when a non-faulty node switches states. Thus, except for when randomized timeouts expire, \mathcal{E} is fully predetermined by the parameters of \mathcal{AS} .¹³ The probability measure on \mathcal{AS} is induced by the random distributions of the randomized timeouts specified by \mathcal{P} .

To avoid confusion, observe that if the clock functions and delays do not follow the model constraints during $[t^-, t^+]$, the respective adversarial space is empty and thus of no concern. This cumbersome definition provides the means to formalize a notion of stabilization that accounts for worst-case drifts and delays and an adversary that knows the full state of the system up to the current time.

We are now in the position to formally state the pulse synchronization problem in our framework. Intuitively, the goal is that after transient faults cease, nodes should with probability one eventually start to issue well-separated, synchronized pulses by switching to a dedicated state *accept*. Thus, as the initial state of the system is arbitrary, specifying an algorithm¹⁴ is equivalent to defining the state machines that run at each node, one of which has a state *accept*.

Definition 2.4 (Self-Stabilizing Pulse Synchronization). *Given a set of nodes $W \subseteq V$ and a set $E \subseteq V \times V$ of channels, we say that protocol \mathcal{P} is a (W, E) -stabilizing pulse synchronization protocol with skew Σ and accuracy bounds T^-, T^+ that stabilizes within time T with probability p iff the following holds. Choose any time interval $[t^-, t^+] \supseteq [t^-, t^- + T + \Sigma]$ and any adversarial space $\mathcal{AS}(W, E, [t^-, t^+], \mathcal{P}, \cdot, \cdot, \cdot, \cdot)$ (i.e., $\mathcal{C}, \Theta, \mathcal{E}_0$, and \mathcal{A} are arbitrary). Then executions from \mathcal{AS} satisfy with probability at least p that there exists a time $t_s \in [t^-, t^- + T]$ so that, denoting by $t_i(k)$ the time when node i switches to a distinguished state *accept* for the k^{th} time after t_s ($t_i(k) = \infty$ if no such time exists), (i) $t_i(1) \in (t_s, t_s + \Sigma)$, (ii) $|t_i(k) - t_j(k)| \leq \Sigma$ if $\max\{t_i(k), t_j(k)\} \leq t^+$, and (iii) $T^- \leq |t_i(k+1) - t_i(k)| \leq T^+$ if $t_i(k) + T^+ \leq t^+$.*

Note that the fact that \mathcal{A} is a deterministic function and, more generally, that we consider each space \mathcal{AS} individually, is no restriction: As \mathcal{P} succeeds for any adversarial space with probability at least p in achieving stabilization, the same holds true for randomized adversarial strategies \mathcal{A} and worst-case drifts and delays.

3 The FATAL Pulse Synchronization Protocol

In this section, we present our self-stabilizing pulse generation algorithm. In order to be suitable for implementation in hardware, it needs to utilize very simple rules only. It is stated in terms of a state machine as introduced in the previous section.

Since the ultimate goal of the pulse generation algorithm is to stabilize a system of DARTS clocks, we introduce an additional port DARTS_i , for each node i , which is driven by node i 's DARTS

¹³This follows by induction starting from the initial configuration \mathcal{E}_0 . Using \mathcal{A} , we can always extend \mathcal{E} to the next time when a correct node switches states, and when correct nodes switch states is fully determined by the parameters of \mathcal{AS} except for when randomized timeouts expire. Note that the induction reaches any finite time within a finite number of steps, as signals switch states finitely often in finite time.

¹⁴We use the terms “algorithm” and “protocol” interchangeably throughout this work.

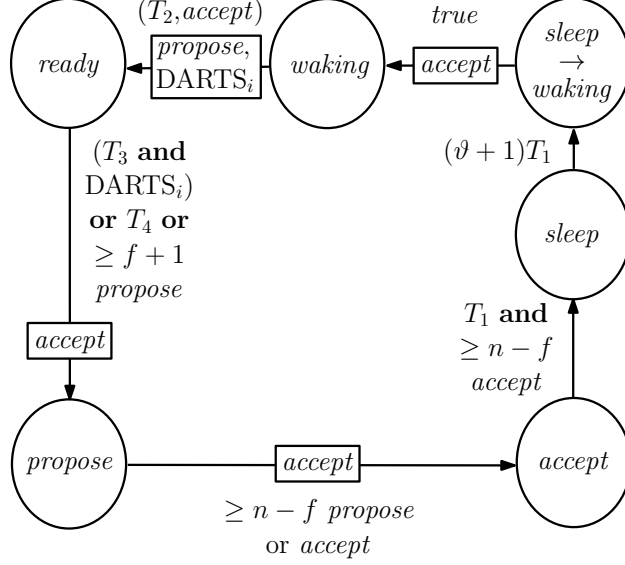


Figure 1: Basic cycle of node i once the algorithm has stabilized.

instance. As for other state signals, its output raises flag $\text{Mem}_{i,\text{DARTS}}$, to which for simplicity we refer to as DARTS_i as well. Note that the DARTS signals are of no concern to the liveness or stabilization of the pulse algorithm itself; rather, it is a control signal from the DARTS component that helps in adjusting the frequency of pulses to the speed of the DARTS clocks once the system as a whole (including the DARTS component) is stable. The pulse algorithm will stabilize independently of the DARTS signal, and the DARTS component will stabilize once the pulse component did so. Therefore we can partition the algorithm's analysis into two parts. When proving the correctness of the algorithm in Section 4, we assume that for each node i , DARTS_i is arbitrary. In Section 7, we will outline how the pulse algorithm and DARTS interact.

3.1 Basic Cycle

The full algorithm makes use of a rather involved interplay between conditions on timeouts, states, and thresholds to converge to a safe state despite a limited number of faulty components. As our approach is thus difficult to present in bulk, we break it down into pieces. Moreover, to facilitate giving intuition about the key ideas of the algorithm, in this section we assume that there are $f < n/3$ faulty nodes, and the remaining $n - f$ nodes are non-faulty within $[0, \infty)$ (where of course the time 0 is unknown to the nodes). We further assume that channels between non-faulty nodes (including loopback channels) are correct within $[0, \infty)$. We start by presenting the basic cycle that is repeated every pulse once a safe configuration is reached (see Figure 1).

We employ graphical representations of the state machine of each node $i \in V$. States are represented by circles containing their names, while transition $(s, s') \in \mathcal{T}$ is depicted as an arrow from s to s' . The guard $tr(s, s')$ is written as a label next to the arrow, and the reset function's value $re(s, s')$ is depicted in a rectangular box on the arrow. To keep labels more simple we make use of some abbreviations. We write T instead of (T, s) if s is the state which node i leaves if the condition involving (T, s) is satisfied. Threshold conditions like “ $\geq f + 1$ s ”, where $s \in \mathbb{S}$, abbreviate Boolean predicates that reach over all of node i 's memory flags $\text{Mem}_{i,j,s}$, where $j \in V$,

and are defined in a straightforward manner. If in such an expression we connect two states by “or”, e.g., “ $\geq n - f$ s or s' ” for $s, s' \in \mathbb{S}$, the summation considers flags of both types s and s' . Thus, such an expression is equivalent to $\sum_{j \in V} \max\{\text{Mem}_{i,j,s}, \text{Mem}_{i,j,s'}\} \geq f + 1$. For any state $s \in \mathbb{S}$, the condition $S_{i,j} = s$, (respectively, $\neg(S_{i,j} = s)$) is written in short as “ j in s ” (respectively, “ j not in s ”). If $j = i$, we simply write “(not) in s ”. We write “true” instead of a condition that is always true (like e.g. “(in s) **or** (not in s)” for an arbitrary state $s \in \mathbb{S}$). Finally, $re(\cdot, \cdot)$ always requires to reset all memory flags of certain types, hence we write e.g. *propose* if all flags $\text{Mem}_{i,j,propose}$ are to be reset.

We now briefly introduce the basic flow of the algorithm once it stabilizes, i.e., once all $n - f$ non-faulty nodes are well-synchronized. Recall that the remaining up to $f < n/3$ faulty nodes may produce arbitrary signals on their outgoing channels. A pulse is locally triggered by switching to state *accept*. Thus, assume that at some time all non-faulty nodes switch to state *accept* within a time window of $2d$, i.e., a valid pulse is generated. Supposing that $T_1 \geq 3\vartheta d$, these nodes will observe, and thus memorize, each other and themselves in state *accept* before T_1 expires. This makes timeout T_1 the critical condition for switching to state *sleep*. From state *sleep*, they will switch to states *sleep* \rightarrow *waking*, *waking*, and finally *ready*, where the timeout (T_2, accept) is determining the time this takes, as it is considerably larger than $\vartheta(\vartheta + 2)T_1$. The intermediate states serve the purpose of achieving stabilization, hence we leave them out for the moment. Note that upon switching to state *ready*, nodes reset their *propose* flags and DARTS_i . Thus, they essentially ignore these signals between the most recent time they switched to *propose* before switching to *accept* and the subsequent time when they switch to *ready*. This ensures that nodes do not take into account outdated information for the decision when to switch to state *propose*. Hence, it is guaranteed that the first node switching from state *ready* to state *propose* again does so because T_4 expired or because T_3 expired and its DARTS memory flag is true. Due to the constraint $\min\{T_3, T_4\} \geq \vartheta(T_2 + 4d)$, we are sure that all non-faulty nodes observe themselves in state *ready* before the first one switches to *propose*. Hence, no node deletes information about nodes that switch to *propose* again after the previous pulse. The first non-faulty node that switches to state *accept* again cannot do so before it memorizes at least $n - f$ nodes in state *propose*, as the *accept* flags are reset upon switching to state *propose*. Therefore, at this time at least $n - 2f \geq f + 1$ non-faulty nodes are in state *propose*. Hence, the rule that nodes switch to *propose* if they memorize $f + 1$ nodes in states *propose* will take effect, i.e., the remaining non-faulty nodes in state *ready* switch to *propose* after less than d time. Another d time later all non-faulty nodes in state *propose* will have become aware of this and switch to state *accept* as well, as the threshold of $n - f$ nodes in states *propose* or *accept* is reached. Thus the cycle is complete and the reasoning can be repeated inductively.

Clearly, for this line of argumentation to be valid, the algorithm could be simpler than stated in Figure 1. We already mentioned that the motivation of having three intermediate states between *accept* and *ready* is to facilitate stabilization. Similarly, there is no need to make use of the *accept* flags in the basic cycle at all; in fact, it adversely affects the constraints the timeouts need to satisfy for the above reasoning to be valid. However, the *accept* flags are much better suited for diagnostic purposes than the *propose* flags, since nodes are expected to switch to *accept* in a small time window and remain in state *accept* for a small period of time only (for all our results, it is sufficient if $T_1 = 4\vartheta d$). Moreover, two different timeout conditions for switching from *ready* to *propose* are unnecessary for correct operation of the pulse synchronization routine. As discussed before, they are introduced in order to allow for a seamless coupling to the DARTS system. We elaborate on this in Section 7.

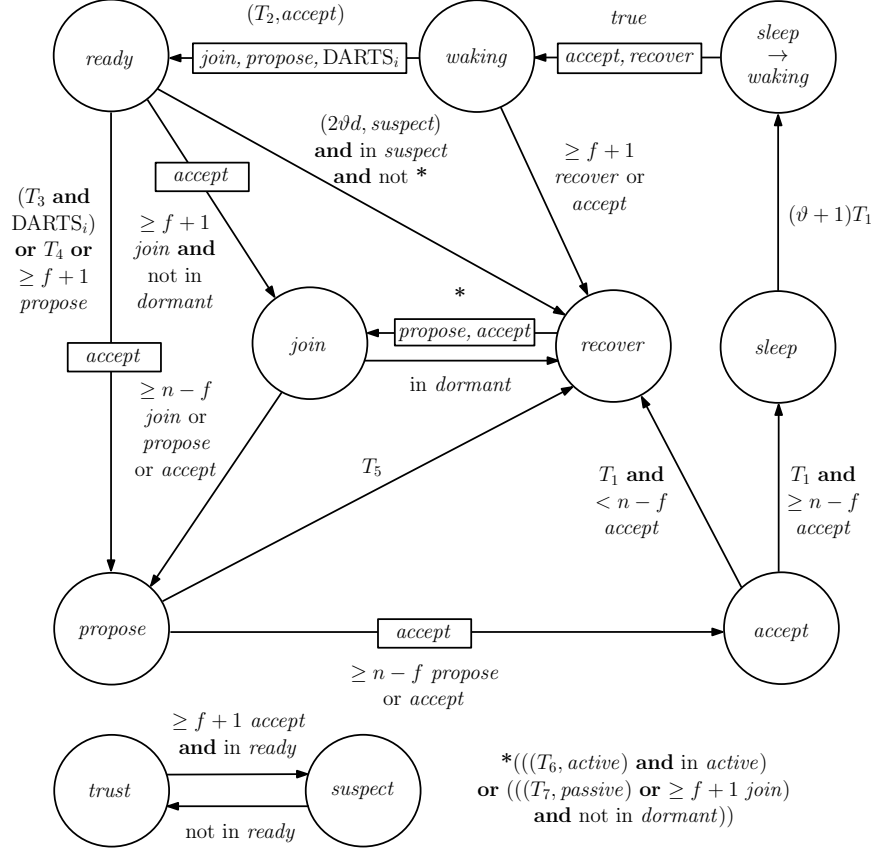


Figure 2: Overview of the core routine of node i 's self-stabilizing pulse algorithm.

3.2 Main Algorithm

We proceed by describing the main routine of the pulse algorithm in full. Alongside the main routine, several other state machines run concurrently and provide additional information to be used during recovery.

The main routine is graphically presented in Figure 2, together with a very simple second component whose sole purpose is to simplify the otherwise overloaded description of the main routine. Except for the states *recover* and *join* and additional resets of memory flags, the main routine is identical to the basic cycle. The purpose of the two additional states is the following: Nodes switch to state *recover* once they detect that something is wrong, that is, non-faulty nodes do not execute the basic cycle as outlined in Section 3.1. This way, non-faulty nodes will not continue to confuse others by sending for example state signals *propose* or *accept* despite clearly being out-of-sync. There are various consistency checks that nodes perform during each execution of the basic cycle. The first one is that in order to switch from state *accept* to state *sleep*, non-faulty nodes need to memorize at least $n - f$ nodes in state *accept*. If this does not happen within T_1 time after switching to state *accept*, by the arguments given in Section 3.1, they could not have entered state *accept* within $2d$ of each other. Therefore, something must be wrong and it is feasible to switch to state *recover*. Next, whenever a non-faulty node is in state *waking*, there should be no non-faulty nodes in states *accept* or *recover*. Considering that the node resets its *accept* and

recover flags upon switching to *waking*, it should not memorize $f + 1$ or more nodes in states *accept* or *recover* at a time when it observes itself in state *waking*. If it does, however, it again switches to state *recover*. Similarly, when in state *ready*, nodes expect others not to be in state *accept* for more than a short period of time, as a non-faulty node switching to *accept* should imply that every non-faulty node switches to *propose* and then to *accept* shortly thereafter. This is expressed by the second state machine comprising two states only. If a node is in state *ready* and memorizes $f + 1$ nodes in state *accept*, it switches to *suspect*. Subsequently, if it remains in state *ready* until a timeout of $2\vartheta d$ expires, it will switch to state *recover*. Last but not least, during a synchronized execution of the basic cycle, no non-faulty node may be in state *propose* for more than a certain amount of time before switching to state *accept*. Therefore, nodes will switch from *propose* to *recover* when timeout T_5 expires.

Nodes can join the basic cycle again via the second new state, called *join*. Since the Byzantine nodes may “play nice” towards $n - 2f$ or more nodes still executing the basic cycle, making them believe that system operation continues as usual, it must be possible to join the basic cycle again without having a majority of nodes in state *recover*. On the other hand, it is crucial that this happens in a sufficiently well-synchronized manner, as otherwise nodes could drop out again because the various checks of consistency detect an erroneous execution of the basic cycle.

In part, this issue is solved by an additional agreement step. In order to enter the basic cycle again, nodes need to memorize $n - f$ nodes in states *join* (the respective nodes detected an inconsistency), *propose* (these nodes continued to execute the basic cycle), or *accept* (there are executions where nodes reset their *propose* flags because of switching to *join* when other nodes already switched to *accept*). Since there are thresholds of $f + 1$ nodes memorized in state *join* both for leaving state *recover* and switching from *ready* to *join*, all nodes will follow the first one switching from *join* to *propose* quickly, just as with the switch from *propose* to *accept* in an ordinary execution of the basic cycle. However, it is decisive that all nodes are in states that permit to participate in this agreement step in order to guarantee success of this approach.

As a result, still a certain degree of synchronization needs to be established beforehand, both among nodes that still execute the basic cycle and those that do not. For instance, if at the point in time when a majority of nodes and channels become non-faulty, some nodes already memorize nodes in *join* that are not, they may switch to state *join* and subsequently *propose* prematurely, causing others to have inconsistent memory flags as well. Again, Byzantine faults may sustain this amiss configuration of the system indefinitely.

So why did we put so much effort in “shifting” the focus to this part of the algorithm? The key advantage is that nodes outside the basic cycle may take into account less reliable information for stabilization purposes. They may take the risk of metastable upsets (as we know it is impossible to avoid these during the stabilization process, anyway) and make use of randomization.

In fact, to make the above scheme work, it is sufficient that all non-faulty nodes agree on a so called *resynchronization point* (formally defined later on), that is, a point in time at which nodes reset the memory flags for states *join* and *sleep* \rightarrow *waking* as well as certain timeouts, while guaranteeing that no node is in these states close to the respective reset times. Except for state *sleep* \rightarrow *waking*, all of these timeouts, memory flags, etc. are not part of the basic cycle at all, thus nodes may enforce consistent values for them when they agree on such a resynchronization point.

Conveniently, the use of randomization also ensures that it is quite unlikely that nodes are in state *sleep* \rightarrow *waking* close to a resynchronization point, as the consistency check of having to memorize $n - f$ nodes in state *accept* in order to switch to state *sleep* guarantees that the time

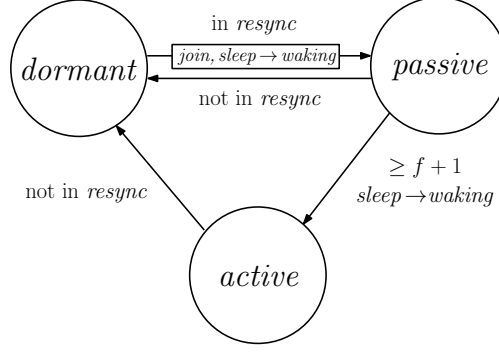


Figure 3: Extension of node i 's core routine.

windows during which non-faulty nodes may switch to *sleep* make up a small fraction of all times only.

Consequently, the remaining components of the algorithm deal with agreeing on resynchronization points and utilizing this information in an appropriate way to ensure stabilization of the main routine. We describe this connection to the main routine first. It is done by another, quite simple state machine, which runs in parallel alongside the core routine. It is depicted in Figure 3.

Its purpose is to reset memory flags in a consistent way and to determine when a node is permitted to switch to *join*. In general, a resynchronization point (locally observed by switching to state *resync*, which is introduced later) triggers the reset of the *join* and *sleep* \rightarrow *waking* flags. If there are still nodes executing the basic cycle, a node may become aware of it by observing $f + 1$ nodes in state *sleep* \rightarrow *waking* at some time. In this case it switches from the state *passive*, which it entered at the point in time when it locally observed the resynchronization point, to the state *active*, which enables an earlier transition to state *join*. This is expressed by the rather involved transition rule $tr(recover, join)$: T_6 is much smaller than T_7 , but T_6 is of no concern until the node switches to state *active* and resets T_6 .¹⁵

It remains to explain how nodes agree on resynchronization points.

3.3 Resynchronization Algorithm

The resynchronization routine is specified in Figure 4 as well. It is a lower layer that the core routine uses for stabilization purposes only. It provides some synchronization that is very similar to that of a pulse, except that such “weak pulses” occur at random times, and may be generated inconsistently after the algorithm as a whole has stabilized. Since the main routine operates independently of the resynchronization routine once the system has stabilized, we can afford the weaker guarantees of the routine: If it succeeds in generating a “good” resynchronization point merely once, the main routine will stabilize deterministically.

Definition 3.1 (Resynchronization Points). *Given $W \subseteq V$, time t is a W -resynchronization point iff each node in W switches to state *supp* \rightarrow *resync* in the time interval $(t, t + 2d)$.*

Definition 3.2 (Good Resynchronization Points). *A W -resynchronization point is called good if*

¹⁵The condition “not in *dormant*” here ensures that the transition is not performed because the node has been in state *resync* a long time ago, but there was no recent switching to *resync*.

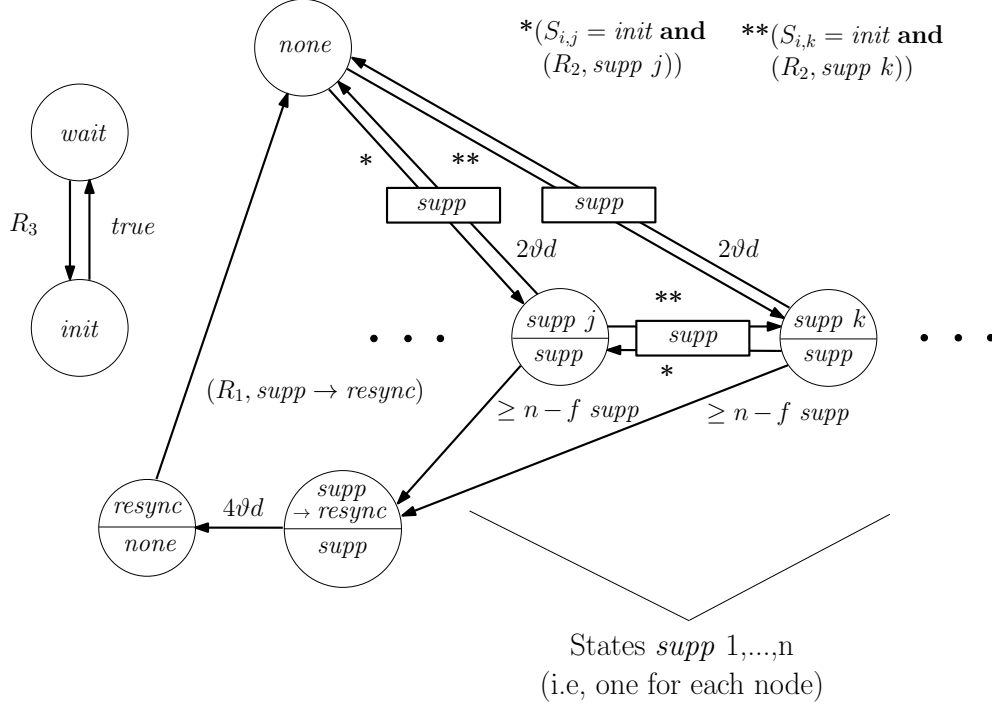


Figure 4: Resynchronization algorithm, comprising two state machines executed in parallel at node i .

no node from W switches to state sleep during $(t - (\vartheta + 3)T_1, t)$ and no node is in state join during $[t - T_1 - d, t + 4d)$.

In order to clarify that despite having a linear number of states ($supp_1, \dots, supp_n$), this part of the algorithm can be implemented using 2-bit communication channels between state machines only, we generalize our description of state machines as follows. If a state is depicted as a circle separated into an upper and a lower part, the upper part denotes the local state, while the lower part indicates the signal state to which it is mapped. A node's memory flags then store the respective signal states only, i.e., remote nodes do not distinguish between states that share the same signal. Clearly, such a machine can be simulated by a machine as introduced in the model section. The advantage is that such a mapping can be used to reduce the number of transmitted state bits; for the resynchronization routine given in Figure 4, we merely need two bits ($init/wait$ and $none/supp$) instead of $\lceil \log(n + 3) \rceil + 1$ bits.

The basic idea behind the resynchronization algorithm is the following: Every now and then, nodes will try to initiate agreement on a resynchronization point. This is the purpose of the small state machine on the left in Figure 4. Recalling that the transition condition “true” simply means that the node switches to state *wait* again as soon as it observes itself in state *init*, it is easy to see that it does nothing else than creating an *init* signal as soon as R_3 expires and resetting R_3 again as quickly as possible. As the time when a node switches to *init* is determined by the randomized timeout R_3 distributed over a large interval (cf. Equality (11)) only, it is impossible to predict when it will expire, even with full knowledge of the execution up to the current point in time. Note that the complete independence of this part of node i 's state from the remaining protocol implies that

faulty nodes are not able to influence the respective times by any means.

Consider now the state machine displayed on the right of Figure 4. To understand how the routine is intended to work, assume that at the time t when a non-faulty node i switches to state *init*, all non-faulty nodes are not in any of the states $\text{supp} \rightarrow \text{resync}$, *resync*, or $\text{supp } i$, and at all non-faulty nodes the timeout $(R_2, \text{supp } i)$ has expired. Then, no matter what the signals from faulty nodes or on faulty channels are, all non-faulty nodes will be in one of the states $\text{supp } j$, $j \in V$, or $\text{supp} \rightarrow \text{resync}$ at time $t + d$. Hence, they will observe each other (and themselves) in one of these states at some time smaller than $t + 2d$. These statements follow from the various timeout conditions of at least $2\vartheta d$ and the fact that observing node i in state *init* will make nodes switch to state $\text{supp } i$ if in *none* or $\text{supp } j$, $j \neq i$. Hence, all of them will switch to state $\text{supp} \rightarrow \text{resync}$ during $(t, t + 2d)$, i.e., t is a resynchronization point. Since t follows a random distribution that is independent of the remaining algorithm and, as mentioned earlier, most of the times nodes cannot switch to state *sleep* and it is easy to deal with the condition on *join* states, there is a large probability that t is a good resynchronization point. Note that timeout R_1 makes sure that no non-faulty node will switch to $\text{supp} \rightarrow \text{resync}$ again anytime soon, leaving sufficient time for the main routine to stabilize.

The scenario we just described relies on the fact that at time t no node is in state $\text{supp} \rightarrow \text{resync}$ or state *resync*. We will choose $R_2 \gg R_1$, implying that $R_2 + 3d$ time after a node switched to state *init* all nodes have “forgotten” about this, i.e., $(R_2, \text{supp } i)$ is expired and they switched back to state *none* (unless other *init* signals interfered). Thus, in the absence of Byzantine faults, the above requirement is easily achieved with a large probability by choosing R_3 as a uniform distribution over some interval $[R_2 + 3d, R_2 + \Theta(nR_1)]$: Other nodes will switch to *init* $\mathcal{O}(n)$ times during this interval, each time “blocking” other nodes for at most $\mathcal{O}(R_1)$ time. If the random choice picks any other point in time during this interval, a resynchronization point occurs. Even if the clock speed of the clock driving R_3 is manipulated in a worst-case manner (affecting the density of the probability distribution with respect to real time by a factor of at most ϑ), we can just increase the size of the interval to account for this.

However, what happens if only *some* of the nodes receive an *init* signal due to faulty channels or nodes? If the same holds for some of the subsequent *supp* signals, it might happen that only a fraction of the nodes reaches the threshold for switching to state $\text{supp} \rightarrow \text{resync}$, resulting in an inconsistent reset of flags and timeouts across the system. Until the respective nodes switch to state *none* again, they will not support a resynchronization point again, i.e., about R_1 time is “lost”. This issue is the reason for the agreement step and the timeouts $(R_2, \text{supp } j)$. In order for any node to switch to state $\text{supp} \rightarrow \text{resync}$, there must be at least $n - 2f \geq f + 1$ non-faulty nodes supporting this. Hence, all of these nodes recently switched to a state $\text{supp } j$ for some $j \in V$, resetting $(R_2, \text{supp } j)$. Until these timeouts expire, $f + 1 \in \Omega(n)$ non-faulty nodes will ignore *init* signals on the respective channels. Since there are $\mathcal{O}(n^2)$ channels, it is possible to choose $R_2 \in \mathcal{O}(nR_1)$ such that this may happen at most $\mathcal{O}(n)$ times in $\mathcal{O}(n)$ time. Playing with constants, we can pick $R_3 \in \mathcal{O}(n)$ maintaining that still a constant fraction of the times are “good” in the sense that R_3 expiring at a non-faulty node will result in a good resynchronization point.

3.4 Timeout Constraints

Condition 3.3 summarizes the constraints we require on the timeouts for the core routine and the resynchronization algorithm to act and interact as intended.

Condition 3.3 (Timeout Constraints). *Define*

$$\lambda := \sqrt{\frac{25\vartheta - 9}{25\vartheta}} \in \left(\frac{4}{5}, 1\right), \quad (1)$$

$\Delta_g := (\vartheta + 3)T_1$, $\Delta_s := T_2/\vartheta - 2T_1 - d$, $\delta_s := 2T_1 + 3d$, and $\tilde{\delta}_s := (\vartheta + 2 - 1/\vartheta)T_1 + 4d$. The timeouts need to satisfy the constraints

$$T_1 \geq \vartheta 4d \quad (2)$$

$$T_2 \geq \vartheta \max \left\{ T_1 + \Delta_g - (4\vartheta^2 + 16\vartheta + 5)d, \left(3\vartheta + 1 - \frac{1}{\vartheta}\right) T_1 + T_5 \right\} \quad (3)$$

$$T_3 \geq \max \left\{ (\vartheta - 1)T_2 + \vartheta(2T_1 + (2\vartheta + 4)d), (2\vartheta^2 + 3\vartheta - 1)T_1 - T_2 + \vartheta(T_6 + 5d) \right\} \quad (4)$$

$$T_4 \geq T_3 \quad (5)$$

$$T_5 \geq \max \left\{ \vartheta(T_4 + 7d) - T_3 + (\vartheta - 1)T_2, (\vartheta^2 + \vartheta - 2)T_1 + \vartheta(T_2 + T_4 + 9d) - T_6 \right\} \quad (6)$$

$$T_6 \geq \vartheta \left(\tilde{\delta}_s - \left(1 - \frac{1}{\vartheta}\right) T_1 + T_2 + 2d \right) > \vartheta \Delta_s \quad (7)$$

$$T_7 \geq \vartheta(T_2 + T_4 + T_5 + \Delta_s + \tilde{\delta}_s - \Delta_g + d) + T_6 - 4d \quad (8)$$

$$R_1 \geq \vartheta \max \left\{ T_7 + (4\vartheta + 8)d, \left(2\vartheta + 4 - \frac{3}{\vartheta}\right) T_1 + 2T_4 + T_5 - \Delta_s - \Delta_g + 17d \right\} \quad (9)$$

$$R_2 \geq \frac{2\vartheta(R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d)(n - f)}{1 - \lambda} \quad (10)$$

$$R_3 = \text{uniformly distributed random variable on } [\vartheta(R_2 + 3d), \vartheta(R_2 + 3d) + 8(1 - \lambda)R_2] \quad (11)$$

$$\lambda \leq \frac{\Delta_s - \Delta_g - \delta_s}{\Delta_s}. \quad (12)$$

We need to show for which values of ϑ this system can be solved. Furthermore, we would like to allow for the largest possible drift of DARTS clocks, which necessitates to maximize the ratio $(T_2 + T_4)/(\vartheta(T_2 + T_3 + 4d))$, that is, the minimal gap between pulses provided that the states of the DARTS signals are zero divided by the maximal time it takes nodes to observe themselves in state *ready* with T_3 expired after a pulse (as then they will respond to DARTS_i switching to one).

Lemma 3.4. *Define $\vartheta_{\max} \approx 1.247$ as the positive solution of $2\vartheta + 1 = \vartheta^3 + \vartheta^2$. Given that $\vartheta < \vartheta_{\max}$, Condition 3.3 can be satisfied with $T_1, \dots, T_7, R_1 \in \mathcal{O}(1)$ and $R_2 \in \mathcal{O}(n)$. The ratio*

$$\frac{(T_2 + T_4)/\vartheta}{T_2 + T_3 + 4d}$$

can be made larger than any constant smaller than

$$\frac{\vartheta^3 + 2\vartheta + 1}{2\vartheta^4 + \vartheta^3}.$$

Proof. First, we identify several redundant inequalities in the system. We have that

$$\begin{aligned} \left(2\vartheta + 2 - \frac{1}{\vartheta}\right) T_1 + T_5 &\stackrel{(6)}{>} 3\vartheta T_1 + T_2 + T_4 - T_6 \\ &\stackrel{(4,5)}{>} 7\vartheta T_1 \\ &> T_1 + \Delta_g - (4\vartheta^2 + 16\vartheta + 5)d, \end{aligned}$$

i.e., the left term in the maximum in Inequality (3) is redundant. The same holds true for the left terms in the maxima in Inequality (4) and Inequality (6), since

$$\begin{aligned} (2\vartheta^2 + 3\vartheta - 1)T_1 - T_2 + \vartheta(T_6 + 5d) &\stackrel{(7)}{>} 3\vartheta T_1 + (\vartheta - 1)T_2 + 4d \\ &\stackrel{(2)}{>} (\vartheta - 1)T_2 + \vartheta(2T_1 + (2\vartheta + 4)d) \end{aligned}$$

and

$$\begin{aligned} \vartheta(T_4 + 7d) - T_3 + (\vartheta - 1)T_2 &\stackrel{(4)}{<} \vartheta(T_2 + T_4 - T_6 + 7d) \\ &< (\vartheta^2 + \vartheta - 2)T_1 + \vartheta(T_2 + T_4 + 9d) - T_6. \end{aligned}$$

Finally, we can eliminate the right term in the maximum in Inequality (9) from the system, as

$$\begin{aligned} T_7 + (4\vartheta + 8)d &\stackrel{(8)}{>} T_2 + T_4 + T_5 + T_6 + 2\tilde{\delta}_s - \Delta_g + 13d \\ &\stackrel{(3)}{>} \left(2\vartheta + 4 - \frac{3}{\vartheta}\right) T_1 + T_4 + 2T_5 + T_6 - \Delta_g + 17d \\ &\stackrel{(6)}{>} \left(2\vartheta + 4 - \frac{3}{\vartheta}\right) T_1 + T_2 + 2T_4 + T_5 - \Delta_g + 17d. \end{aligned}$$

Next, it is not difficult to see that the right hand sides of all inequalities are strictly increasing in T_1 (except for Inequality (12), whose right hand side decreases with T_1), implying that w.l.o.g. we may set $T_1 := 4\vartheta d$. Similarly, we demand that Inequality (8), Inequality (9), and Inequality (10) are satisfied with equality, i.e.,

$$\begin{aligned} T_7 &= \vartheta(T_2 + T_4 + T_5) + T_6 - (4\vartheta^2 + 4)d \\ R_1 &= \vartheta T_7 + (4\vartheta^2 + 8\vartheta)d \\ R_2 &= \frac{2\vartheta(R_1 + T_2/\vartheta + (4\vartheta^2 + 16\vartheta + 9)d)(n - f)}{1 - \lambda} \\ R_3 &= \text{uniformly distributed random variable on } [\vartheta(R_2 + 3d), \vartheta(R_2 + 3d) + 8(1 - \lambda)R_2]. \end{aligned}$$

We set $T_4 := \alpha T_3$ for a parameter

$$\alpha \in \left[1, \frac{2\vartheta + 1}{\vartheta^3 + \vartheta^2}\right],$$

implying that Inequality (5) holds by definition. The remaining simpler system is as follows.

$$T_2 \geq (8\vartheta^3 + 8\vartheta^2 - 4\vartheta)d + \vartheta T_5 \quad (13)$$

$$T_3 \geq (8\vartheta^3 + 12\vartheta^2 + \vartheta)d - T_2 + \vartheta T_6 \quad (14)$$

$$T_5 \geq (4\vartheta^3 + 4\vartheta^2 + \vartheta)d + \vartheta(T_2 + \alpha T_3) - T_6 \quad (15)$$

$$T_6 \geq (4\vartheta^2 + 6\vartheta - 4)d + T_2 \quad (16)$$

$$\sqrt{\frac{25\vartheta - 9}{25\vartheta}} \leq \frac{T_2/\vartheta - (4\vartheta^2 + 28\vartheta + 4)d}{T_2/\vartheta - (8\vartheta + 1)d}.$$

Note the above equalities do not affect this system and can be resolved iteratively once the other variables are fixed. We observe that the right hand side of Inequality (13) is increasing in T_5 , the

right hand side of Inequality (15) is increasing in T_3 , and neither T_3 nor T_5 are present in any further inequalities. Hence, we rule that Inequality (14) and Inequality (15) shall be satisfied with equality, i.e.,

$$\begin{aligned} T_3 &= (8\vartheta^3 + 12\vartheta^2 + \vartheta)d - T_2 + \vartheta T_6 \\ T_5 &= (\alpha(8\vartheta^4 + 12\vartheta^3 + \vartheta^2) + (4\vartheta^3 + 4\vartheta^2 + \vartheta))d - (\vartheta\alpha - 1)T_2 + (\vartheta^2\alpha - 1)T_6 \end{aligned}$$

and arrive at the subsystem

$$\begin{aligned} T_2 &\geq \frac{(\alpha(8\vartheta^5 + 12\vartheta^4 + \vartheta^3) + (4\vartheta^4 + 12\vartheta^3 + 9\vartheta^2 - 4\vartheta))d + (\vartheta^3\alpha - \vartheta)T_6}{1 + \vartheta - \vartheta^2\alpha} \\ T_6 &\geq (4\vartheta^2 + 6\vartheta - 4)d + T_2 \\ T_2 &\geq \frac{(4\vartheta^3 + 20\vartheta^2 + 3\vartheta)d}{1 - \sqrt{(25\vartheta - 9)/(25\vartheta)}}, \end{aligned} \tag{17}$$

where we used that $1 + \vartheta - \vartheta^2\alpha > 0$. Now we can see that Inequality (17) is also increasing in T_6 , set

$$T_6 := (4\vartheta^2 + 6\vartheta - 4)d + T_2,$$

and obtain

$$T_2 \geq \frac{(\alpha(12\vartheta^5 + 18\vartheta^4 - 3\vartheta^3) + (4\vartheta^4 + 8\vartheta^3 + 3\vartheta^2))d}{1 + 2\vartheta - (\vartheta^3 + \vartheta^2)\alpha} \tag{18}$$

$$T_2 \geq \frac{25(1 + \sqrt{(25\vartheta - 9)/(25\vartheta)})(4\vartheta^4 + 20\vartheta^3 + 3\vartheta^2)d}{9}, \tag{19}$$

exploiting that $1 + 2\vartheta - (\vartheta^3 + \vartheta^2)\alpha > 0$.

Since α and thus ϑ are constantly bounded (and we treat d as constant as well), we have a feasible solution for $T_2 \in \mathcal{O}(1)$ (considering asymptotic with respect to n). Resolving the equalities we derived for the other variables, we see that $T_1, \dots, T_7, R_1 \in \mathcal{O}(1)$ and $R_2 \in \mathcal{O}(n)$ as claimed.

It remains to determine the maximal ratio $(T_2 + T_4)/(\vartheta(T_2 + T_3 + 4d)) = (T_2 + \alpha T_3)/(\vartheta(T_2 + T_3 + 4d))$ we can ensure. Obviously, for any value of α , fixing either T_2 or T_3 implies that we want to minimize T_2 or maximize T_3 , respectively. Have a look at Inequalities (13)–(16) again. The solution we constructed minimized T_3 and subsequently T_2 , parametrized by feasible values of α . Increase now T_3 by $x \in \mathbb{R}^+$ in Inequality (14). Consequently, we may increase T_6 in Inequality (16) by x/ϑ compared to our previous solution (where we minimized all inequalities). Hence, we need to increase T_5 by $(\vartheta\alpha - 1/\vartheta)x$ according to Inequality (15), and finally T_2 by $\vartheta(\vartheta\alpha - 1/\vartheta)x$. Thus, for any feasible α and any $\varepsilon > 0$, we can achieve that $T_2 \leq (\vartheta^2\alpha - 1 + \varepsilon)T_3$ if we just choose x large enough. We conclude that we can get arbitrarily close to the ratio

$$\frac{(\alpha + (\vartheta^2\alpha - 1))T_3}{\vartheta(1 + (\vartheta^2\alpha - 1))T_3} = \frac{\vartheta^2\alpha + \alpha - 1}{\vartheta^3\alpha}.$$

Inserting the supremum of admissible values for α , this expression becomes

$$\frac{(2\vartheta + 1)(\vartheta^2 + 1) - (\vartheta^3 + \vartheta^2)}{\vartheta^3(2\vartheta + 1)} = \frac{\vartheta^3 + 2\vartheta + 1}{2\vartheta^4 + \vartheta^3}.$$

This shows the last claim of the lemma, concluding the proof. \square

4 Analysis

In this section we derive skew bounds Σ , as well as accuracy bounds T^-, T^+ , such that the presented protocol is a (W, E) -stabilizing pulse synchronization protocol, for proper choices of the set of nodes W and the set of channels E , with skew Σ and accuracy bounds T^-, T^+ that stabilizes within time $T(k) \in \mathcal{O}(kn)$ with probability $1 - 1/2^{k(n-f)}$, for any $k \in \mathbb{N}$.

To start our analysis, we need to define the basic requirements for stabilization. Essentially, we need that a majority of nodes is non-faulty and the channels between them are correct. However, the first part of the stabilization process is simply that nodes “forget” about past events that are captured by their timeouts. Therefore, we demand that these nodes indeed have been non-faulty for a time period that is sufficiently large to ensure that all timeouts have been reset at least once after the considered set of nodes became non-faulty.

Definition 4.1 (Coherent States). *The subset of nodes $W \subseteq V$ is called coherent during the time interval $[t^-, t^+]$, iff during $[t^- - (\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2) - d, t^+]$ all nodes $i \in W$ are non-faulty, and all channels $S_{i,j}$, $i, j \in W$, are correct.*

We will show that if a coherent set of at least $n - f$ nodes fires a pulse, i.e., switches to *accept* in a tight synchrony, this set will generate pulses deterministically and with controlled frequency, as long the set remains coherent. This motivates the following definitions.

Definition 4.2 (Stabilization Points). *We call t a W -stabilization point (quasi-stabilization point) iff all nodes $i \in W$ switch to *accept* during $[t, t + 2d)$ ($[t, t + 3d)$).*

Throughout this section, we assume the set of coherent nodes W with $|W| \geq n - f$ to be fixed and consider all nodes in and channels originating from $V \setminus W$ as (potentially) faulty. As all our statements refer to nodes in W , we will typically omit the word “non-faulty” when referring to the behaviour or states of nodes in W , and “all nodes” is short for “all nodes in W ”. Note, however, that we will still clearly distinguish between channels originating at faulty and non-faulty nodes, respectively, to nodes in W .

As a first step, we observe that at times when W is coherent, indeed all nodes reset their timeouts, basing the respective state transition on proper perception of nodes in W .

Lemma 4.3. *If the system is coherent during the time interval $[t^-, t^+]$, any (randomized) timeout (T, s) of any node $i \in W$ expiring at a time $t \in [t^-, t^+]$ has been reset at least once since time $t^- - (\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2)$. If t' denotes the time when such a reset occurred, for any $j \in W$ it holds that $S_{i,j}(t') = S_j(\tau_{j,i}^{-1}(t'))$, i.e., at time t' , i observes j in a state j attained when it was non-faulty.*

Proof. According to Condition 3.3, the largest possible value of any (randomized) timeout is $\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2$. Hence, any timeout that is in state 1 at a time smaller than $t^- - (\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2)$ expires before time t_1 or is reset at least once. As by the definition of coherency all nodes in W are non-faulty and all channels between such nodes are correct during $[t^- - (\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2) - d, t^+]$, this implies the statement of the lemma. \square

Phrased informally, any corruption of timeout and channel states eventually ceases, as correct timeouts expire and correct links remember no events that lie d or more time in the past. Proper cleaning of the memory flags is more complicated and will be explained further down the road.

Throughout this section, we will assume for the sake of simplicity that the system is **coherent at all times** and use this lemma implicitly, e.g. we will always assume that nodes from W will observe all other nodes from W in states that they indeed had less than d time ago, expiring of randomized timeouts at non-faulty nodes cannot be predicted accurately, etc. We will discuss more general settings in Section 5.

We proceed by showing that once all nodes in W switch to *accept* in a short period of time, i.e., a W -quasi-stabilization point is reached, the algorithm guarantees that synchronized pulses are generated deterministically with a frequency that is bounded both from above and below.

Theorem 4.4. *Suppose t is a W -quasi-stabilization point. Then*

- (i) *all nodes in W switch to *accept* exactly once within $[t, t + 3d)$, and do not leave *accept* until $t + 4d$, and*
- (ii) *there will be a W -stabilization point $t' \in (t + (T_2 + T_3)/\vartheta, t + T_2 + T_4 + 5d)$ satisfying that no node in W switches to *accept* in the time interval $[t + 3d, t')$ and that*
- (iii) *each node i 's, $i \in W$, core state machine (Figure 1) is metastability-free during $[t + 4d, t' + 4d)$.*

Proof. Proof of (i): Due to Inequality (2), a node does not leave the state *accept* earlier than $T_1/\vartheta \geq 4d$ time after switching to it. Thus, no node can switch to *accept* twice during $[t, t + 3d)$. By definition of a quasi-stabilization point, every node does switch to *accept* in the interval $[t, t + 3d) \subset [t, t + T_1/\vartheta)$. This proves Statement (i).

Proof of (ii): For each $i \in W$, let $t_i \in [t, t + 3d)$ be the time when i switches to *accept*. By (i) t_i is well-defined. Further let t'_i be the infimum of times in (t_i, ∞) when i switches to *recover*, *join*, or *propose*.¹⁶ In the following, denote by $i \in W$ a node with minimal t'_i .

We will show that all nodes switch to *propose* via states *sleep*, *sleep* \rightarrow *waking*, *waking*, and *ready* in the presented order. By (i) nodes do not leave *accept* before $t + 4d$. Thus at time $t + 4d$, each node in W is in state *accept* and observes each other node in W in *accept*. Hence, each node in W memorizes each other node in W in *accept* at time $t + 4d$. For each node $j \in W$, let $t_{j,s}$ be the time node j 's timeout T_1 expires first after t_j . Then $t_{j,s} \in (t_j + T_1/\vartheta, t_j + T_1 + d)$.¹⁷ Since $|W| \geq n - f$, each node j switches to state *sleep* at time $t_{j,s}$. Hence, by time $t + T_1 + 4d$, no node will be observed in state *accept* anymore (until the time when it switches to *accept* again).

When a node $j \in W$ switches to state *waking* at the minimal time t_w larger than t_j , it does not do so earlier than at time $t + T_1/\vartheta + (1 + 1/\vartheta)T_1 = t + (1 + 2/\vartheta)T_1 > t + T_1 + 5d$. This implies that all nodes in W have already left *accept* at least d time ago, since they switched to it at their respective times $t_j < t + T_1 + 4d$. Moreover, they cannot switch to *accept* again until t'_i as it is minimal and nodes need to switch to *propose* before switching to *accept*. Hence, nodes in W are not observed in state *accept* during $(t + T_1 + 5d, t'_i]$, in particular not by node j . Furthermore, nodes in W are not observed in state *recover* during $(t_w - d, t'_i]$. As it resets its *accept* and *recover* flags upon switching to *waking*, j will hence neither switch from *waking* to *recover* nor from *trust* to *suspect* during $(t_w, t'_i]$, and thus also not from *ready* to *recover*.

Now consider node i . By the previous observation, it will not switch from *waking* to *recover*, but to *ready*, following the basic cycle. Consequently, it must wait for timeout T_2 to expire, i.e.,

¹⁶Note that we follow the convention that $\inf \emptyset = \infty$ if the infimum is taken with respect to a (from above) unbounded subset of \mathbb{R}_0^+ .

¹⁷The upper bound comprises an additive term of d since T_1 is reset at some time from $(t_j, t_j + d)$.

cannot switch to *ready* earlier than at time $t + T_2/\vartheta$. As nodes in W clear their *join* flags upon switching to state *ready*, by definition of t'_i node i cannot switch from *ready* to *join*, but has to switch to *propose*. Again, by definition of i , it cannot do so before timeouts T_3 or T_4 expire, i.e., before time

$$t + \frac{T_2}{\vartheta} + \frac{\min\{T_3, T_4\}}{\vartheta} \stackrel{(5)}{=} t + \frac{T_2 + T_3}{\vartheta} \stackrel{(4)}{>} t + T_2 + 5d. \quad (20)$$

All other nodes in W will switch to *waking*, and for the first time after t_j , observe themselves in state *waking* at a time within $(t + T_1 + 4d, t + T_1(2 + \vartheta) + 7d)$. Recall that unless they memorize at least $f + 1$ nodes in *accept* or *recover* while being in state *waking*, they will all switch to state *ready* by time

$$\max\{t + T_2 + 4d, t + (\vartheta + 2)T_1 + 7d\} \stackrel{(3)}{=} t + T_2 + 4d. \quad (21)$$

As we just showed that $t'_i > t + T_2 + 5d$, this implies that at time $t + T_2 + 5d$ all nodes are observed in state *ready*, and none of them leaves before time t'_i .

Now choose t' to be the infimum of times from $(t + (T_2 + T_3)/\vartheta, t + T_2 + T_4 + 4d]$ when a node in W switches to state *accept*.¹⁸ Because of Inequality (20), t' is the first time any node $j \in W$ may switch to *accept* again after its respective time t_j . We will next show that no node $j \in W$ can switch to *recover* within $[t_j, t' + 2d]$. Since at time t'_i node j does not memorize other nodes from W in state *accept*, it will also not do so during $[t'_i, t']$. Hence, it cannot switch from *ready* to *recover* during $[t'_i, t' + 2d]$ since it cannot be in state *suspect* during $[t'_i, t']$. By Inequality (20), j cannot switch to *propose* within $[t_j, t + (T_2 + T_3)/\vartheta]$, and thus its timeout T_5 cannot expire until time

$$t + \frac{T_2 + T_3 + T_5}{\vartheta} \stackrel{(6)}{\geq} t + T_2 + T_4 + 7d \geq t' + 3d, \quad (22)$$

making it impossible for j to switch from *propose* to *recover* at a time within $[t_j, t' + 3d]$. What is more, a node from W that switches to *accept* must stay there for at least $T_1/\vartheta > 3d$ time. Thus, by definition of t' , no node $j \in W$ can switch from *accept* to *recover* at a time within $[t_j, t' + 3d]$. Hence, no node $j \in W$ can switch to state *recover* after t_j , but earlier than time $t' + 2d$. As nodes reset their *join* flags upon switching to state *ready*, it follows that no node in W can switch to other states than *propose* or *accept* during $[t + T_2 + 4d, t' + 2d]$. In particular, no node in W resets its *propose* flags during $[t + T_2 + 5d, t' + 2d] \supset [t'_i, t' + 2d]$.

If at time t' a node in W switches to state *accept*, $n - 2f \geq f + 1$ of its *propose* flags corresponding to nodes in W are true, i.e., in state 1. As the node reset its *propose* flags at the most recent time when it switched to *ready* and no nodes from W have been observed in *propose* between this time and t'_i , it holds that $f + 1$ nodes in W switched to state *propose* during $[t'_i, t']$. Since we established that no node resets its *propose* flags during $[t'_i, t' + 2d]$, it follows that all nodes are in state *propose* by time $t' + d$. Consequently, all nodes in W will observe all nodes in W in state *propose* before time $t' + 2d$ and switch to *accept*, i.e., $t' \in (t + (T_2 + T_3)/\vartheta, t + T_2 + T_4 + 4d)$ is a stabilization point. Statement (ii) follows.

On the other hand, if at time t' no node in W switches to state *accept*, it follows that $t' = t + T_2 + T_4 + 4d$. As all nodes observe themselves in state *ready* by time $t + T_2 + 5d$, they switch to *propose* before time $t + T_2 + T_4 + 5d = t' + d$ because T_4 expired. By the same reasoning as in the previous case, they switch to *accept* before time $t' + 2d$, i.e., Statement (ii) holds as well.

¹⁸Note that since we take the infimum on $(t + (T_2 + T_3)/\vartheta, t + T_2 + T_4 + 4d]$, we have that $t' \leq t + T_2 + T_4 + 4d$.

Proof of (iii): We have shown that within $[t_j, t' + 2d]$, any node $j \in W$ switches to states along the basic cycle only. Moreover, such nodes switch to *accept* at some time in $[t', t' + 2d]$. Since $T_1 \geq 4\vartheta d$, this implies that no node observing itself in *accept* after time t' will leave this state before time $t' + 4d$. To show the correctness of Statement (iii), it is thus sufficient to prove that, whenever j switches from state s of the basic cycle to s' of the basic cycle during time $[t_j + d, t' + 2d] \supset [t + 4d, t' + 2d]$, the transition from s to *join* or *recover* is disabled from the time it switches to s' until it observes itself in this state. We consider transitions $tr(accept, recover)$, $tr(waking, recover)$, $tr(ready, recover)$, $tr(ready, join)$, and $tr(propose, recover)$ one after the other:

1. $tr(accept, recover)$: We showed that node j 's $tr(accept, sleep)$ is satisfied before time $t + 4d \leq t + T_1/\vartheta$, i.e., before $tr(accept, recover)$ can hold, and no node resets its *accept* flags less than d time after switching to state *sleep*. When j switches to state *accept* again at or after time t' , T_1 will not expire earlier than time $t' + 4d$.
2. $tr(waking, recover)$: As part of the reasoning in (ii), we derived that $tr(waking, recover)$ does not hold at nodes from W observing themselves in state *waking*.
3. $tr(ready, recover)$ and $tr(ready, join)$: Similarly, we proved that at no node in W , condition $tr(ready, recover)$ or $tr(ready, join)$ can hold during $(t + (T_2 + T_3)/\vartheta, t' + 2d)$, and nodes in W are in state *ready* during $(t + (T_2 + T_3)/\vartheta, t' + d)$ only.
4. $tr(propose, recover)$: Finally, the additional slack of d in Inequality (22) ensures that T_5 does not expire at any node in W switching to state *accept* during $(t', t' + 2d)$ earlier than time $t' + 3d$.

Since $[t_j, t' + 4d] \supset [t + 3d, t' + 4d]$, Statement (iii) follows. \square

Inductive application of Theorem 4.4 shows that by construction of our algorithm, nodes in W provably do not suffer from metastability upsets once a W -quasi-stabilization point is reached, as long as all nodes in W remain non-faulty and the channels connecting them correct. Unfortunately, it can be shown that it is impossible to ensure this property during the stabilization period, thus rendering a formal treatment infeasible. This is not a peculiarity of our system model, but a threat to any model that allows for the possibility of metastable upsets as encountered in physical chip designs. However, it was shown that, by proper chip design, the probability of metastable upsets can be made arbitrarily small [13]. **In the remainder of this work, we will therefore assume that all non-faulty nodes are metastability-free in all executions.**

The next lemma reveals a very basic property of the main algorithm that is satisfied if no nodes may switch to state *join* in a given period of time. It states that in order for any non-faulty node to switch to state *sleep*, there need to be $f + 1$ non-faulty nodes supporting this by switching to state *accept*. Subsequently, these nodes cannot do so again for a certain time window. In particular, this implies that during the respective time window no node may switch to *sleep*.

Lemma 4.5. *Assume that at time t_s , some node from W switches to sleep and no node from W is in state *join* during $[t_s - T_1 - d, t^+]$. Then there is a subset $A \subseteq W$ of at least $n - 2f$ nodes such that*

- (i) *each node from A has been in state *accept* at some time in the interval $(t_s - T_1 - d, t_s)$ and*

(ii) no node from A is in state *propose* or switches to state *accept* during the time interval

$$(t_s, \min \{t_s + \Delta_s, t^+\}).$$

Proof. In order to switch to *sleep* at time t_s , a node must have observed $n - 2f$ non-faulty nodes in state *accept* at times from $(t_s - T_1, t_s]$, since it resets its *accept* flags at the time $t_a \geq t_s - T_1$ (that is minimal with this property) when it switched to state *accept*. Each of these nodes must have been in state *accept* at some time from $(t_s - T_1 - d, t_s)$, showing the existence of a set $A \subseteq W$ satisfying Statement (i).

We will next prove Statement (ii). Consider a node $i \in A$. In order to switch to *propose* or again to *accept*, i must switch to *join* first or wait for T_2 to expire after switching to state *accept* some time after $t_s - 2T_1 - d$. However, by assumption the first option is impossible until time t^+ , since no nodes are in state *join* during $[t_s - T_1 - d, t^+]$. Therefore, j will not be in state *propose* or switch to state *accept* again until $t_s - 2T_1 + T_2/\vartheta - d = t_s + \Delta_s$ or t^+ , respectively, whatever is smaller. This proves Statement (ii). \square

Granted that nodes are not in state *join*, this implies that the time windows during which nodes may switch to *sleep* and *sleep* \rightarrow *waking*, respectively, are well-separated.

Corollary 4.6. *Assume that during $[t^- - T_1 - d, t^+]$ no node from W is in state *join*, where $t^+ - t^- \leq \Delta_s$. Then*

- (i) *any time interval $[t_a, t_b] \subseteq [t^-, t^+]$ of minimum length containing all switches of nodes in W from *accept* to *sleep* during $[t^-, t^+]$ has length at most $2T_1 + 3d$, and*
- (ii) *granted that no node from W switches to state *sleep* during $(t^- - (\vartheta + 1)T_1 - d, t^-)$, any time interval $[t_a, t_b] \subseteq [t^-, t^+ + (1 + 1/\vartheta)T_1]$ of minimum length containing all times in $[t^-, t^+ + (1 + 1/\vartheta)T_1]$ when a node from W switches to *sleep* \rightarrow *waking* has length at most $\tilde{\delta}_s$.*

Proof. Consider Statement (i) first. If there is no node from W that switches from *accept* to *sleep* during $[t^-, t^+]$, the statement is trivially satisfied.

Otherwise, choose any such interval $[t_a, t_b]$. Since $[t_a, t_b] \neq \emptyset$ is minimal, both at time t_a and t_b some nodes from W switch to *sleep*. Assume by means of contradiction that $t_b - t_a > 2T_1 + 3d$. Due to the constraints on t^- and t^+ , we have that $t_b \leq t_a + \Delta_s$. Moreover, during $[t_a - T_1 - d, t_b] \subseteq [t^- - T_1 - d, t^+]$ no node from W is in state *join*. Thus, we can apply Lemma 4.5 to t_a and see that at least $n - 2f \geq f + 1$ nodes from W do not switch to *accept* in the time interval

$$(t_a, t_a + \Delta_s) \supset (t_b - (2T_1 + 3d), t_b].$$

As nodes from W leave state *accept* as soon as T_1 expires, these nodes are not in state *accept* during $[t_b - (T_1 + 2d), t_b]$, implying that they are not observed in this state during $[t_b - (T_1 + d), t_b]$. It follows that no node in W can observe more than $n - f - 1$ different nodes in state *accept* during $[t_b - (T_1 + d), t_b]$. As nodes from W clear their *accept* flags upon switching to *accept* and leave state *accept* after less than $T_1 + d$ time, we conclude that no node from W switches to state *sleep* at time t_b . This is a contradiction, implying that the assumption that $t_b - t_a > 2T_1 + 3d$ must be wrong and therefore Statement (i) must be true.

To obtain Statement (ii), observe first that any node from W switching to state *sleep* at some time $t \leq t^- - (\vartheta + 1)T_1 - d$ switches to state *sleep* \rightarrow *waking* before time t^- . Subsequently, it

needs to switch to state *sleep* again in order to be in state *sleep* \rightarrow *waking* at or later than time t^- . On the other hand, every node that switches to *sleep* after time t^+ will not switch to *sleep* \rightarrow *waking* again before time $t^+ + (1 + 1/\vartheta)T_1$. Hence, any node switching to state *sleep* \rightarrow *waking* during the considered interval must switch to *sleep* during $[t^-, t^+]$. Applying Statement (i) to $[t^-, t^+]$ yields that nodes from W can only switch to *sleep* within a time interval of length at most $2T_1 + 3d$. Considering the fastest and slowest possible transitions from *sleep* to *sleep* \rightarrow *waking* we obtain that nodes from W can switch to *sleep* \rightarrow *waking* within a time interval of length at most $2T_1 + 3d + (\vartheta + 1)T_1 + d - (1 + 1/\vartheta)T_1 = \tilde{\delta}_s$. Statement (ii) follows. \square

We are now ready to advance to proving that good resynchronization points are likely to occur within bounded time, no matter what the strategy of the Byzantine faulty nodes and channels is. To this end, we first establish that in any execution, at most of the times a node switching to state *init* will result in a good resynchronization point. This is formalized by the following definition.

Definition 4.7 (Good Times). *Given an execution \mathcal{E} of the system, denote by \mathcal{E}' any execution satisfying that $\mathcal{E}'|_{[0,t]} = \mathcal{E}|_{[0,t]}$, where at time t a node $i \in W$ switches to state *init* in \mathcal{E}' . Time t is good in \mathcal{E} with respect to W provided that for any such \mathcal{E}' it holds that t is a good W -resynchronization point in \mathcal{E}' .*

The previous statement thus boils down to showing that in any execution, the majority of the times is good.

Lemma 4.8. *Given any execution \mathcal{E} and any time interval $[t^-, t^+]$, the volume of good times in \mathcal{E} during $[t^-, t^+]$ is at least*

$$\lambda^2(t^+ - t^-) - \frac{11(1 - \lambda)R_2}{10\vartheta}.$$

Proof. Assume w.l.o.g. that $|W| = n - f$ (otherwise consider a subset of size $n - f$) and abbreviate

$$\begin{aligned} N &:= \left(\frac{\vartheta(t^+ - t^-)}{R_2} + \frac{11}{10} \right) (n - f) \\ &\geq \left\lceil \frac{\vartheta(t^+ - t^-) + R_2/10}{R_2} \right\rceil (n - f) \\ &\stackrel{(10)}{\geq} \left\lceil \frac{\vartheta(t^+ - t^-) + \vartheta(R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d)/(5(1 - \lambda))}{R_2} \right\rceil (n - f) \\ &\stackrel{(1)}{\geq} \left\lceil \frac{\vartheta(t^+ - t^- + (R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d))}{R_2} \right\rceil (n - f) \\ &\stackrel{(3)}{\geq} \left\lceil \frac{\vartheta(t^+ - t^- + R_1 + T_1 + 4d + \Delta_g)}{R_2} \right\rceil (n - f). \end{aligned}$$

The proof is in two steps: First we construct a measurable subset of $[t^-, t^+]$ that comprises good times only. In a second step a lower bound on the volume of this set is derived.

Constructing the set: Consider an arbitrary time $t \in [t^-, t^+]$, and assume a node $i \in W$ switches to state *init* at time t . When it does so, its timeout R_3 expires. By Lemma 4.3 all timeouts of node i that expire at times within $[t^-, t^+]$, have been reset at least once until time t^- . Let t_{E3} be the maximum time not later than t when R_3 was reset. Due to the distribution of R_3 we know that

$$t_{E3} \stackrel{(11)}{\leq} t - (R_2 + 3d).$$

Thus, node i is not in state *init* during time $[t - (R_2 + 2d), t)$, and no node $j \in W$ observes i in state *init* during time $[t - (R_2 + d), t)$. Thereby any node j 's, $j \in W$, timeout $(R_2, \text{supp } i)$ corresponding to node i is expired at time t .

We claim that the condition that no node from W is in or observed in one of the states *resync* or *supp* \rightarrow *resync* at time t is sufficient for t being a W -resynchronization point. To see this, assume that the condition is satisfied. Thus all nodes $j \in W$ are in states *none* or *supp* k for some $k \in \{1, \dots, n\}$ at time t . By the algorithm, they all will switch to state *supp* i or state *supp* \rightarrow *resync* during $(t, t + d)$. It might happen that they subsequently switch to another state *supp* k' for some $k' \in V$, but all of them will be in one of the states with signal *supp* during $(t + d, t + 2d]$. Consequently, all nodes will observe at least $n - f$ nodes in state *supp* during $(t', t + 2d)$ for some time $t' < t + 2d$. Hence, those nodes in W that were still in state *supp* i (or *supp* k' for some k') at time $t + d$ switch to state *supp* \rightarrow *resync* before time $t + 2d$, i.e., t is a W -resynchronization point.

We proceed by analyzing under which conditions t is a good W -resynchronization point. Recall that in order for t to be good, it has to hold that no node from W switches to state *sleep* during $(t - \Delta_g, t)$ or is in state *join* during $(t - T_1 - d, t + 4d)$.

We begin by characterizing subsets of good times within $(t_r, t'_r) \subset [t^-, t^+]$, where t_r and t'_r are times such that during (t_r, t'_r) no node from W switches to state *supp* \rightarrow *resync*. Due to timeout

$$R_1 \stackrel{(9)}{\geq} (4\vartheta + 2)d,$$

we know that during $(t_r + R_1 + 2d, t'_r)$, no node from W will be in, or be observed in, states *supp* \rightarrow *resync* or *resync*. Thus, if a node from W switches to *init* at a time within $(t_r + R_1 + 2d, t'_r)$, it is a W -resynchronization point. Further, all nodes in W will be in state *dormant* during $(t_r + R_1 + 2d, t'_r + 4d)$. Thus all nodes in W will be observed to be in state *dormant* during $(t_r + R_1 + 3d, t'_r + 4d)$, implying that they are not in state *join* during $(t_r + R_1 + 3d, t'_r + 4d)$. In particular, any time $t \in (t_r + R_1 + T_1 + 4d, t'_r)$ satisfies that no node in W is in state *join* during $(t - T_1 - d, t + 4d)$.

Further define t_a to be the infimum of times from $(t_r + R_1 + T_1 + 4d, t'_r]$ when a node from W switches to state *sleep*. By Corollary 4.6, no node from W switches to state *sleep* during $(t_a + \delta_s, \min\{t_a + \Delta_s, t'_r\})$. Hence, if $t_a < \infty$, all times in both $(t_r + R_1 + T_1 + 4d + \Delta_g, t_a)$ and $(t_a + \delta_s + \Delta_g, \min\{t_a + \Delta_s, t'_r\})$ are good.

In case $t_a < t'_r - \Delta_s$ we can repeat the reasoning, defining that t'_a is the infimum of times from $[t_a + \Delta_s, t'_r]$ when a node switches to state *sleep*. By analogous arguments as before we see that all times in the sets $[t_a + \Delta_s, t'_a)$ and $(t'_a + \delta_s + \Delta_g, \min\{t'_a + \Delta_s, t'_r\})$ are good.

By induction on the times t_a, t'_a, \dots, t_a^k (halting once $t_a^k \geq t'_r - \Delta_s$), we infer that the total volume of times from (t_r, t'_r) as well as from $(t_r + R_1 + T_1 + 4d + \Delta_g, t'_r)$ that is good is at least

$$\left\lfloor \frac{t'_r - (t_r + R_1 + T_1 + 4d + \Delta_g)}{\Delta_s} \right\rfloor (\Delta_s - \Delta_g - \delta_s) > \frac{t'_r - (t_r + R_1 + T_1 + 4d + \Delta_g + \Delta_s)}{\Delta_s} (\Delta_s - \Delta_g - \delta_s). \quad (23)$$

In other words, up to a constant loss in each interval (t_r, t'_r) , a constant fraction of the times are good.

Volume of the set: In order to infer a lower bound on the volume of good times during $[t^-, t^+]$, we subtract from $[t^-, t^+]$ all intervals $[t_r, t_r + R_1 + T_1 + 4d + \Delta_g]$, where a node from W switches to $supp \rightarrow resync$ at a time t_r within $[t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$. Formally define

$$\bar{G} = \bigcup_{\substack{t_r \in [t^- - (R_1 + T_1 + 4d + \Delta_g), t^+] \\ \exists i \in W: i \text{ switches to } supp \rightarrow resync \text{ at } t_r}} [t_r, t_r + R_1 + T_1 + 4d + \Delta_g].$$

What remains is the set $[t^-, t^+] \setminus \bar{G}$, that has as subset the union of intervals $(t_r + R_1 + T_1 + 4d + \Delta_g, t'_r) \subseteq [t^-, t^+]$, where t_r and t'_r are times at which a node from W switches to $supp \rightarrow resync$ and no node from W switches to $supp \rightarrow resync$ within (t_r, t'_r) . Note that for each such interval we already know it contains a certain amount of good times because of Inequality (23). In order to lower bound the good times in $[t^-, t^+]$, it is thus feasible to lower bound the volume and number of connected components (i.e., maximal intervals) of any subset of $[t^-, t^+] \setminus \bar{G}$.

Observe that any node in W does not switch to state *init* more than

$$\left\lceil \frac{t^+ - t^- + R_1 + T_1 + 4d + \Delta_g}{R_3} \right\rceil \stackrel{(11)}{\leq} \left\lceil \frac{t^+ - t^- + R_1 + T_1 + 4d + \Delta_g}{R_2 + d} \right\rceil \leq \frac{N}{n - f} \quad (24)$$

times during $[t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$.

Now consider the case that a node in W switches to state $supp \rightarrow resync$ at a time t satisfying that no node in W switched to state *init* during $(t - (8\vartheta + 6)d, t)$. This necessitates that this node observes $n - f$ of its channels in state *supp* during $(t - (2\vartheta + 1)d, t)$, at least $n - 2f \geq f + 1$ of which originate from nodes in W . As no node from W switched to *init* during $(t - (8\vartheta + 6)d, t)$, every node that has not observed a node $i \in V \setminus W$ in state *init* at a time from $(t - (8\vartheta + 4)d, t)$ when $(R_2, supp\ i)$ is expired must be in a state whose signal is *none* during $(t - (2\vartheta + 2)d, t)$ due to timeouts. Therefore its outgoing channels are not in state *supp* during $(t - (2\vartheta + 1)d, t)$. By means of contradiction, it thus follows that for each node j of the at least $f + 1$ nodes (which are all from W), there exists a node $i \in V \setminus W$ such that node j resets timeout $(R_2, supp\ i)$ during the time interval $(t - (8\vartheta + 4)d, t)$.

The same reasoning applies to any time $t' \notin (t - (8\vartheta + 6)d, t)$ satisfying that some node in W switches to state $supp \rightarrow resync$ at time t' and no node in W switched to state *init* during $(t' - (8\vartheta + 6)d, t')$. Note that the set of the respective at least $f + 1$ events (corresponding to the at least $f + 1$ nodes from W) where timeouts $(R_2, supp\ i)$ with $i \in V \setminus W$ are reset and the set of the events corresponding to t are disjoint. However, the total number of events where such a timeout can be reset during $[t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$ is upper bounded by

$$|V \setminus W| |W| \left\lceil \frac{t^+ - t^- + R_1 + T_1 + 4d + \Delta_g}{R_2/\vartheta} \right\rceil < (f + 1)N, \quad (25)$$

i.e., the total number of channels from nodes not in W ($|V \setminus W|$ many) to nodes in W multiplied by the number of times the associated timeout can expire at the receiving node in W during $[t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$.

With the help of inequalities (24) and (25), we can show that \bar{G} can be covered by less than $2N$ intervals of size $(R_1 + T_1 + 4d + \Delta_g) + (8\vartheta + 6)d$ each. By Inequality (24), there are no more than N times $t \in [t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$ when a non-faulty node switches to *init* and thus may cause others to switch to state $supp \rightarrow resync$ at times in $[t, t + (8\vartheta + 6)d]$. Similarly,

Inequality (25) shows that the channels from $V \setminus W$ to W may cause at most $N - 1$ such times $t \in [t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$, since any such time requires the existence of at least $f + 1$ events where timeouts $(R_2, \text{supp } i)$, $i \in V \setminus W$, are reset at nodes in W , and the respective events are disjoint. Thus, all times $t_r \in [t^- - (R_1 + T_1 + 4d + \Delta_g), t^+]$ when some node $i \in W$ switches to $\text{supp} \rightarrow \text{resync}$ are covered by at most $2N - 1$ intervals of length $(8\vartheta + 6)d$.

This results in a cover $\bar{G}' \supseteq \bar{G}$ consisting of at most $2N - 1$ intervals that satisfies that

$$\text{vol}(\bar{G}) \leq \text{vol}(\bar{G}') < 2N(R_1 + T_1 + \Delta_g + (8\vartheta + 10)d).$$

Summing over the at most $2N$ intervals that remain in $[t^-, t^+] \setminus \bar{G}'$ and using Inequality (23), we conclude that the volume of good times during $[t^-, t^+]$ is at least

$$\begin{aligned} & \frac{t^+ - t^- - 2N(R_1 + T_1 + (8\vartheta + 10)d + \Delta_g + \Delta_s)}{\Delta_s} (\Delta_s - \Delta_g - \delta_s) \\ = & \frac{t^+ - t^- - 2N(R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d)}{\Delta_s} (\Delta_s - \Delta_g - \delta_s) \\ \stackrel{(12)}{\geq} & \lambda \left(t^+ - t^- - 2 \left(\frac{\vartheta(t^+ - t^-)}{R_2} + \frac{11}{10} \right) (n - f)(R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d) \right) \\ = & \lambda \left(1 - \frac{2\vartheta(R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d)(n - f)}{R_2} \right) (t^+ - t^-) \\ & - \frac{11\lambda(R_1 + (\vartheta + 2)T_1 + T_2/\vartheta + (8\vartheta + 9)d)(n - f)}{5} \\ \stackrel{(10)}{\geq} & \lambda^2(t^+ - t^-) - \frac{11(1 - \lambda)R_2}{10\vartheta}, \end{aligned}$$

as claimed. The lemma follows. \square

We are now in the position to prove our second main theorem, which states that a good resynchronization point occurs within $\mathcal{O}(R_2)$ time with overwhelming probability.

Theorem 4.9. *Denote by $\hat{E}_3 := \vartheta(R_2 + 3d) + 8(1 - \lambda)R_2 + d$ the maximal value the distribution R_3 can attain plus the at most d time until R_3 is reset whenever it expires. For any $k \in \mathbb{N}$ and any time t , with probability at least $1 - (1/2)^{k(n-f)}$ there will be a good W -resynchronization point during $[t, t + (k + 1)\hat{E}_3]$.*

Proof. Assume w.l.o.g. that $|W| = n - f$ (otherwise consider a subset of size $n - f$). Fix some node $i \in W$ and denote by t_0 the infimum of times from $[t, t + (k + 1)\hat{E}_3]$ when node i switches to *init*. We have that $t_0 < t + \hat{E}_3$. By induction, it follows that node i will switch to state *init* at least another k times during $[t, t + (k + 1)\hat{E}_3]$ at the times $t_1 < t_2 < \dots < t_k$. We claim that each such time t_j , $j \in \{1, \dots, k\}$, has an independently by $1/2$ lower bounded probability of being good and therefore being a good W -resynchronization point.

We prove this by induction on j : As induction hypothesis, suppose for some $j \in \{1, \dots, k - 1\}$, we showed the statement for $j' \in \{1, \dots, j - 1\}$ and the execution of the system is fixed until time t_{j-1} , i.e., $\mathcal{E}|_{[0, t_{j-1}]}$ is given. Now consider the set of executions that are extensions of $\mathcal{E}|_{[0, t_{j-1}]}$ and have the same clock functions as \mathcal{E} . For each such execution \mathcal{E}' it holds that $\mathcal{E}'|_{[0, t_{j-1}]} = \mathcal{E}|_{[0, t_{j-1}]}$, and all nodes' clocks make progress in \mathcal{E}' as in \mathcal{E} . Clearly each such \mathcal{E}' has its own time $t_j < t + (j + 1)\hat{E}_3$

when R_3 expires next after t_{j-1} at node i , and i switches to *init*. We next characterize the distribution of the times t_j .

As the rate of the clock driving node i 's R_3 is between 1 and ϑ , $t_j > t_{j-1}$ is within an interval, call it $[t^-, t^+]$, of size at most

$$t^+ - t^- \leq 8(1 - \lambda)R_2,$$

regardless of the progress that i 's clock C makes in any execution \mathcal{E}' .

Certainly we can apply Lemma 4.8 also to each of the \mathcal{E}' , showing that the volume of times from $[t^-, t^+]$ that are *not* good in \mathcal{E}' is at most

$$(1 - \lambda^2)(t^+ - t^-) + \frac{11(1 - \lambda)R_2}{10\vartheta}.$$

Since clock C can make progress not faster than at rate ϑ and the probability density of R_3 is constantly $1/(8(1 - \lambda)R_2)$ (with respect to the clock function C), we obtain that the probability of t_j not being a good time is upper bounded by

$$\frac{(1 - \lambda^2)(t^+ - t^-) + 11(1 - \lambda)R_2/(10\vartheta)}{8(1 - \lambda)R_2/\vartheta} \leq \vartheta(1 - \lambda^2) + \frac{11}{80} \stackrel{(1)}{<} \vartheta \frac{9}{25\vartheta} + \frac{7}{50} = \frac{1}{2}.$$

Here we use that the time when R_3 expires is independent of $\mathcal{E}'|_{[0, t_{j-1}]}$.

We complete our reasoning as follows. Given $\mathcal{E}|_{[0, t_{j-1}]}$, we permit an adversary to choose \mathcal{E}' , including random bits of all nodes and full knowledge of the future, with the exception that we deny it control or knowledge of the time t_j when R_3 expires at node i , i.e., \mathcal{E}' is an imaginary execution in which R_3 does not expire at i at any time greater than t_{j-1} . Note that for the good W -resynchronisation points we considered, the choice of \mathcal{E}' does not affect the probability that t_1, \dots, t_{j-1} are good W -resynchronization points: The conditions referring to times greater than a W -resynchronisation point t , i.e., that all nodes in W switch to state *supp* \rightarrow *resync* during $(t, t+2d)$ and no node in W shall be in state *join* during $(t - T_1 - d, t + 4d)$, are already fully determined by the history of the system until time t . As we fixed \mathcal{E}' , the behaviour of the clock driving R_3 is fixed as well. Next, we determine the time t_j when R_3 expires according to its distribution, given the behaviour of node i 's clock. The above reasoning shows that time t_j is good in \mathcal{E}' with probability at least $1/2$, independently of $\mathcal{E}'|_{[0, t_{j-1}]} = \mathcal{E}|_{[0, t_{j-1}]}$. We define that $\mathcal{E}|_{[0, t_j]} = \mathcal{E}'|_{[0, t_j]}$ and in \mathcal{E} node i switches to state *init* (because R_3 expired). As — conditional to the clock driving R_3 and t_{j-1} being specified — t_j is independent of $\mathcal{E}|_{[0, t_j]}$, \mathcal{E} is indistinguishable from \mathcal{E}' until time t_j . Because t_j is good with probability at least $1/2$ independently of $\mathcal{E}'|_{[0, t_{j-1}]} = \mathcal{E}|_{[0, t_{j-1}]}$, so it is in \mathcal{E} . Hence, in \mathcal{E} t_j is a good W -resynchronization point with probability $1/2$, independently of $\mathcal{E}|_{[0, t_{j-1}]}$. Since \mathcal{E}' was chosen in an adversarial manner, this completes the induction step.

In summary, we showed that for *any* node in W and *any* execution (in which we do not manipulate the times when R_3 expires at the respective node), starting from the second time during $[t, t + (k+1)\hat{E}_3]$ when R_3 expires at the respective node, there is a probability of at least $1/2$ that the respective time is a good W -resynchronization point. Since we assumed that $|W| = n - f$ and there are at least k such times for each node in W , this implies that having no good W -resynchronization point during $[t, t + (k+1)\hat{E}_3]$ is as least as unlikely as $k(n - f)$ unbiased and independent coin flips all showing tail, i.e., $(1/2)^{k(n-f)}$. This concludes the proof. \square

Having established that eventually a good W -resynchronization point t_g will occur, we turn to proving the convergence of the main routine. We start with a few helper statements wrapping up

that a good resynchronization point guarantees proper reset of flags and timeouts involved in the stabilization process of the main routine.

Lemma 4.10. *Suppose t_g is a good W -resynchronization point. Then*

- (i) *each node $i \in W$ switches to passive at a time $t_i \in (t_g + 4d, t_g + (4\vartheta + 3)d)$ and observes itself in state dormant during $[t_g + 4d, \tau_{i,i}(t_i))$,*
- (ii) *$\text{Mem}_{i,j,\text{join}}|_{[\tau_{i,i}(t_i), t_{\text{join}}]} \equiv 0$ for all $i, j \in W$, where $t_{\text{join}} \geq t_g + 4d$ is the infimum of all times greater than $t_g - T_1 - d$ when a node from W switches to join,*
- (iii) *$\text{Mem}_{i,j,\text{sleep} \rightarrow \text{waking}}|_{[\tau_{i,i}(t_i), t_s]} \equiv 0$ for all $i, j \in W$, where $t_s \geq t_g + (1 + 1/\vartheta)T_1$ is the infimum of all times greater or equal to t_g when a node from W switches to sleep \rightarrow waking,*
- (iv) *no node from W resets its sleep \rightarrow waking flags during $[t_g + (1 + 1/\vartheta)T_1, t_g + R_1/\vartheta]$, and*
- (v) *no node from W resets its join flags due to switching to passive during $[t_g + (1 + 1/\vartheta)T_1, t_g + R_1/\vartheta]$.*

Proof. All nodes in W switch to state *supp* \rightarrow *resync* during $(t_g, t_g + 2d)$ and switch to state *resync* when their timeout of $\vartheta 4d$ expires, which does not happen until time $t_g + 4d$. Once this timeout expired, they switch to state *passive* as soon as they observe themselves in state *resync*, i.e., by time $t_g + (4\vartheta + 3)d$. Hence, every node $i \in W$ does not observe itself in state *resync* within $[t_g + 3d, \tau_{i,i}(t_i))$, and therefore is in state *dormant* during $[t_g + 3d, \tau_{i,i}(t_i)]$. This implies that it observes itself in state *dormant* during $[t_g + 4d, \tau_{i,i}(t_i))$, completing the proof of Statement (i).

Moreover, from the definition of a good W -resynchronization point we have that no nodes from W are in state *join* at times in $[t_g - T_1 - d, t_{\text{join}})$. Statement (ii) follows, as every node from W resets its *join* flags upon switching to state *passive* at time t_i .

Regarding Statement (iii), observe first that no nodes from W are in state *sleep* \rightarrow *waking* during $(t_g - d, t_g + (1 + 1/\vartheta)T_1)$ for the following reason: By definition of a good W -resynchronization point no node from W switches to *sleep* during $(t_g - \Delta_g, t_g) \supseteq (t_g - (\vartheta + 1)T_1 - 3d, t_g)$. Any node in W that is in states *sleep* or *sleep* \rightarrow *waking* at time $t_g - (\vartheta + 1)T_1 - 3d$ switches to state *waking* before time $t_g - d$ due to timeouts. Finally, any node in W switching to *sleep* at or after time t_g will not switch to state *sleep* \rightarrow *waking* before time $t_g + (1 + 1/\vartheta)T_1$. The observation follows.

Since nodes in W reset their *sleep* \rightarrow *waking* flags at some time from

$$[t_i, \tau_{i,i}(t_i)] \subset (t_g + 3d, t_g + (4\vartheta + 4)d) \stackrel{(2)}{\subseteq} (t_g + 3d, t_g + (1 + 1/\vartheta)T_1),$$

Statement (iii) follows.

Statements (iv) and (v) follow from the fact that all nodes in W switch to state *passive* until time

$$t_g + (3 + 4\vartheta)d \stackrel{(2)}{\leq} t_g + \left(1 + \frac{1}{\vartheta}\right)T_1 - d,$$

while timeout $(R_1, \text{supp} \rightarrow \text{resync})$ must expire first in order to switch to *dormant* and subsequently *passive* again. \square

Before we proceed, in the next lemma we make the basic yet crucial observation that after a good W -resynchronization point t_g , no node from W will switch to state *join* until either time $t_g + T_7/\vartheta + 4d$ or T_6/ϑ time after the first non-faulty node switched to *sleep* \rightarrow *waking* again after t_g . By proper choice of T_6 and $T_7 > T_6$, this will guarantee that nodes from W do not switch to *join* prematurely during the final steps of the stabilization process.

Lemma 4.11. *Suppose t_g is a good W -resynchronization point. Denote by t_s the infimum of times greater than t_g when a node in W switches to state *sleep* \rightarrow *waking* and by t_{join} the infimum of times greater than $t_g - T_1 - d$ when a node in W switches to state *join*. Define $t^+ := t_g + \Delta_s - \Delta_g + \tilde{\delta}_s + T_2 + T_4 + T_5 + d$. Then, starting from time $t_g + 4d$, $\text{tr}(\text{recover}, \text{join})$ is not satisfied at any node in W until time*

$$\min \left\{ t_s + \frac{T_6}{\vartheta}, t_g + \frac{T_7}{\vartheta} + 4d \right\} \geq \min\{t_s + \Delta_s, t^+\}$$

and t_{join} is larger than this time.

Proof. By Statements (ii) and (iii) of Lemma 4.10 and Inequality (2), we have that $t_s \geq t_g + T_1 + 4d \geq t_g + (4\vartheta + 4)d$ and $t_{\text{join}} \geq t_g + 4d$. Consider a node $i \in W$ not observing itself in state *dormant* at some time $t \in [t_g + 4d, t_{\text{join}}]$. According to Statements (i) and (ii) of Lemma 4.10, the threshold condition of $f+1$ nodes memorized in state *join* cannot be satisfied at such a node. By statements (i) and (iii) of the lemma, the threshold condition of $f+1$ nodes memorized in state *sleep* \rightarrow *waking* cannot be satisfied unless $t > t_s$. Hence, if at time t a node from W satisfies that it observes itself in state *active* and T_6 expired, we have that $t > t_s + T_6/\vartheta$. Moreover, by Statement (i) of Lemma 4.10, we have that if T_7 is expired at any node in W at time t , it holds that $t > t_g + T_7/\vartheta + 4d$. Altogether, we conclude that $\text{tr}(\text{recover}, \text{join})$ is not satisfied at any node in W during

$$\left[t_g + 4d, \min \left\{ t_s + \frac{T_6}{\vartheta}, t_g + \frac{T_7}{\vartheta} + 4d \right\} \right] \stackrel{(7,8)}{\supseteq} [t_g + 4d, \min\{t_s + \Delta_s, t^+\}].$$

In particular, t_{join} must be larger than the upper boundary of this interval, concluding the proof. \square

Before we can move on to proving eventual stabilization, we need one last key lemma. Essentially, it states that after a good W -resynchronization point, any node in W switches to *recover* or to *sleep* \rightarrow *waking* within bounded time, and all nodes in W doing the latter will do so in rough synchrony, i.e., within a time window of $\tilde{\delta}_s$. Using the previous lemma, we can show that this happens before the transition to *join* is enabled for any node.

Lemma 4.12. *Suppose t_g is a good W -resynchronization point and use the notation of Lemma 4.11. Then either*

- (i) $t_s < t^+ - \Delta_s$ and any node in W switches to state *sleep* \rightarrow *waking* at some time in $[t_s, t_s + \tilde{\delta}_s]$ or is observed in state *recover* during $[t_s + T_1 + T_5, t_{\text{join}}]$ or
- (ii) all nodes in W are observed in state *recover* during $[t^+, t_{\text{join}}]$.

Proof. By Lemma 4.11, it holds that

$$t_{\text{join}} > \min\{t_s + \Delta_s, t^+\}. \tag{26}$$

For any node in W , consider the supremum t of all times smaller or equal to $t_g - \Delta_g$ when it switched to *sleep*. After that, it observed itself in state *waking* before time

$$t + (\vartheta + 1)T_1 + 3d \leq t_g - T_1 - d \quad (27)$$

(w.l.o.g. assuming that the node has ever been in state *sleep* since it became non-faulty). By definition of a good W -resynchronization point, nodes in W are not in state *join* during $(t_g - T_1 - d, t_{join})$ and do not switch to state *sleep* during $(t_g - \Delta_g, t_s)$. Continuing to execute the basic cycle after time $t_g - d > t_g - T_1 - d$ thus necessitates that the node is in one of the states *waking*, *ready*, *propose* or *accept* at time $t_g - d$.

Assume that it is in state *waking* (we just showed that if not, it already was in *waking* by time $t_g - d$). As timeout T_2 cannot have been reset later than time $t - T_1/\vartheta + d \leq t_g - \Delta_g - T_1/\vartheta + d$ at the respective node, it observes itself in state *ready* by time $t_g - \Delta_g - T_1/\vartheta + T_2 + 2d$, in state *propose* by time $t_g - \Delta_g - T_1/\vartheta + T_2 + T_4 + 3d$, in state *accept* by time $t_g - \Delta_g - T_1/\vartheta + T_2 + T_4 + T_5 + 4d$, in state *sleep* by time $t_g - \Delta_g + (1 - 1/\vartheta)T_1 + T_2 + T_4 + T_5 + 5d$, and must switch to *sleep* \rightarrow *waking* before time $t^+ - \Delta_s$.

We next distinguish between two cases:

Case 1: Assume that $t_s < t^+ - \Delta_s$. We already established that no node in W observes itself in states *sleep* or *sleep* \rightarrow *waking* at time t_s , and by Inequality (27), any node in W observing itself in states *waking* or *ready* reset its *accept* flags after time $t_g - T_1 - d$. Denote by $t'_s \in (t_s - (\vartheta + 1)T_1 - d, t_s - (1 + 1/\vartheta)T_1)$ the minimal time greater or equal to t_g when a node from W switches to state *sleep*; by the timeout condition for switching from *sleep* to *sleep* \rightarrow *waking* and the definitions of t_s and good W -resynchronization points, such a time exists. According to Lemma 4.5, at least $f + 1$ nodes have been in state *accept* at times in $(t'_s - T_1 - d, t'_s)$. By Statements (i) and (iii) of Lemma 4.10, all nodes are in state *passive* until at least time t_s . Hence, any nodes from W observing themselves in state *waking* or *ready* at time $t'_s + d$ satisfy $tr(waking, recover)$ or $tr(unsuspect, suspect)$, respectively. Consequently, they will leave these states no later than time

$$t'_s + (2\vartheta + 2)d \leq t_s - \left(1 + \frac{1}{\vartheta}\right)T_1 + (2\vartheta + 2)d \stackrel{(2)}{\leq} t_s - 4d.$$

It follows that any nodes from W that are in state *propose* at time t_s observe themselves in this state since at least time $t_s - 3d$, implying that they switch to states *accept* or *recover* by time $t_s + T_5 - 3d$. After switching to *accept*, a node from W switches to *sleep* and subsequently to *sleep* \rightarrow *waking* within another $(2\vartheta + 1)T_1 + 2d$ or is observed in state *recover* after less than $T_1 + 2d$ time. Thus, as

$$t_{join} > t_s + \Delta_s - (\vartheta - 1/\vartheta)T_1 - d \stackrel{(3)}{>} t_s + T_1 + T_5 - d,$$

all nodes in W that do not switch to state *sleep* \rightarrow *waking* during

$$[t_s, t_s + (2\vartheta + 1)T_1 + T_5 - d] \stackrel{(3)}{\subseteq} \left[t_s, t_s + \Delta_s - \left(\vartheta - \frac{1}{\vartheta}\right)T_1 - d \right] \subseteq \left[t_s, t'_s + \Delta_s + \left(1 + \frac{1}{\vartheta}\right)T_1 \right]$$

are observed in state *recover* at time $t_s + T_1 + T_5$. Because $t_{join} > t_s + \Delta_s - (\vartheta - 1/\vartheta)T_1 - d$ and no nodes from W switch to state *sleep* during $(t_g - \Delta_g, t_s)$, we can apply Statement (ii) of Corollary 4.6 to conclude that no nodes from W switch to state *sleep* \rightarrow *waking* during

$$\left(t_s + \tilde{\delta}_s, t'_s + \Delta_s + \left(1 + \frac{1}{\vartheta}\right)T_1 \right],$$

i.e., any node from W that does not switch to state *sleep* \rightarrow *waking* during $[t_s, t_s + \tilde{\delta}_s]$ is observed in state *recover* during $[t_s + T_1 + T_5, t_{join}]$. Statement (i) follows.

Case 2: Assume $t_s \geq t^+ - \Delta_s$. Then by Inequality (26), $t_{join} \geq t^+$ holds. By definition of t_s , the first node in W switching to *sleep* \rightarrow *waking* after t_g does so at time t_s , and by the arguments given above, no node from W executing the basic cycle does so later than $t^+ - \Delta_s < t^+ - d$. Hence, it is observed in state *recover* during $[t^+, t_{join}]$, as it cannot leave *recover* through *join* before time t_{join} . Hence Statement (ii) holds and the proof concludes. \square

We have everything in place for proving that a good resynchronization point leads to stabilization within $R_1/\vartheta - 3d$ time.

Theorem 4.13. *Suppose t_g is a good W -resynchronization point. Then there is a quasi-stabilization point during $(t_g, t_g + R_1/\vartheta - 3d]$.*

Proof. For simplicity, assume during this proof that $R_1 = \infty$, i.e., by Statement (i) of Lemma 4.10 all nodes in W observe themselves in states *passive* or *active* at times greater or equal to $t_g + (4\vartheta + 4)d$. We will establish the existence of a quasi-stabilization point at a time larger than t_g and show that it is upper bounded by $t_g + R_1/\vartheta - 3d$. Hence this assumption can be made w.l.o.g., as the existence of the quasi-stabilization point depends on the execution up to time $t_g + R_1/\vartheta$ only, and R_1 cannot expire before this time at any node in W . We use the notation of Lemma 4.11. By Statements (ii) of Lemma 4.10 and Inequality (2), we have that $t_s \geq t_g + T_1 + 4d \geq t_g + (4\vartheta + 4)d$. By Lemma 4.11, it holds that $t_{join} > \min\{t_s + \Delta_s, t^+\}$. We differentiate several cases.

Case 1: Assume $t_s \geq t^+ - \Delta_s$. According to Lemma 4.10, all nodes in W switched to state *passive* during $(t_g + 4d, t_g + (3 + 4\vartheta)d)$, implying that at any node in W , T_7 will expire at some time from $(t_g + T_7/\vartheta + 4d, t_g + T_7 + (4\vartheta + 4)d)$. By Lemma 4.12 we have that all non-faulty nodes are observed in state *recover* during $[t^+, t_{join}]$. By Statement (v) of Lemma 4.10, no node in W resets its *join* flags after time t^+ before it switches to state *propose*, returning to the basic cycle. Thus, any node from W will switch to state *join* before time $t_g + T_7 + (4\vartheta + 4)d$ and switch to *propose* as soon as it memorizes all non-faulty nodes in state *join*. Denote by $t_p \in (t_g + T_7/\vartheta + 4d, t_g + T_7 + (4\vartheta + 5)d)$ the minimal time when a node from W switches from *join* to *propose*. Certainly, nodes in W do not switch from *waking* to *ready* during $(t_p, t_p + 2d)$ and therefore also not reset their *join* flags before time $t_p + 3d$. As nodes in W reset their *propose* and *accept* flags upon switching to state *join*, some node in W must memorize $n - 2f \geq f + 1$ non-faulty nodes in state *join* at time t_p . According to Statement (ii) of Lemma 4.10, these nodes must have switched to state *join* at or after time t_{join} . Hence, all nodes in W will memorize them in state *join* by time $t_p + d$ and thus have switched to state *join*. Hence, all nodes in W will switch to state *propose* before time $t_p + 2d$ and subsequently to state *accept* before time $t_p + 3d$, i.e. $t_p \leq t_g + T_7 + (4\vartheta + 5)d$ is a quasi-stabilization point.

Case 2a: Assume $t_s < t^+ - \Delta_s$ and $< f + 1$ nodes in W switch to *sleep* \rightarrow *waking* during $[t_s, t_s + \tilde{\delta}_s]$. We then have that $t^+ \stackrel{(3)}{>} t_s + T_1 + T_5$. According to Lemma 4.12, any node in W that does not switch to state *sleep* \rightarrow *waking* is observed in state *recover* during $[t_s + T_1 + T_5, t_{join}]$. Thus, any node in W will observe at least $n - 2f \geq f + 1$ nodes from W in state *recover* during $[t_s + T_1 + T_5, t_{join}]$. As nodes in W reset their *propose* flags when switching to state *ready* and

$$t_s + T_1 + T_5 \stackrel{(3,4)}{\leq} t_s + \frac{T_2 + T_3}{\vartheta} - (\vartheta + 2)T_1 - (2\vartheta + 4)d,$$

a node from W switching to state *sleep* \rightarrow *waking* at or after time t_s cannot switch to *propose* via states *waking* and *ready* before time $t_s + T_1 + T_5 + (2\vartheta + 1)d$. Any node in W switching to state *sleep* \rightarrow *waking* during $[t_s, t_s + \tilde{\delta}_s]$ will observe itself in state *waking* before time

$$t_s + \tilde{\delta}_s + 2d \stackrel{(2,3)}{\leq} t_s + \Delta_s - (2\vartheta + 2)d \leq t^+ - (2\vartheta + 2)d.$$

By Lemma 4.11, $tr(recover, join)$ cannot be satisfied at any node in W until time $\min\{t_s + \Delta_s, t^+\}$. Thus, we have that no node from W switches from *ready* to *join* during $[t_s, t_{join})$ by definition of t_{join} and any node in W that observes itself in states *ready* and *suspect* will switch to state *recover* once $(2\vartheta d, suspect)$ expires. In summary, any node in W switching to state *sleep* \rightarrow *waking* at some time in $[t_s, t_s + \tilde{\delta}_s]$ will switch from *waking* to *recover* or from *unsuspect* to *suspect* by time $t_s + \Delta_s - (2\vartheta + 2)d$, and in the latter case it cannot leave state *ready* before switching to state *recover* due to $tr(ready, recover)$ being satisfied. As the latter happens before time $t_s + \Delta_s - d < t_{join} - d$, all nodes in W are observed in state *recover* during $[t_s + \Delta_s, t_{join}]$. From here we can argue analogously to the first case, i.e., there exists a quasi-stabilization point $t_p \leq t_g + T_7 + (4\vartheta + 5)d$.

Case 2b: Assume $t_s < t^+ - \Delta_s$ and $\geq f + 1$ nodes in W switch to *sleep* \rightarrow *waking* during $[t_s, t_s + \tilde{\delta}_s]$. By Statements (ii) and (iv) of Lemma 4.10, no node from W resets its *sleep* \rightarrow *waking* flags at or after time $t_s \geq t_g + (1 + 1/\vartheta)T_1$. Hence, by Statement (i) of the lemma, all nodes in W switch to *active* during $(t_s, t_s + \tilde{\delta}_s + d)$. Between T_6/ϑ and $T_6 + d$ time later T_6 will expire. We have that

$$t_s + \frac{T_6}{\vartheta} < t^+ - \Delta_s + \frac{T_6}{\vartheta} \stackrel{(8)}{\leq} t_g + \frac{T_7}{\vartheta} + 4d.$$

Thus, according to Lemma 4.11, $t_{join} > t_s + T_6/\vartheta$. On the other hand, at the latest once T_6 expires, $tr(recover, join)$ holds at every node.

By time

$$t_s + \frac{T_6}{\vartheta} \stackrel{(7)}{\geq} t_s + \tilde{\delta}_s - \left(1 - \frac{1}{\vartheta}\right) T_1 + T_2 + 2d,$$

the nodes in W that switched to state *sleep* \rightarrow *waking* observe themselves in state *ready* because of timeouts or are in state *recover*. By Statement (v) of Lemma 4.10, after this time no node in W resets its *join* flags again before it runs through the basic cycle again and switches to state *ready*.

Hence, all nodes in W will switch to states *join* or *propose* until time

$$\begin{aligned} & \max \left\{ t_s + \tilde{\delta}_s - \left(1 + \frac{1}{\vartheta}\right) T_1 + T_2 + T_4 + 2d, t_s + \tilde{\delta}_s + T_6 + 3d \right\} + d \\ & \stackrel{(4,5)}{=} t_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right) T_1 + T_2 + T_4 + 7d, \end{aligned}$$

where we accounted for an additional delay of d due to a possible transition from *ready* to *recover* just before time $t_s + \tilde{\delta}_s + T_6 + 2d$ and, if no node from W switches from *ready* to *join*, all nodes in W needing to be observed in state *join* for a node in W to switch to state *propose*. It follows that a minimal time $t_p \in (t_s + T_6/\vartheta, t_s + (\vartheta + 1 - 2/\vartheta)T_1 + T_2 + T_4 + 7d)$ exists when a node from W switches to state *propose*. Again, we distinguish two cases.

Case 2b-I: Assume that some node in W switches from state *join* to state *propose* at time t_p . Thus, there must be at least $n - 2f \geq f + 1$ non-faulty nodes in state *join* at time $t_p - \varepsilon$ (for some arbitrarily small $\varepsilon > 0$), as any *propose* or *accept* flag corresponding to a non-faulty node has been

reset at a time t satisfying that the respective node has not been observed in one of these states during $[t, t_p]$. Thus, all nodes in W will switch to states *join* or *propose* before time $t_p + d$. At time $t_p + 2d$, they will observe all non-faulty nodes in one of the states *join*, *propose*, or *accept*, i.e., they switch to state *propose* before time $t_p + 2d$. Finally, they will observe all non-faulty nodes in states *propose* or *accept* before time $t_p + 3d < t_p + T_1/\vartheta$ and switch to state *accept*. As t_p is minimal, we conclude that all nodes in W switched to state *accept* during $(t_p, t_p + 3d)$, i.e., t_p is a quasi-stabilization point.

Case 2b-II: Otherwise, some node in W switched from state *ready* to state *propose* at time t_p . As we have that

$$t_s + \tilde{\delta}_s + T_6 + 4d \stackrel{(4)}{\leq} t_s - (\vartheta + 1)T_1 + \frac{T_2 + T_3}{\vartheta} - d,$$

T_6 is expired at all nodes in W since time $t_p - 2d$, i.e., $tr(recover, join)$ is satisfied at all nodes in W since time $t_p - 2d$. Hence, all nodes in W are observed in states *ready* or *join* at time t_p , and no node from W may switch to state *recover* again or reset its *propose* flags before switching to *resync* or *accept* first after time t_p .

Denote by t_a the infimum of times greater than t_p when a node from W switches to *accept* and assume for the moment that no node from W may switch from *propose* to *recover* before switching to *accept* first after time t_p . As nodes in W reset their *propose* flags upon switching to states *ready* or *join*, there must be $n - 2f \geq f + 1$ non-faulty nodes that switched to state *propose* during $[t_p, t_a)$ (unless $t_a = \infty$, which will be ruled out shortly). Thus, all nodes in W leave state *ready* before time $t_a + d$, and are observed in states *propose* or *join* before time $t_a + 2d$. Recalling that all nodes in W switch to states *join* or *propose* until time

$$t_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right)T_1 + T_2 + T_4 + 7d,$$

we get that indeed all nodes in W are observed in one of these states after time t_p and before time

$$\min \left\{ t_a + 2d, t_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right)T_1 + T_2 + T_4 + 8d \right\}.$$

Thus, at any node from W , $tr(join, propose)$ will be satisfied before this time, and it will be observed in state *propose* less than d time later. It follows that all nodes in W switch to state *accept* before time

$$t_q + 3d := \min \left\{ t_a + 3d, t_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right)T_1 + T_2 + T_4 + 9d \right\},$$

i.e., t_q is a quasi-stabilization point. As we made the assumption that no node from W switches from *propose* to *recover* before switching to *accept*, we need to show that T_5 does not expire at any node from W in state *propose* until time $t_q + 3d$. This holds true because

$$t_p + \frac{T_5}{\vartheta} > t_s + \frac{T_5 + T_6}{\vartheta} \stackrel{(6)}{\geq} t_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right)T_1 + T_2 + T_4 + 9d \geq t_q + 3d.$$

It remains to check that in all cases, the obtained quasi-synchronisation point t_q occurs no later than time $t_g + R_1/\vartheta - 3d$. In Cases 1 and 2a, we have that

$$t_q \leq t_g + T_7 + (4\vartheta + 5)d \stackrel{(9)}{\leq} t_g + \frac{R_1}{\vartheta} - 3d.$$

In Case 2b, it holds that

$$\begin{aligned}
t_q &\leq t_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right) T_1 + T_2 + T_4 + 9d \\
&\leq t^+ - \Delta_s + \left(\vartheta + 1 - \frac{2}{\vartheta}\right) T_1 + T_2 + T_4 + 9d \\
&\stackrel{(9)}{\leq} t_g + \frac{R_1}{\vartheta} - 3d.
\end{aligned}$$

We conclude that indeed all nodes in W switch to *accept* within a window of less than $3d$ time before at any node in W , R_1 expires and it leaves state *resync*, concluding the proof. \square

Finally, putting together our main theorems and Lemma 3.4, we deduce that the system will stabilize from an arbitrary initial state provided that a subset of $n - f$ nodes remains coherent for a sufficiently large period of time.

Corollary 4.14. *Suppose that $\vartheta < \vartheta_{\max} \approx 1.247$ as given in Lemma 3.4. Let $W \subseteq V$, where $|W| \geq n - f$, and define for any $k \in \mathbb{N}$*

$$T(k) := (k + 2)(\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2 + d) + R_1/\vartheta.$$

Then, for any $k \in \mathbb{N}$, the proposed algorithm is a (W, W^2) -stabilizing pulse synchronization protocol with skew $2d$ and accuracy bounds $(T_2 + T_3)/\vartheta - 2d$ and $T_2 + T_4 + 7d$ stabilizing within time $T(k)$ with probability at least $1 - 1/2^{k(n-f)}$. It is feasible to pick timeouts such that $T(k) \in \mathcal{O}(kn)$ and $T_2 + T_4 + 7d \in \mathcal{O}(1)$.

Proof. The satisfiability of Condition 3.3 with $T(k) \in \mathcal{O}(kn)$ and $T_2 + T_4 + 7d \in \mathcal{O}(1)$ follows from Lemma 3.4. Assume that t^+ is sufficiently large for $[t^- + T(k) + 2d, t^+]$ to be non-empty, as otherwise nothing is to show. By definition, W will be coherent during $[t_c^-, t^+]$, with $t_c^- = t^- + \vartheta(R_2 + 3d) + 8(1 - \lambda)R_2 + d$. According to Theorem 4.9, there will be some good W -resynchronization point $t_g \in [t_c^-, t_c^- + (k + 1)(\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2 + d)]$ with probability at least $1 - 1/2^{k(n-f)}$. If this is the case, Theorem 4.13 shows that there is a W -stabilization point $t \in [t_g, t^- + T(k)]$. Applying Theorem 4.4 inductively, we derive that the algorithm is a (W, E) -stabilizing pulse synchronization protocol with the bounds as stated in the corollary that stabilizes within time $T(k)$ with probability at least $1 - 1/2^{k(n-f)}$. \square

5 Generalizations and Extensions

5.1 Synchronization Despite Faulty Channels

Theorem 4.13 and our notion of coherency require that all involved nodes are connected by correct channels only. However, it is desirable that non-faulty nodes synchronize even if they are not connected by correct channels. To capture this, the notions of coherency and stability can be generalized as follows.

Definition 5.1 (Weak Coherency). *We call the set $C \subseteq V$ weakly coherent during $[t^-, t^+]$, iff for any node $i \in C$ there is a subset $C' \subseteq C$ that contains i , has size $n - f$, and is coherent during $[t^-, t^+]$.*

In particular, if there are in total at most f nodes that are faulty or have faulty outgoing channels, then the set of non-faulty nodes is (after some amount of time) weakly coherent.

Corollary 5.2. *For each $k \in \mathbb{N}$ let $T'(k) := T(k) - ((\vartheta(R_2 + 3d) + 8(1 - \lambda)R_2 + d))$, where $T(k)$ is defined as in Corollary 4.14. Suppose the subset of nodes $C \subseteq V$ is weakly coherent during the time interval $[t^-, t^+] \supseteq [t^- + T'(k) + T_2 + T_4 + 8d, t^+] \neq \emptyset$. Then, with probability at least $1 - (f + 1)/2^{k(n-f)}$, there is a C -quasi-stabilization point $t \leq t^- + T'(k) + T_2 + T_4 + 5d$ such that the system is weakly C -coherent during $[t, t^+]$.*

Proof. By the definition of weak coherency, every node in C is in some coherent set $C' \subseteq C$ of size $n - f$. Hence, for any such C' it holds that we can cover all nodes in C by at most $1 + |V \setminus C'| \leq f + 1$ coherent sets $C_1, \dots, C_{f+1} \subseteq C$. By Corollary 4.14 and the union bound, with probability at least $1 - (f + 1)/2^{k(n-f)}$, for each of these sets there will be at least one stabilization point during $[t^-, t^- + T'(k) - (T_2 + T_4 + 5d)]$. Assuming that this is indeed true, denote by $t_{i_0} \in [t^-, t^- + T'(k) - (T_2 + T_4 + 5d)]$ the time

$$\max_{i \in \{1, \dots, f+1\}} \{\max\{t \leq t^- + T'(k) - (T_2 + T_4 + 5d) \mid t \text{ is a } C_i\text{-stabilization point}\}\},$$

where $i_0 \in \{1, \dots, f+1\}$ is an index for which the first maximum is attained and t_{i_0} is the respective maximal time, i.e., t_{i_0} is a C_{i_0} -stabilization point.

Define $t'_{i_0} \in (t_{i_0}, t^- + T'(k)]$ to be minimal such that it is another C_{i_0} -stabilization point. Such a time must exist by Theorem 4.4. Since the theorem also states that no node from C_{i_0} switches to state *accept* during $[t_{i_0} + 2d, t'_{i_0})$ and $C_i \cap C_{i_0} \neq \emptyset$, there can be no C_i -stabilization point during $(t_{i_0} + 2d, t'_{i_0} - 2d)$ for any $i \in \{1, \dots, f+1\}$. Applying the theorem once more, we see that there are also no C_i -stabilization points during $(t'_{i_0} + 2d, t'_{i_0} + (T_2 + T_3)/\vartheta - 2d)$ for any $i \in \{1, \dots, f+1\}$. On the other hand, the maximality of t_{i_0} implies that every C_i had a stabilization point by time t_{i_0} . Applying Theorem 4.4 to the latest stabilization point until time t_{i_0} for each C_i , we see that it must have another stabilization point before time $t_{i_0} + T_2 + T_4 + 5d$. We have that

$$\frac{2(T_2 + T_3)}{\vartheta} - 2d \stackrel{(3)}{>} \frac{T_2 + T_3 + T_5}{\vartheta} \stackrel{(6)}{>} T_2 + T_4 + 5d,$$

i.e., all C_i have stabilization points within a short time interval of $(t'_{i_0} - 2d, t'_{i_0} + 2d)$. Arguing analogously about the previous stabilization points of the sets C_i (which exist because t_{i_0} is maximal), we infer that all C_i had their previous stabilization point during $(t_{i_0} - 2d, t_{i_0} + 2d)$.

Now suppose t_a is the minimal time in $(t'_{i_0} - 2d, t'_{i_0} + 2d)$ when a node from C switches to *accept* and this node is in set C_i for some $i \in \{1, \dots, f+1\}$. As usual, there must be at least $f + 1$ non-faulty nodes from C_i in state *propose* at time t_a and by time $t_a + d$ all nodes from C_i will be in either of the states *propose* or *accept*. As $|C_i \cap C_j| \geq f + 1$ for any $j \in \{1, \dots, f+1\}$ all nodes in C_j will observe $f + 1$ nodes in state *propose* at times in $(t_a, t_a + 2d)$. We have that $t_a \geq t_{i_0} + (T_2 + T_3)/\vartheta - 2d$ according to Theorem 4.4. As no nodes switched to state *accept* during $(t_{i_0} + 2d, t_a)$ and none of them switch to state *recover* (cf. Theorem 4.4), it follows from the Inequality

$$(T_2 + T_3)/\vartheta - 4d \stackrel{(4)}{>} T_2 + 2T_1 \stackrel{(2)}{>} T_2 + 5d$$

that all nodes from C_j observe themselves in one of the states *ready* or *propose* at time t_a . Hence, they will switch from *ready* to *propose* if they still are in *ready* before time $t_a + 2d$. Less than d time later, all nodes in C_j will memorize C_j in state *propose* and therefore switch to *accept* if not done so yet. Since j was arbitrary, it follows that t_a is a C -quasi-stabilization point. \square

Corollary 5.3. *Suppose C is weakly coherent during $[t^-, t^+]$ and $t \in [t^-, t^+ - (T_2 + T_4 + 8d)]$ is a C -quasi-stabilization point. Then*

- (i) *all nodes from C switch to accept exactly once within $[t, t + 3d)$ and*
- (ii) *there will be a C -quasi-stabilization point $t' \in [t + (T_2 + T_3)/\vartheta, t + T_2 + T_4 + 5d)$ satisfying that no nodes switch to accept in the time interval $[t + 3d, t')$*
- (iii) *and each node i 's, $i \in W$, state of the basic cycle (Figure 1) is metastability-free during $[t + 4d, t' + 4d)$*

Proof. Analogously to the proofs of Theorem 4.4 and Corollary 5.2. □

We point out that one cannot get stronger results by the proposed technique. Even if there are merely $f + 1$ failing channels, this can e.g. effectively render a node faulty (as it may never see $n - f$ nodes in states *propose* or *accept*) or exclude the existence of a coherent set of size $n - f$ (if the channels connect $f + 1$ disjoint pairs of nodes, there can be no subset of $n - f$ nodes whose induced subgraph contains correct channels only). Stronger resilience to channel faults would necessitate to propagate information over several hops in a fault-tolerant manner, imposing larger bounds on timeouts and weaker synchronization guarantees.

Combination of Corollary 5.2 and Corollary 5.3 finally yields:

Corollary 5.4. *Suppose that $\vartheta < \vartheta_{\max} \approx 1.247$ as given in Lemma 3.4. Let $C \subseteq V$ be such that, for each $i \in C$, there is a set $C_i \subseteq C$ with $|C_i| = n - f$, and let $E = \bigcup_{i \in C} C_i^2$. Then the proposed algorithm is a (C, E) -stabilizing pulse synchronization protocol with skew $3d$ and accuracy bounds $(T_2 + T_3)/\vartheta - 3d$ and $T_2 + T_4 + 8d$ stabilizing within time $T(k) + T_2 + T_4 + 5d$ with probability at least $1 - (f + 1)/2^{k(n-f)}$, for any $k \in \mathbb{N}$.*

Proof. Analogously to the proof of Corollary 4.14 □

5.2 Late Joining and Fast Recovery

An important aspect of combining self-stabilization with Byzantine fault-tolerance is that the system can remain operational when facing a limited number of transient faults. If the affected components stabilize quickly enough, this can prevent future faults from causing system failure. In an environment where transient faults occur according to a random distribution that is not too far from being uniform (i.e., one deals not primarily with bursts), the mean time until failure is therefore determined by the time it takes to recover from transient faults. Thus, it is of significant interest that a node that starts functioning according to the specifications again synchronizes as fast as possible to an existing subset of correct nodes making a quasi-stabilization point. Moreover, it is of interest that a node that has been shut down temporarily, e.g. for maintenance, can join the operational system again quickly.

In the presented form, the algorithm suffers from the drawback that a node in state *recover* may be caught there until the next good resynchronization point. Since Byzantine faults of a certain pattern may deterministically delay this for $\Omega(n)$ time, we would like to modify the algorithm in a way ensuring that a non-faulty node can synchronize to others more quickly if a quasi-stabilization point is reached.

This can be done in a simple manner. Whenever a node switches to state *none*, it stays until a new timeout (R_1, none) expires. When switching to *none*, it switches also to *passive*, resets its

join and *sleep* \rightarrow *waking* flags, and repeats to reset its *sleep* \rightarrow *waking* flags whenever a timeout of $(\vartheta - 1)(\vartheta + 2)T_1 + \vartheta 5d$ expires. Thus, it will not switch to state *active* because of outdated information. On the other hand, it will not miss the next occurrence of a set C , that are weakly coherent since a C -quasi-stabilization point, switching to state *sleep* \rightarrow *waking* within a time window of $(1 - 1/\vartheta)(\vartheta + 2)T_1 + \vartheta 5d$, as it will reset its *sleep* \rightarrow *waking* flags at most once in this window, whereas $|C| \geq n - f \geq 2f$. Subsequently, it will switch to state *join* at an appropriate time to enter the basic cycle again at the occurrence of the next C -stabilization point. Since nodes refrain from leaving state *none* for a constant period of time only, this way stabilization time in face of severe failures can still be kept linear, while in a stable system, nodes recovering from faults or joining late stabilize in constant time.

Corollary 5.5. *The pulse synchronization routine can be modified such that it retains all shown properties, \hat{E}_3 increases by a constant factor, and it holds that, for any node i in V , if there is a C -quasi-stabilization point at some time $t < t^-$, so that C is weakly coherent during $[t, t^+]$, and $(C \cup \{i\})$ -coherent during $[t^-, t^+]$, then there exists a $(C \cup \{i\})$ -quasi-stabilization point at some time $t' \leq t^- + \mathcal{O}(1)$, so that $(C \cup \{i\})$ is weakly coherent during $[t', t^+]$.*

Proof Sketch. Essentially, the fact that $n - f$ nodes continue to execute the basic cycle narrows down the possibilities in the proof of Theorem 4.13 to Case 2b-II, where the threshold for leaving state *join* will be achieved close to the next C -stabilization point due to the involved threshold conditions. Since the nodes in C execute the basic cycle, they are not affected by the re-synchronisation subroutine at all. Thus, v stabilizes independently of this subroutine provided that it resets its *join* and *sleep* \rightarrow *waking* flags in an appropriate fashion. We explained above how this is done and why a consistent reset of the *sleep* \rightarrow *waking* flags is achieved. The *join* flags are not an issue since at most $n - |C| \leq f$ channels can attain state *join*. As a node switches to state *none* again in constant time whenever it leaves, the node will stabilize in constant time provided that there is a C -quasi-stabilization point from where on C is weakly coherent until time t^+ . On the other hand, we can easily adapt the re-synchronisation subroutine, Lemma 4.8, Theorem 4.9, and Condition 3.3 to allow for the additional time nodes are non-responsive with respect to the re-synchronisation subroutine, increasing \hat{E}_3 by a constant factor only. \square

5.3 Stronger Adversary

So far, our analysis considered a fixed set C of coherent (or weakly coherent) nodes. But what happens if whether a node becomes faulty or not is not determined upfront, but depends on the execution? Phrased differently, does the algorithm still stabilize quickly with a large probability if an adversary may “corrupt” up to f nodes, but may decide on its choices as time progresses, fully aware of what happened so far? Since we operate in a system where all operations take positive time, it might even be the case that a node might fail just when it is about to perform a certain state transition, and would not have done so if the execution had proceeded differently. Due to the way we use randomization, this however makes little difference for the stabilization properties of the algorithm.

Corollary 5.6. *Suppose at every time t , an adversary has full knowledge of the state of the system up to and including time t , and it might decide on in total up to f nodes (or all channels originating from a node) becoming faulty at arbitrary times. If it picks a node at time t , it fully controls its actions after and including time t . Furthermore, it controls delays and clock drifts of non-faulty*

components within the system specifications, and it initializes the system in an arbitrary state at time 0. For any $k \in \mathbb{N}$, define

$$t_k := 2(k+2)(\vartheta(R_2 + 3d) + 8(1-\lambda)R_2 + d) + R_1/\vartheta + T_2 + T_4 + 5d.$$

Then the set of all non-faulty nodes have reached a quasi-stabilization point by time $T(k)$ from where on they are weakly coherent, with probability at least

$$1 - (f+1)e^{-k(n-f)/2}.$$

Proof. We need to show that Theorem 4.9 holds for the modified time interval $[t, t + (k+2)\hat{E}_3]$ with the modified probability of at least $1 - e^{-k(n-f)/2}$. If this is the case, we can proceed as in Corollaries 5.2 and 5.3.

We start to track the execution from time 0. Whenever a node switches to state *init* at a good time, the adversary must corrupt it in order to prevent subsequent deterministic stabilization. In the proof of Theorem 4.9, we showed that for any non-faulty node, there are at least $k+1$ different times until $2\vartheta(k+2)\hat{E}_3$ when it switches to *init* that have an independently by $1/2$ lower bounded probability to be good. Since Lemma 4.8 holds for *any* execution where we have at most f faults, the adversary corrupting some node at time t affects the current and future trials of that node only, while the statement still holds true for the non-corrupted nodes. Thus, the probability that the adversary may prevent the system from stabilizing until time t_k is upper bounded by the probability that $(k+1)(n-f)$ independent and unbiased coin flips show f or less times tail. Chernoff's bound states for the random variable X counting the number of tails in this random experiment that for any $\delta \in (0, 1)$,

$$P[X < (1-\delta)\mathbb{E}[X]] < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right)^{\mathbb{E}[X]} < e^{-\delta\mathbb{E}[X]}.$$

Inserting $\delta = k/(k+1)$ and $\mathbb{E}[X] = (k+1)(n-f)/2$, we see that the probability that

$$P[X \leq f] \leq P[X < (n-f)/2] < e^{-k(n-f)/2},$$

as claimed. □

6 Implementation Issues

In this section, we briefly survey some core aspects of the VLSI implementation of the pulse synchronization algorithm, which is currently being developed. Thereby we focus on the three major building blocks: (1) asynchronous state machines, (2) memory flags with thresholds and (3) watchdog timers.

The pulse synchronization algorithm at every node consists of several simple state machines that execute asynchronously and concurrently. There are several types of conditions that can trigger state transitions:

- (i) The state machines of a certain number ($1, \geq f+1$, or $\geq n-f$) of remote nodes reached some particular state, indicated by memory flags.
- (ii) Some local state machine reached a particular state.

(iii) A watchdog timer expires.

These conditions may also be combined (using AND or OR).

We will employ standard Huffman-type asynchronous state machines [25] for implementing our state machines, as they fit nicely to the Θ -Model already used in DARTS.¹⁹ Analyzing the transition conditions of all the five state machines (Figures 2, 3 and 4) of a single node reveals that we need to communicate six different states (*recover*, *accept*, *join*, *propose*, *sleep* \rightarrow *waking* and “other”) of the core state machine (Figure 2) and two states each (*supp*, *none* and *init*, *wait*) for the two state machines making up the resynchronization algorithm from every node to every node. There are several possibilities for implementing this communication. For example, both a simple high-speed serial protocol and a parallel five bit bundled data bus with a strobe signal are viable alternatives, each offering different trade-offs between implementation complexity, speed, area consumption, etc.

We note, however, that any method for communicating states is complicated by the fact that state occupancy times may be very short in an asynchronous state machine: Reaching a state must always be faithfully conveyed to all remote nodes even if it is almost immediately left again. In addition, the core state machine may undergo various sequences of state transitions, implying that we cannot use a state encoding where only a single bit changes between successive states. Care must hence be taken in order not to trigger hazardous intermediate state occupancies at the receiver when communicating some multi-bit state change. Both problems can be handled using suitable bounded delay conditions.

Remote Memory Flags and Thresholds

Figure 5 shows the principle of implementing remote memory flags, which are the basic mechanism required for type (i) state transition conditions at node i . For every remote node j , it consists of a hazard-free demultiplexer that decodes the communicated state of node j ’s state machines, a resettable memory flag per state that remembers whether node j has ever reached the respective state since the most recent flag reset, and optionally a threshold module that combines the corresponding flag outputs for all remote nodes. Note that every memory flag is implemented as a (resettable) Muller C Gate²⁰ here, but could also be built by using a flip-flop.

Implementing local state input transition conditions (ii) is pretty much straightforward, as one simply needs to incorporate (single) state signals from local state machines here. Note that every transition condition comprises the node observing itself in a particular state, which also falls into this category. To avoid metastable upsets in the asynchronous state machine (see below), it may be necessary to add memory flags for local signals as well.

Watchdog Timers

Our implementation of the watchdog timers, which are required for realizing state transition conditions (iii), will rest upon a single local clock generator (we will use a simple ring oscillator, i.e., a single inverter with feedback and a prescaler) per node that drives all watchdog timers, instead of a crystal oscillator, because of the possibility to integrate it on-chip. However, the oscillator frequency of such ring oscillators vary heavily with operating conditions, in particular with supply

¹⁹The Θ -Model assumes that we can enforce a certain ratio between slowest and fastest end-to-end delay along critical signaling paths.

²⁰A Muller C Gate retains its current output value when its inputs are different, and sets its output to the common input otherwise.

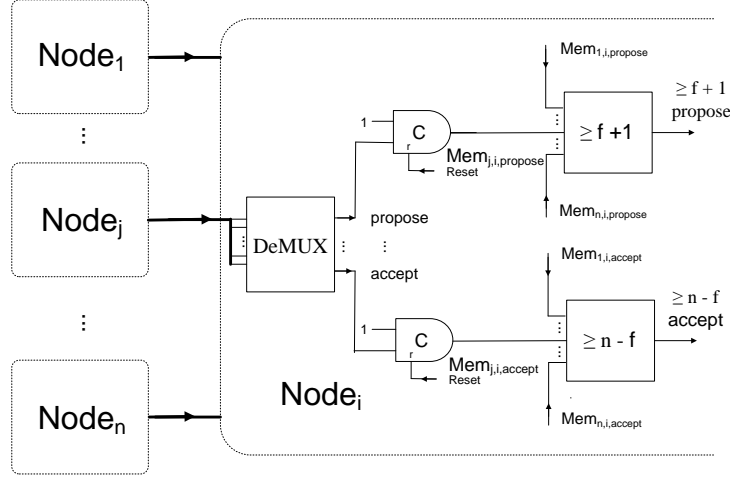


Figure 5: Implementation principle of remote memory flags and thresholds.

voltage and temperature, as well as with process conditions. The resulting (two-sided) clock drift ξ (with respect to supply voltage, temperature and process variation) is typically in the range of 7% to 9% for uncompensated ring oscillators and can be lowered down to 1% to 2% by proper compensation techniques [32]. The two-sided clock drifts map to $\vartheta = (1 + \xi)/(1 - \xi)$ bounds of 1.15 to 1.19 and 1.02 to 1.04, respectively. Recalling from Lemma 3.4 that $\vartheta_{\max} \approx 1.247$, one sees that both uncompensated and compensated ring oscillators are suitable for implementation of the pulse synchronization protocol's watchdog timers. However, care must be taken when the protocol is used to stabilize DARTS: to compensate a typical drift of 15% of DARTS clocks, one must ensure that ϑ is smaller than roughly 1.064 (cf. Section 7). Thus, here, only compensated ring oscillators are sufficiently accurate. Note, however, that these are conservative bounds, assuming that the synchronization protocol and DARTS drift into different directions. Considering that a large share of the drift in both systems is due to variations in temperature, it seems reasonable to assume that, in the long term, both drift into the same direction.

As shown in Figure 6, every watchdog timer consists of a resettable up-counter and a timeout register, which holds the timeout value. A comparator compares the counter value and the timeout register after every clock tick, and raises a stable expiration output signal if the counter value is greater or equal to the register value. The asynchronous reset of the counter, which also resets the timeout output signal, is used to re-trigger the watchdog.

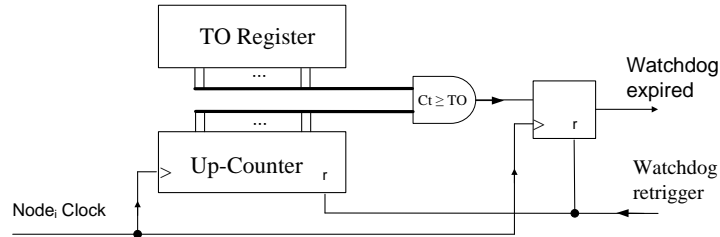


Figure 6: Implementation principle of watchdog timers.

As for the watchdog timer with random timeout R_3 in the resynchronization algorithm, the

simplest implementation would load a uniformly distributed random value into the timeout register whenever the watchdog is re-triggered. Depending on the implementation technology, such random values can be generated either via true random sources (thermal noise) or pseudo-random sources (LFSRs) clocked by another ring oscillator. If we could guarantee that the content of the timeout and the random source can, by no means, read or probed somehow by anybody, such an implementation satisfies the model requirements.²¹ Alternatively, one could use random sampling per clock tick, which avoids storing the future timeout value and also converges to uniformly distributed timeouts for sufficiently large values of R_3 .

Combined State Transition Conditions

Combining different state transition conditions (i)–(iii) via AND/OR requires some care, since an asynchronous state machine requires stable input signals in order not to become metastable during its state transition. Combining several conditions (i) does not cause any problems, since the memory flags ensure that all outputs are stable. Non-stable signals, like “ $T_1 \text{ AND } < n - f \text{ accept}$ ” require sampling via a flip-flop clocked by a stable signal. For example, the status of $< n - f = \neg \geq n - f$ is sampled when the signal reporting expiration of T_1 is issued. Similarly, it might happen that conditions requiring conflicting state transitions are satisfied at the same time, e.g., (T_2, accept) might expire simultaneously with the threshold of “ $\geq f + 1 \text{ recover or accept}$ ” being reached.

Obviously, both of the above situations could create a metastable upset, either of the sampling flip-flop, or directly of the register(s) holding the node’s state. Fortunately, Theorem 4.4 revealed that this can happen during stabilization only. In regular operation, e.g. the critical threshold of $\geq n - f \text{ accept}$ is always reached before T_1 expires. Thus, the former is acceptable, as metastable upsets occur rarely and increase convergence time only. Moreover, to further decrease the probability of a metastable upset that might affect stabilization time, it is perfectly feasible to insert a synchronizer or an elastic pipeline after the sampling flip-flop for capturing metastability [13]. This additional precaution merely increases the latency by a constant delay, which due to being restricted to the pulse synchronization component will not adversely affect the final precision and accuracy of the stabilized DARTS clocks.

7 Coupling of DARTS and Pulse Synchronization Algorithm

In this section, we describe how the self-stabilizing pulse synchronization protocol could be coupled with DARTS clocks. As this requires certain implementation details, we also sketch some ideas that might be used in a prototype implementation. The joint system provides a high-precision self-stabilizing Byzantine fault-tolerant clocking system for multi-synchronous GALS.

The coupling between the pulse synchronization protocol and DARTS clocks involves two directions:

1. The pulse synchronization protocol primarily monitors the operation of the DARTS clocks. As long as DARTS ticks are generated correctly, it must not interfere with the DARTS tick generation rules at all.

²¹Note that in practice this is a reasonable assumption, as even the node itself does not access this value except for checking whether the timer expired and the computational power of the system is very limited.

2. If DARTS clocks become inconsistent w.r.t. the behavior of the pulse synchronization protocol, the latter must interfere with the regular DARTS tick generation, possibly up to resetting DARTS clocks.

To assist the reader, we first provide a very brief overview of the original DARTS and its implementation.

7.1 DARTS Overview

DARTS clocks (called TG-Algs in the sequel) are instances of a simple synchronizer [35] for the Θ -Model based on consistent broadcasting [31]. They generate ticks $\text{Tick}(0)$, $\text{Tick}(1)$, $\text{Tick}(2)$, \dots approximately simultaneously at all correct nodes. Since actual DARTS ticks are just binary clock signal transitions, which cannot carry tick numbers, the original algorithm had to be modified significantly in order to be implementable in asynchronous digital logic. Figure 7 shows a schematic of a single TG-Alg for a 5-node system.

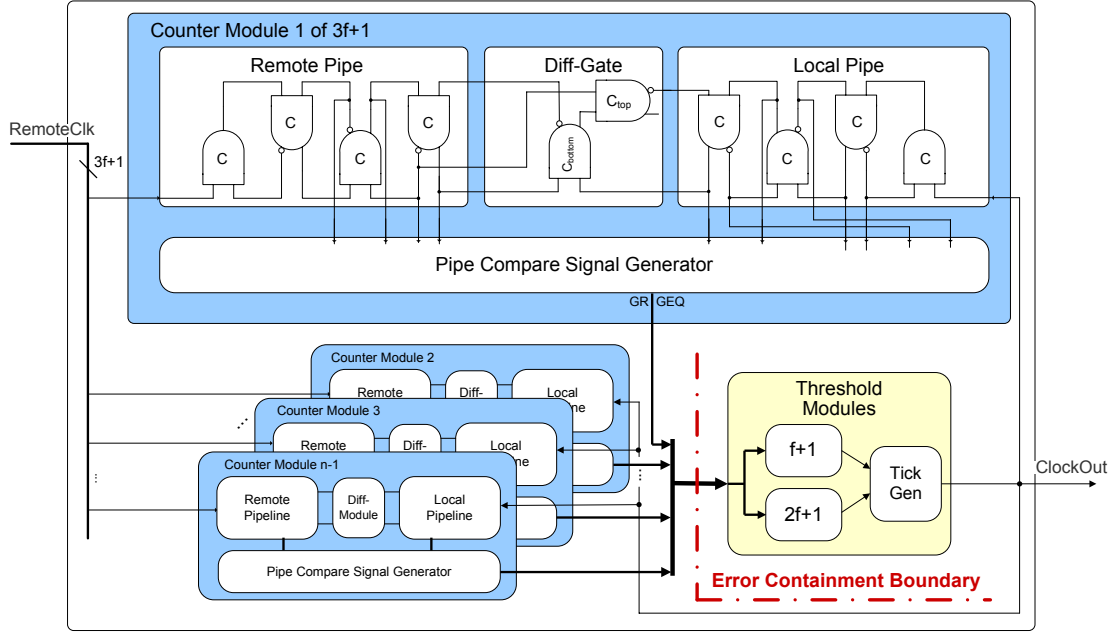


Figure 7: Schematic of DARTS TG-Alg Implementation

Key components of a TG-Alg are counter modules, one per remote TG-Alg, which just count the difference between the number of ticks generated locally and remotely. They are implemented using a pair of elastic pipelines [33], which implement FIFO buffers for signal transitions. Matching ticks in both pipelines, which are obviously irrelevant for the difference, are removed by the connecting Diff-Gate. The status (> 0 , ≥ 0) of all counter modules is fed into two threshold modules, whose outputs trigger the generation of the next local tick. A detailed discussion of the implementation can be found in [10].

The correctness proof and performance analysis in [16, 17, 15] revealed that correct TG-Algs indeed generate synchronized clock ticks, in the presence of up to f Byzantine faulty TG-Algs in a system with $n \geq 3f + 2$ nodes: For any two correct nodes p , q , the number of clock ticks

generated by p and q by time t do not differ by more than a (very small) constant π , and the frequency of any correct clock (and thus the maximum drift ρ) is within a certain range. In addition, expressions (in the order of π) for the maximum size of the elastic pipelines in the counter modules were established, which guarantee overflow-free operation. Experiments with both FPGA and ASIC prototype implementations demonstrated that DARTS clocks indeed offer close to perfect synchronization and very reliable operation.

Nevertheless, as already mentioned, (almost) simultaneous start-up of all TG-Algs and at most f failures during the whole life-time of the system are mandatory preconditions for these results to hold. DARTS neither supports late joining or recovery of TG-Algs, nor recovery from more than f failures.

7.2 Required Extensions for Coupling DARTS and Pulse Synchronization

The major obstacle for supporting late joining of TG-Algs, removing spuriously generated ticks in the pipelines etc. are the anonymous clock ticks used in DARTS: Since they are just signal transitions on a single wire, they cannot encode any information except their occurrence time. The most important extension of DARTS is hence to add an additional bundled data wire to the clock signal, which carries 1 bit of data. This way, single ticks can be *marked* with a 1, distinguishing them from ordinary non-marked ticks that carry a 0.

We will actually mark every T -th DARTS tick, for some suitably chosen T . Such a marked tick kT , $k \geq 0$, is to be understood as the start of the $(k + 1)$ -st DARTS *round*, which consists of the marked DARTS tick kT and $T - 1$ subsequent unmarked ticks $kT + 1, kT + 2, \dots, (k + 1)T - 1$; the marked tick $(k + 1)T$ starts the next DARTS round. Note that the resulting DARTS ticks can be interpreted as a discrete, bounded clock operating modulo T . As DARTS rounds at any two correct TG-Algs are synchronous, marked ticks must always match in the pipelines of every counter, i.e., the Diff-Gate must always remove pairs of matching marked (or non-marked) ticks and can hence detect and remove any inconsistency.

The actual coupling between the instance of the pulse synchronization protocol and the DARTS clock running at node i is accomplished by means of two signals, namely, DARTS_i and PULSE_i :

- DARTS_i reports DARTS rounds to the pulse synchronization protocol. The rising edge of the DARTS_i signal, which may trigger a switch from *ready* to *propose*; is issued when the DARTS clock of node i generates tick $kT - X$, for some fixed $X < T$. The falling edge of DARTS_i reports the occurrence of the marked tick kT .
- PULSE_i reports the generation of a pulse to the DARTS clock. Its rising edge is issued on the transition to *accept*, and its falling edge signals the expiration of a fixed timeout T_y that is reset at the time the rising edge is transmitted.

The basic idea underlying the coupling of the pulse synchronization protocol and DARTS is to align marked ticks and pulses as follows: If the system operates normally, every correct node i first reaches some DARTS tick $kT - X$ and issues $\text{DARTS}_i = 1$. Next, a pulse is generated at node i by the pulse synchronization protocol, which thus sets $\text{PULSE}_i = 1$. Subsequently, the DARTS marked tick kT occurs, which is signaled by $\text{DARTS}_i = 0$. Finally, the pulse timeout T_y and hence $\text{PULSE}_i = 0$ occurs. Normal operation thus expects that the DARTS marked tick (= the falling edge of DARTS_i) occurs within the time window where PULSE_i is 1. Provided that the timeout used for

generating this window²² is chosen sufficiently large, namely, $\vartheta\rho(\pi + 2d + 1)$, this interleaving can indeed be guaranteed in normal operation.

As long as this is the case, we just let DARTS generate its ticks using its standard rules. Should a DARTS clock fail, however, such that PULSE_i and DARTS_i are not properly interleaved, then we will force marking the next DARTS tick (and possibly resetting the TG-Alg, if needed) upon the falling edge of PULSE_i . DARTS ticks and pulses (as well as marked DARTS ticks at different nodes) will hence only be re-aligned in case of errors or desynchronization: As long as DARTS clocks work correctly, any two correct TG-Algs will mark tick kT within the DARTS synchronization precision.

Provided that X and T_y are suitably chosen, it is not difficult to prove that the joint system consisting of pulse synchronization protocol and DARTS clocks will stabilize: After some unstable period, the pulse synchronization algorithm will stabilize, which we have proved to happen independently of the (non-)operation of DARTS clocks. When the pulse synchronization protocol eventually starts to generate synchronous pulses, the DARTS clocks will start to recover in a guided (synchronized) manner. When all correct DARTS clocks are eventually synchronized to within the intrinsic DARTS precision, the system will perpetually ensure the above interleaving at all correct nodes.

Some additional observations:

- (1) Since the DARTS precision is typically considerably smaller than the worst case pulse synchronization precision, the underlying DARTS clocks may be viewed as a “precision amplifier” (as well as a clock multiplier, see Section 7.2).
- (2) There is no need to specify properties possibly achieved by DARTS clocks during their own recovery. We only require that they eventually reach full synchronization in the presence of synchronous pulses at all correct nodes. In practice, DARTS clocks will typically also gradually improve their synchronization precision during this interval.
- (3) Although the pulse synchronization algorithm stabilizes even when the DARTS clocks behave arbitrary, it is nevertheless the case that it achieves better pulse synchronization precision when the DARTS clocks are fully synchronized.
- (4) One might ask why we did not just use the k -th rising edge of PULSE_i to mark the very next DARTS tick generated by the TG-Alg at node i . This simple solution has several major drawbacks. First, the pulse synchronization precision is typically worse than the synchronization precision provided by DARTS. Thus, every pulse would result in a temporary deterioration of the DARTS synchronization quality. Second, marked ticks are not necessarily generated exactly every T DARTS ticks. And last but not least, since DARTS clocks and pulse synchronization execute completely asynchronously, marking DARTS ticks at pulse occurrence times would create the potential of metastability every kT DARTS ticks, even if there is no failure at all.

7.3 DARTS \Rightarrow pulse synchronization

To implement this part of the coupling, every DARTS clock signals the upcoming occurrence of marked tick kT to its local instance of the pulse synchronization protocol. This is accomplished by the rising edge of DARTS_i , the dedicated DARTS *signal*, which is generated upon DARTS tick

²²We remark that it is vital not to rely on the DARTS clock here.

$KT - X$. If all correct nodes happen to do this within some time window when they are (w.r.t. the pulse algorithm) in state *ready* with $T_3 < T_4$ already expired, all correct nodes will switch to state *propose* within π time.²³ Subsequently, they will all switch to state *accept* within d time. To make sure that indeed all correct nodes are in state *ready* with T_3 already expired, up to small additional terms of $\mathcal{O}(d)$, we must choose the minimal duration of a DARTS round to be larger than $T_2 + T_3 + 4d$, while $(T_2 + T_4)/\vartheta$ is to exceed its maximal duration. Assuming that $\rho < 1.15$, Lemma 3.4 shows that this is feasible up to $\vartheta \approx 1.064$, which is clearly within the reach of ring oscillators [32].²⁴

7.4 Pulse synchronization \Rightarrow DARTS

This part of the coupling between DARTS and the pulse synchronization protocol requires two mechanisms:

- (1) A way to force a marked DARTS tick at node i upon the occurrence of the falling edge of PULSE_i , provided that no marked tick (i.e., the falling edge of DARTS_i) has been generated while PULSE_i was 1. This may also include recovering from a complete stall of the DARTS tick generation.
- (2) A way to recover accurate DARTS synchronization after forcing marked ticks, which may also include the need to get rid of any information from the preceding unstable period.

To achieve (1), we use a simple asynchronous circuit that supervises the interleaving of PULSE_i and DARTS_i , and generates a “force marking” signal if DARTS_i does not occur in time. Note that this device can be built in a way that entirely avoids metastability in case of normal operation. In an unstable period, however, it may happen that force marking occurs exactly at the time when DARTS generates its marked tick, so a metastable upset or two very close marked ticks (a forced and a regularly generated one) are possible.

There are several variants for implementing the forced marking itself, including the simplest variant of just resetting the TG-Alg in order to generate marked tick 0. The need for possibly resetting a TG-Alg originates from the fact that stateful TG-Alg components may deadlock due to earlier failures. For example, a deadlocked pipe will never propagate ticks from its input to the Diff-Gate. Unfortunately, resetting TG-Algs complicates DARTS recovery considerably: If a TG-Alg reset would also reset the remote pipes of its counters, it might lose “fresh” marked ticks generated by remote TG-Algs. Hence, remote pipes should only be reset when the *remote* node is reset. However, since a remote node might never observe a discrepancy between DARTS rounds and pulses, this approach might end up in the pipe not being reset at all. This is problematic as it might effectively render the node faulty despite all its components being operational. Luckily, we may utilize the fact that solving (2) under the assumption that correct pipes are not deadlocked yields a trivial means to distinguish a locked pipe from an operational one: If the Diff-Gate cannot remove any ticks within a certain time interval after a (correct) pulse, the pipe must have deadlocked and can safely be reset. At the next pulse, all pipes will have recovered from previous deadlocks and a solution to (2) assuming deadlock-free pipes will succeed.

²³In contrast to the model we employed for our analysis, we neglect the local signaling delay here, as it is smaller than the time to generate a single tick.

²⁴This is true regardless of the additional term of $\mathcal{O}(d)$, as the bound is derived from an asymptotic statement.

To explain how we achieve (2), we start with the observation that our way of marking every T -th tick implies that, for any two correct DARTS TG-Algs, it will always be a marked tick kT from a remote node that is matched by the local marked tick kT in every counter of Figure 7 when the Diff-Gate removes it. That is, if ever a marked tick is matching a non-marked tick in a counter, ticks have been lost or spuriously generated somewhere, or local and remote node are severely out of synchrony.

Assume for the moment that we could generate exactly one marked tick at every correct node, we made sure that no such tick is in the system before this happens, and that we have elastic pipelines of infinite size. The following simple strategy would eventually establish matching DARTS ticks: Whenever a Diff-Gate encounters a marked tick in one pipe matched by an ordinary tick in the other, it removes the ordinary tick only. At the pipe level, this rule implies that whatever the state of the pipes was before the marked ticks were generated, they will be cleared before the matching pair of marked ticks is removed. Since all DARTS tick generation rules ensure that no TG-Alg generates any tick $kT+1$, $kT+2$, \dots based on information from the previous DARTS round $k-1$ (consisting of DARTS ticks up to $kT-1$) all counter states will be valid as soon as the matching marked ticks kT have been removed. As DARTS essentially generates ticks based on comparing the number of locally and remotely generated ticks, this is enough to ensure stabilization of the DARTS system; full DARTS precision will be achieved quickly because nodes “catch up”, i.e., generate tick numbers that at least $f+1$ correct DARTS clocks already reached, faster than “new ticks”, i.e., ones that no correct node generated yet, may occur.

The issue of finite-size pipes is (largely) solved by the pulse synchronization protocol: Pulses and hence marked ticks are generated close to each other, in a time window of at most $2d + T_y \in \mathcal{O}(d)$ (provided that T_y is not unnecessarily large). Hence, apart from implementation issues, pipes that can accommodate all ticks that may be generated within this time window are sufficient for not losing any valid DARTS tick.

In reality, however, we cannot always expect the “single marked tick” setting described above: Elastic pipelines may initially be populated with arbitrarily many marked ticks from the unstable period. We must hence make sure that all these marked ticks (and the white ticks in between) are eventually removed, and that we do not generate new marked ticks close to each other. The pulse synchronization protocol will ensure that forced ticks are separated by T DARTS ticks, and our implementations of (1) and (2) will ensure with a large probability that a forced marked tick will not be generated close to a marked tick generated regularly by DARTS. Under these conditions, it is a relatively easy task to clear all superfluous marked ticks between pulses.

For example, we could asynchronously reset the whole data flip-flop chain that holds the markings of the ticks (not the ticks themselves!) currently in a pipe shortly after the rising flank of PULSE_i . Enlarging X and T_y slightly, we can be sure that all TG-Algs will remove spurious markings from their pipes before any marked tick associated with the respective correct pulse is generated. Although this could generate metastability in the Diff-Gate, namely, when the tick at the head of the pipe is a marked tick and the Diff-Gate is about to act when the pipe is reset upon arrival of a new marked tick arrives, this cannot happen during normal operation.

References

- [1] M. Ben-Or, D. Dolev, and E. N. Hoch. Fast self-stabilizing byzantine tolerant digital clock synchronization. In *Proc. 27th symposium on Principles of Distributed Computing (PODC)*, pages 385–394, 2008.
- [2] A. Berman and I. Keidar. Low-Overhead Error Detection for Networks-on-Chip. In *The 27th International Conference on Computer Design (ICCD)*, 2009.
- [3] R. Bhamidipati, A. Zaidi, S. Makineni, K. Low, R. Chen, K.-Y. Liu, and J. Dalgrehn. Challenges and Methodologies for Implementing High-Performance Network Processors. *Intel Technology Journal*, 6(3):83–92, 2002.
- [4] D. M. Chapiro. *Globally-Asynchronous Locally-Synchronous Systems*. PhD thesis, Stanford University, 1984.
- [5] C. Constantinescu. Trends and Challenges in VLSI Circuit Reliability. *IEEE Micro*, 23(4):14–19, 2003.
- [6] A. Daliot and D. Dolev. Self-Stabilizing Byzantine Pulse Synchronization. *CoRR*, abs/cs/0608092, 2006.
- [7] A. Daliot, D. Dolev, and H. Parnas. Self-Stabilizing Pulse Synchronization Inspired by Biological Pacemaker Networks. In *Proc. 6th Symposium on Self-Stabilizing Systems (SSS)*, 2003.
- [8] C. Dike and E. Burton. Miller and Noise Effects in a Synchronizing Flip-Flop. *IEEE Journal of Solid-State Circuits*, SC-34(6):849–855, 1999.
- [9] S. Dolev and J. L. Welch. Self-stabilizing clock synchronization in the presence of byzantine faults. *Journal of the ACM*, 51(5):780–799, 2004.
- [10] M. Ferringier, G. Fuchs, A. Steininger, and G. Kempf. VLSI Implementation of a Fault-Tolerant Distributed Clock Generation. *IEEE Symposium on Defect and Fault-Tolerance in VLSI Systems (DFT)*, pages 563–571, 2006.
- [11] M. J. Fischer and N. A. Lynch. A Lower Bound for the Time to Assure Interactive Consistency. *Information Processing Letters*, 14:183–186, 1982.
- [12] E. G. Friedman. Clock Distribution Networks in Synchronous Digital Integrated Circuits. *Proceedings of the IEEE*, 89(5):665–692, 2001.
- [13] G. Fuchs, M. Függer, and A. Steininger. On the Threat of Metastability in an Asynchronous Fault-Tolerant Clock Generation Scheme. In *Proc. 15th Symposium on Asynchronous Circuits and Systems (ASYNC)*, pages 127–136, Chapel Hill, N. Carolina, USA, 2009.
- [14] M. Függer. *Analysis of On-Chip Fault-Tolerant Distributed Algorithms*. PhD thesis, Technische Universität Wien, Institut für Technische Informatik, 2010.
- [15] M. Függer, A. Dielacher, and U. Schmid. How to Speed-Up Fault-Tolerant Clock Generation in VLSI Systems-on-Chip via Pipelining. In *Proc. 8th European Dependable Computing Conference (EDCC)*, pages 230–239, 2010.

- [16] M. Függer and U. Schmid. Reconciling Fault-Tolerant Distributed Computing and Systems-on-Chip. Research Report 13/2010, Technische Universität Wien, Institut für Technische Informatik, 2010.
- [17] M. Függer, U. Schmid, G. Fuchs, and G. Kempf. Fault-Tolerant Distributed Clock Generation in VLSI Systems-on-Chip. In *Proc. 6th European Dependable Computing Conference (EDCC)*, pages 87–96, 2006.
- [18] M. J. Gadlage, P. H. Eaton, J. M. Benedetto, M. Carts, V. Zhu, and T. L. Turflinger. Digital Device Error Rate Trends in Advanced CMOS Technologies. *IEEE Transactions on Nuclear Science*, 53(6):3466–3471, 2006.
- [19] E. Hoch, D. Dolev, and A. Daliot. Self-stabilizing Byzantine Digital Clock Synchronization. In *Proc. 8th Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2006)*, volume 4280, pages 350–362, 2006.
- [20] Internat. Technology Roadmap for Semiconductors, 2007.
- [21] D. J. Kinniment, A. Bystrov, and A. V. Yakovlev. Synchronization Circuit Performance. *IEEE Journal of Solid-State Circuits*, SC-37(2):202–209, 2002.
- [22] M. Malekpour. A Byzantine-Fault Tolerant Self-stabilizing Protocol for Distributed Clock Synchronization Systems. In *Proc. 9th Conference on Stabilization, Safety, and Security of Distributed Systems (SSS)*, pages 411–427, 2006.
- [23] L. Marino. General Theory of Metastable Operation. *IEEE Transactions on Computers*, C-30(2):107–115, 1981.
- [24] C. Metra, S. Francescantonio, and T. Mak. Implications of Clock Distribution Faults and Issues with Screening them During Manufacturing Testing. *IEEE Transactions on Computers*, 53(5):531–546, 2004.
- [25] C. J. Myers. *Asynchronous Circuit Design*. John Wiley & Sons, Inc., 2001.
- [26] M. Pease, R. Shostak, and L. Lamport. Reaching Agreement in the Presence of Faults. *Journal of the ACM*, 27:228–234, 1980.
- [27] T. Polzer, T. Handl, and A. Steininger. A Metastability-Free Multi-synchronous Communication Scheme for SoCs. In *Proc. 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2009)*, pages 578–592, 2009.
- [28] C. L. Portmann and T. H. Y. Meng. Supply Noise and CMOS Synchronization Errors. *IEEE Journal of Solid-State Circuits*, SC-30(9):1015–1017, 1995.
- [29] P. J. Restle and others;. A Clock Distribution Network for Microprocessors. *IEEE Journal of Solid-State Circuits*, 36(5):792–799, 2001.
- [30] Y. Semiat and R. Ginosar. Timing Measurements of Synchronization Circuits. In *Proc. 9th Symposium on Asynchronous Circuits and Systems (ASYNC)*, 2003.

- [31] T. K. Srikanth and S. Toueg. Optimal Clock Synchronization. *Journal of the ACM*, 34(3):626–645, 1987.
- [32] K. Sundaresan, P. Allen, and F. Ayazi. Process and temperature compensation in a 7-MHz CMOS clock oscillator. *IEEE J. Solid-State Circuits*, 41(2):433–442, 2006.
- [33] I. E. Sutherland. Micropipelines. *Communications of the ACM, Turing Award*, 32(6):720–738, 1989.
- [34] P. Teehan, M. Greenstreet, and G. Lemieux. A Survey and Taxonomy of GALS Design Styles. *IEEE Design and Test of Computers*, 24(5):418–428, 2007.
- [35] J. Widder and U. Schmid. The Theta-Model: Achieving Synchrony without Clocks. *Distributed Computing*, 22(1):29–47, 2009.