SAT-Based Abductive Diagnosis

Roxane Koitz[1*] and **Franz Wotawa**[1]
[1]Graz University of Technology, Graz, Austria
e-mail: {rkoitz, wotawa}@ist.tugraz.at

Abstract

Increasing complexity and magnitude of technical systems demand an accurate fault localization in order to reduce maintenance costs and system down times. Resting on solid theoretical foundations, model-based diagnosis provides techniques for root cause identification by reasoning on a description of the system to be diagnosed. Practical implementations in industries, however, are sparse due to the initial modeling effort and the computational complexity. In this paper, we utilize a mapping function automating the modeling process by converting fault information available in practice into propositional Horn logic sentences to be used in abductive model-based diagnosis. Furthermore, the continuing performance improvements of SAT solvers motivated us to investigate a SAT-based approach to abductive diagnosis. While an empirical evaluation did not indicate a computational benefit over an ATMS-based algorithm, the potential to diagnose more expressive models than Horn theories encourages future research in this area.

1 Introduction

Fault identification of technical systems is becoming increasingly difficult due to their rising complexity and scale. Economic and safety considerations have put accurate diagnosis not only into research focus but has led to a growing interest in practice as well.

Model-based diagnosis has been presented as a method to derive root causes for observable anomalies utilizing a description of the system to be diagnosed [1, 2]. Reiter [1] proposed a component-oriented model encompassing the correct system behavior and structure. Discrepancies, i.e. conflicts, arise when the observed and expected system performance diverge. Based on the minimal conflict sets, root causes for the inconsistencies are obtained by hitting set computation. Hence, fault diagnosis is a two step process, where first contradicting assumptions on component health, given a set of symptoms and the model, are identified. Then the sets intersecting all conflict sets are computed which constitute the diagnoses. At the same time [2] presents the General Diagnosis Engine (GDE) for multiple fault identification, drawing on the connection between inconsistencies and causes as well. Their approach employs an assumption-based truth maintenance system (ATMS) to detect conflicts and thereon compute diagnoses. Over the years much work has concentrated on model-based diagnosis applications in various domains, such as space probes [3] or the automotive industry [4].

Besides the consistency-based approach, a second method emerged within the field of model-based diagnosis, which exploits the concept of entailment to infer explanations for given observables. While related to the more traditional technique based on consistency, abductive model-based diagnosis requires a system formalization representing faults and their manifestations [5].

Even though based on a well-defined theory, a widespread acceptance of model-based diagnosis among industries has not been accounted for yet. Two main contributing factors can be identified: the initial model development and the computational complexity of diagnosis [6]. In order to diminish the modeling effort, [7] formulates a conversion of failure assessments available in practice into a propositional logic representation suitable for abductive diagnosis. Failure mode and effect analysis (FMEA) is an established reliability evaluation method utilized in various industrial fields. It considers possible component faults as well as their implications on the system's behavior [8]. Whereas there has been extensive research on the automatic generation of FMEAs from system models [9], we argue in favor of the inverse process. As these assessments report on failures and how they reveal themselves in the artifact's behavior, they provide knowledge requisite for abductive reasoning. In this paper, we present a compilation of FMEAs to models which can be used in abductive diagnosis.

Apart from discovering inconsistencies, an ATMS is capable of inferring abductive diagnoses. However, it may face computational challenges and is restricted to operate on propositional Horn clauses. In the case of the models we are extracting from the FMEAs, this is not a limitation so far. Nevertheless, as we anticipate to exploit more expressive representations, a different approach is required.

The performance of Boolean satisfiability (SAT) solvers has improved immensely over the last years and

[*]Authors are listed in alphabetical order.

several applications of SAT solvers in practice have proven successful. Furthermore, we are able to encode a greater variety of models in SAT. Thus, we propose a SAT-based approach to abductive diagnosis and empirically compare its performance to a procedure dependent on an ATMS.

The remainder of this paper is structured as follows. After formally providing the theoretical background on abductive diagnosis as well as relevant definitions in the context of SAT, we formulate the modeling process based on FMEAs and give information on the properties of the obtained system descriptions. In Section 5 we describe our SAT-based approach to abductive diagnosis and present an algorithm computing explanations for a given abduction problem. An empirical evaluation comparing our method to an ATMS-based diagnosis engine follows in Section 6. Subsequently, we provide some concluding remarks and give an outlook on future research possibilities.

2 Related Work

Mechanizing logic-based abduction has been an active research field for several decades with different approaches for generating explanations emerging, such as proof tree completion [10] and consequence finding [11]. While the former exploits a refutation proof involving hypotheses, the latter computes causes as logical consequences of the theory. As resolution is not consequence finding complete, [12] devised a procedure based on linear resolution which is sound and complete for consequence finding for propositional as well as first order logic.

While the number of practical applications in the context of abductive model-based diagnosis is rather small, in [13] the authors describe abductive reasoning in environmental decision support systems.

Most recently [14] present a SAT encoding for consistency-based diagnosis. The system description is compiled into a Boolean formula, such that the formula's satisfying assignments correspond to the solutions of the diagnosis problem. Based on the encoding, a SAT solver directly computes the diagnoses. In order to improve the solver's performance, the authors utilize several preprocessing techniques. An empirical comparison of their approach to other model-based diagnosis algorithms indicates that their SAT encoding yields performance benefits. Contrasting these results, [15] propose a translation to Max-SAT which could not outperform the stochastic model-based diagnosis algorithm SAFARI [16].

In [17] the authors present an algorithm which ties constraint solving to diagnosis, thus renders the detection of inconsistencies and subsequent hitting set computation unnecessary. Another direct approach by [18] computes minimal diagnoses for over-constrained problems by finding the sets of constraints to be relaxed in order to restore consistency. For Boolean formulas, those relaxations correspond to Minimal Correction Subsets (MCSes). Their hitting set dual, minimal unsatisfiable subsets (MUSes), constitute the set of subformulas explaining the unsatisfiability, i.e. refer to conflicts. While there are several algorithms for efficiently computing MCSes, most recently [19] develop three techniques for reducing the number of SAT solver calls for existing methods as well as a novel algorithm for MCSes computation.

As stated by [20] the complexity of abduction suspends of a polynomial-time transformation to SAT. Thus, in their work the authors present a fixed-parameter tractable transformation from propositional abduction to SAT exploiting backdoors and describe how to use their transformations to enumerate all solutions for a given abduction instance.

3 Preliminaries

This section provides a brief introduction to abductive model-based diagnosis. In particular, we describe the propositional Horn clause abduction problem (PHCAP) which provides the basis for our research. Note that throughout the paper we consider the closed-world assumption. In addition to the background on abductive model-based diagnosis, we formally define MUSes and MCSes.

3.1 Abductive Diagnosis

In contrast to the traditional consistency-based approach, abductive model-based diagnosis depends on a stronger relation between faults and observable symptoms, namely entailment. Hence, whereas consistency-based diagnosis reasons on the description of the correct system operation, abductive reasoning requires the model to capture the behavior in presence of a fault. By exploiting the notion of entailment and the causal links between defects and their corresponding effects, we can reason about explanations for observed anomalies. In general, abductive diagnosis is an NP-hard problem. However, there are certain subsets of logic, such as propositional definite Horn theory, which are tractable [21]. On these grounds we consider the PHCAP as defined in [22], which represents the connections between causes and effects as propositional Horn sentences. Similar to [22], we define a knowledge base as a set of Horn clauses over a finite set of propositional variables.

Definition 1 (**Knowledge base (KB)**). *A knowledge base (KB) is a tuple (A,Hyp,Th) where A denotes the set of propositional variables, Hyp* $\subseteq$ *A the set of hypotheses, and Th the set of Horn clause sentences over A.*

The set of hypotheses contains the propositions, which can be assumed to either be true or false and refer to possible causes. In order to form an abduction problem, a set of observations has to be considered for which explanations are to be computed.

Definition 2. ***(Propositional Horn Clause Abduction Problem (PHCAP))*** *Given a knowledge base (A,Hyp,Th) and a set of observations Obs* $\subseteq$ *A then the tuple (A,Hyp,Th,Obs) forms a Propositional Horn Clause Abduction Problem (PHCAP).*

Definition 3 (**Diagnosis; Solution of a PHCAP**). *Given a PHCAP (A,Hyp,Th,Obs). A set* $\Delta \subseteq$ *Hyp is a solution if and only if* $\Delta \cup Th \models Obs$ *and* $\Delta \cup Th \not\models \bot$. *A solution* Δ *is parsimonious or minimal if and only if no set* $\Delta' \subset \Delta$ *is a solution.*

A solution to a PHCAP is equivalent to an abductive diagnosis, as it comprises the set of hypotheses

explaining the observations. Even though Definition 3 does not impose the constraint of minimality on a solution, in practice only parsimonious explanations are of interest. Hence, we refer to minimal diagnoses simply as diagnoses. Notice that finding solutions for a given PHCAP is NP-complete [22].

As aforementioned an ATMS derives abductive explanations for propositional Horn theories, thus it can be utilized to find solutions to a PHCAP. Based on a graph structure where hypotheses, observations, and contradiction are represented as nodes, the Horn clause sentences defined in Th determine the directed edges in the graph. Each node is assigned a label containing the set of hypotheses said node can be inferred from. By updating the labels, the ATMS maintains consistency.

Algorithm `abductiveExplanations` exploits an ATMS and returns consistent abductive explanations for a set of observations [23]. In case the observation consists of a single effect, the label of the corresponding proposition already contains the abductive diagnoses. To account for multiple observables, i.e. $Obs = \{o_1, o_2, \ldots, o_n\}$, an individual implication is added, such that $o_1 \wedge o_2 \ldots \wedge o_n \rightarrow obs$, where obs is a new proposition not yet considered in A. Every set contained in the label of obs constitutes a solution to the particular PHCAP.

Algorithm 1 abductiveExplanations [23]

procedure ABDUCTIVEEXPLANATIONS
(A, Hyp, Th, Obs)
  Add Th to $ATMS$
  Add $\left(\bigwedge_{o \in Obs} o \rightarrow obs\right)$ to $ATMS$ ▷ $obs \notin A$
  return the label of obs
end procedure

3.2 Minimal Unsatisfiable Subset and Minimal Correction Subset

We assume standard definitions for propositional logic [24]. A propositional formula ϕ in CNF, defined over a set of Boolean variables $X = \{x_1, x_2, \ldots x_n\}$, is a conjunction of m clauses $(C_1, C_2, \ldots, C_m)$. A clause $C_i = (l_1, l_2, \ldots, l_k)$ is a disjunction of literals, where each literal l is either a Boolean variable or its complement. A truth assignment is a mapping $\mu : X \Rightarrow \{0, 1\}$ and a satisfying assignment for ϕ is a truth assignment μ such that ϕ evaluates to 1 under μ. Given a formula ϕ, the decision problem SAT consists of deciding whether there is a satisfying assignment for the formula.

In case ϕ is unsatisfiable there are subsets of ϕ, which are of special interest in the diagnosis context, namely the MUSes and MCSes. A Minimal Unsatisfiable Subset (MUS) comprises a subset of clauses which cannot be satisfied simultaneously. Notice that every proper subset of MUS is satisfiable. A Minimal Correction Subset (MCS) is the set of clauses which corrects the unsatisfiable formula, i.e. by removing any MCS the formula becomes satisfiable.

Given an unsatisfiable formula ϕ, an MUS and MCS are defined as follows [25]:

Definition 4. ***(Minimal Unsatisfiable Subset (MUS))*** *A subset $U \subseteq \phi$ is an MUS if U is unsatisfiable and $\forall C_i \in U, U \setminus (\{C_i\})$ is satisfiable.*

Definition 5. ***(Minimal Correction Subset (MCS))*** *A subset $M \subseteq \phi$ is an MCS if $\phi \setminus M$ is satisfiable and $\forall C_i \in M, \phi \setminus (M \setminus \{C_i\})$ is unsatisfiable.*

Since an MCS is a set of clauses correcting the unsatisfiable formula when removed, a single clause of an MUS is an MCS for this MUS. Note that the hitting set duality of MUSes and MCSes has been established [26].

Example. Consider the unsatisfiable formula ϕ in CNF.

$$\phi = \overbrace{(\neg a \vee \neg b \vee c)}^{C_1} \wedge \overbrace{(\neg c \vee d)}^{C_2} \wedge \overbrace{(c)}^{C_3} \wedge \overbrace{(\neg d)}^{C_4}$$

It is apparent that the combination of clauses C_2, C_3 and C_4 results in ϕ being unsatisfiable, hence

$$\text{MUSes}(\phi) = \{\{C_2, C_3, C_4\}\}.$$

By hitting set computation we arrive at the following set of MCSes:

$$\text{MCSes}(\phi) = \{\{C_2\}, \{C_3\}, \{C_4\}\}.$$

Removing any MCS of ϕ results in the formula being satisfiable.

It is worth noticing that utilizing subsets of unsatisfiable formulas has been proposed in regard to consistency-based diagnosis. In this context, a diagnosis is defined as the set of components which assumed faulty retains the consistency of the system. Thus, a consistency-based diagnosis corresponds to an MCS. For instance, [18] presents a direct diagnosis method computing MCSes for over-constrained systems. In conflict-directed algorithms, as proposed by Reiter [1], the minimal conflicts, arising from the deviations of the modeled to the experienced behavior, equate to the MUSes. In Section 5 we discuss our abductive diagnosis approach based on MUSes and MCSes.

4 Modeling Methodology

As mentioned before model-based diagnosis depends on a formal description of the system to be examined. The generation of appropriate models, however, is still an issue preventing a wide industrial adoption, since the modeling process is time-consuming and typically demanding for system engineers.

Therefore, we present a modeling methodology relying on FMEAs available in practice. An FMEA comprises a systematic component-oriented analysis of possible faults and the way they manifest themselves in the artifact's behavior and functionality [8]. This type of assessment is gaining importance and has become a mandatory task in certain industries, especially for systems that require a detailed safety analysis. Due to the knowledge capturing the causal dependencies between specific fault modes and symptoms, an FMEA provides information suitable for abductive reasoning [7].

Definition 6 (**FMEA**). *An FMEA is a set of tuples (C, M, E) where $C \in COMP$ is a component, $M \in MODES$ is a fault mode, and $E \subseteq PROPS$ is a set of effects.*

Running Example. In order to illustrate our modeling process, we use the converter of an industrial wind turbine as our running example [27]. Table 1

illustrates a simplified FMEA neglecting all parts affiliated with reliability analysis, such as severity ratings. Each row specifies a particular failure mode, (i.e. Corrosion, Thermo-mechanical fatigue (TMF) or High-cycle fatigue (HCF)) of a subsystem and determines its corresponding symptoms, such as *P_turbine* referring to a deviation between expected and measured turbine power output.

Component	Fault Mode	Effect
Fan	Corrosion	T_cabinet, P_turbine
Fan	TMF	T_cabinet, P_turbine
IGBT	HCF	T_inverter_cabinet, T_nacelle, P_turbine

Table 1: Excerpt of the FMEA of the converter

Consider the FMEA of the converter in Table 1. We can map the columns to their corresponding representations from Definition 6. The entries in the column *Component* constitute the elements of $COMP$, the entries in *Fault Mode* of $MODES$ and $PROPS$ subsumes the entries of *Effect*.

$$COMP = \{\ Fan, IGBT\ \}$$

$$MODES = \{\ Corrosion, TMF, HCF\ \}$$

$$PROPS = \left\{ \begin{array}{c} T\_cabinet, P\_turbine, \\ T\_inverter\_cabinet, T\_nacelle \end{array} \right\}$$

Through Definition 6 we obtain $FMEA_{Converter} =$

$$\left\{ \begin{array}{c} (Fan, Corrosion, \{T\_cabinet, P\_turbine\}), \\ (Fan, TMF, \{T\_cabinet, P\_turbine\}), \\ (IGBT, HCF, \{T\_inverter\_cabinet, T\_nacelle, \\ P\_turbine\}) \end{array} \right\}$$

Since the FMEA already represents the relation between defects and their manifestations the conversion to a suitable abductive KB is straightforward. It is worth noting that FMEAs usually consider single faults; thus, the resulting diagnostic system holds the single fault assumption. Let HC be the set of horn clauses. We define a mapping function $\mathfrak{M} : 2^{FMEA} \mapsto HC$ generating a corresponding propositional Horn clause for each entry of the FMEA [7].

Definition 7 (**Mapping function $\mathfrak{M}$**). *Given an FMEA, the function $\mathfrak{M}$ is defined as follows:*

$$\mathfrak{M}(FMEA) =_{def} \bigcup_{t \in FMEA} \mathfrak{M}(t)$$

where $\mathfrak{M}(C, M, E) =_{def} \{mode(C, M) \rightarrow e \,|e \in E\}$.

We utilize the proposition $mode(C, M)$ to denote that component C experiences fault mode M. Thus, the set of component-fault mode couples forms the set of hypotheses.

$$Hyp =_{def} \bigcup_{(C,M,E) \in FMEA} \{mode(C, M)\}.$$

In regard to the running example the following elements compose the set Hyp:

$$Hyp = \left\{ \begin{array}{c} mode(Fan, Corrosion), \\ mode(Fan, TMF), \\ mode(IGBT, HCF) \end{array} \right\}$$

The set of propositional variables A is defined as the union of all effects stored in the FMEA as well as all hypotheses, that is the set of component-fault mode pairs, i.e.:

$$A =_{def} \bigcup_{(C,M,E) \in FMEA} E \cup \{mode(C, M)\}$$

Continuing our converter example:

$$A = \left\{ \begin{array}{c} T\_cabinet, P\_turbine, \\ T\_inverter\_cabinet, T\_nacelle, \\ mode(Fan, Corrosion), \\ mode(Fan, TMF), \\ mode(IGBT, HCF) \end{array} \right\}$$

Applying $\mathfrak{M}$ results in the following set of propositional Horn clauses representing Th and thus completing $KB_{Converter}$:

$$Th = \left\{ \begin{array}{c} mode(Fan, Corrosion) \rightarrow T\_cabinet, \\ mode(Fan, Corrosion) \rightarrow P\_turbine, \\ mode(Fan, TMF) \rightarrow T\_cabinet, \\ mode(Fan, TMF) \rightarrow P\_turbine, \\ mode(IGBT, HCF) \rightarrow T\_inverter\_cabinet, \\ mode(IGBT, HCF) \rightarrow T\_nacelle, \\ mode(IGBT, HCF) \rightarrow P\_turbine \end{array} \right\}$$

On account of the mapping function $\mathfrak{M}$ and the underlying structure of the FMEAs, the compiled models feature a certain topology. First, the set of hypotheses and symptoms are disjoint sets. Second, since there is a causal link from faults to effects but not vice versa, the descriptions exhibit a forward and acyclic structure. Specifically, each implication connects one hypothesis to one effect, thus are bijunctive clauses. In order to account for impossible observations, we append additional implications to KB stating that an effect and its negation cannot occur simultaneously, i.e. $e \wedge \neg e \models \bot$.

The question remains whether the generated models are suitable for the diagnostic task. Abductive explanations are consistent by definition and complete given an exhaustive search. Thus, the appropriateness of the system description is determined by whether a single fault diagnosis can be obtained given all necessary information is available.

Definition 8. ***(One Single Fault Diagnosis Property (OSFDP))*** *Given a KB (A, Hyp, Th). KB fulfills the OSFDP if the following hold:*

$\forall m \in Hyp : \exists Obs \subseteq A : \{m\}$ *is a diagnosis of* (A, Hyp, Th, Obs) *and* $\neg \exists m' \in Hyp : m' \neq m$ *such that* $\{m'\}$ *is a diagnosis for the same PHCAP.*

The property ensures that under the assumption enough knowledge is available all single fault diagnoses can be distinguished and subsequently unnecessary replacement activities are avoided. To verify whether the OSFDP holds or not, we compute the set of propositions $\delta(h)$ implied by each hypothesis h and the theory. It is not fulfilled if we can record for two or more hypotheses the same $\delta(h)$. [7] describes a polynomial algorithm testing for the property. Note that the OSFDP check can be done on side of the FMEA before compiling the model. This is advantageous as the absence of the property indicates that internal variables or observations have not been considered in the FMEA.

Assume the set of hypotheses $\{h_1, h_2, \ldots, h_n\}$ share the same $\delta(h)$. We cannot distinguish $h_1, h_2, \ldots, h_n$ from one another and thus all corresponding components have to be repaired or replaced in case they are part of the diagnosis. Therefore, we can treat them as a unit by replacing $h_1, h_2, \ldots, h_n$ with a new hypothesis h'. Once all indistinguishable hypotheses have been removed, the KB satisfies the OSFDP. Regarding the hypotheses, which cannot be differentiated, as one cause during diagnosis has an effect on the computational effort as fewer hypotheses are to be considered.

Algorithm `distinguishHypotheses` replaces all indistinguishable causes and ensures that after termination the given KB satisfies the OSFDP. Evidently, the algorithm's complexity is determined by the three nested loops, hence $O(|Hyp|^2|A - Hyp|)$. Since there is a finite number of hypotheses and effects possibly included in $\delta(h)$ the algorithm must terminate.

Algorithm 2 distinguishHypotheses

```
procedure DISTINGUISHHYPOTHESES (A, Hyp, Th)
    Ψ[|Hyp|] ← Hyp
    for all h1 ∈ Ψ do
        for all h2 ∈ Ψ do
            if h1 ≠ h2 then
                if δ(h1) = δ(h2) and δ(h1) ≠ ∅ then
                    Create new hypothesis h'    ▷ h' ∉ Hyp
                    Add h' to Ψ
                    Add h' to A
                    for all e ∈ δ(h1) do
                        Add (h' → e) to Th
                        Remove (h1 → e) from Th
                        Remove (h2 → e) from Th
                    end for
                    Remove h1 ∧ h2 from Ψ
                    Remove h1 ∧ h2 from A
                end if
            end if
        end for
    end for
    return KB(A, Ψ, Th)
end procedure
```

Our running example of the converter does not fulfill the OSFDP, since $mode(Fan, Corrosion)$ and $mode(Fan, TMF)$ are not distinguishable. By removing both hypotheses and introducing $h' = mode((Fan, Corrosion), (Fan, TMF))$ the property is fulfilled.

Notice that abductive diagnosis is premised on the assumption that the model is complete; thus, we presume that all significant fault modes for each contributing part of the system have been contemplated in the FMEA. Furthermore, we expect on the one hand that the symptoms described within the FMEA are detectable in order to constitute observations. On the other hand, the automated mapping demands a consistent effect denotation throughout the analysis.

5 Abductive Diagnosis via SAT

Although an ATMS derives abductive diagnoses, it is limited to propositional Horn theories and subject to performance issues. Both problems have been accommodated through ATMS extensions and focus strategies. Nevertheless, the advances in the development of SAT solvers and their application to a vast number of different AI problems and industrial domains have motivated us to consider a SAT-based approach for abductive diagnosis.

Recall Definition 3 of a diagnosis: Δ is an abductive explanation if $\Delta \cup Th \models Obs$ and $\Delta \cup Th \not\models \bot$. Through logical equivalence we recast the first condition to $\Delta \cup Th \cup \{\neg Obs\} \models \bot$, where $\{\neg Obs\}$ denotes the set containing the complement of each observation in Obs, i.e. $\forall o \in Obs : \neg o \in \{\neg Obs\}$ [10]. In general, we can state the relation as follows: given the theory and assuming the hypotheses to be true whereas stating the absence of a set of observations, results in an inconsistency due to the fact that the causes entail the effects, i.e. $Hyp \cup Th \cup \{\neg Obs\} \models \bot$. Thus, we draw on this relationship and reformulate the problem of generating minimal abductive explanations for a set of observations to computing minimal unsatisfiable subformulas.

Since MUSes contain several unsatisfiable subsets irrelevant for the diagnostic task, we define the set $MUSes_{Hyp}$, which only contains subset minimal MUS comprising clauses referring to hypotheses:

Definition 9. *($MUSes_{Hyp}$) Let $MUSes$ be the set of MUSes of $Hyp \cup Th \cup \{\neg Obs\}$, then $\forall M \in MUSes_{Hyp} : \exists U \in MUSes : M = U \cap Hyp$ and $\neg\exists M' \in MUSes_{Hyp} : M' \subset M$.*

Corollary 1. *Given a $PHCAP(A, Hyp, Th, Obs)$, let $MUSes_{Hyp}$ be the set of interesting MUSes. A set $\Delta \subseteq Hyp$ is a minimal abductive diagnosis if $\exists M \in MUSes_{Hyp} : \Delta = M$ and $\Delta \cup Th \not\models \bot$.*

Proof. We can restate the problem of computing inconsistencies to finding the set of prime implicates of $Th \wedge Hyp \wedge \{\neg Obs\}$. By definition, the prime implicates are equivalent to the MUSes of said formula. □

Deriving a minimal abductive explanation corresponds to computing a minimal subset of the hypotheses, which cannot be simultaneously satisfied with the theory and the negation of observations.

We devised the algorithm `satAB`, which computes the set of abductive diagnoses for a given PHCAP based on MUS enumeration. First, in order to take advantage of the MUSes, which correspond to the solutions of the PHCAP, we create an unsatisfiable CNF encoding of the problem. Since the Th consists of Horn clauses a conversion into CNF is straightforward. Note that we are, however, not limited to Horn clause models, as we can create a CNF representation based on Tseitin transformation [28]. We refer to the set of clauses associated with the theory as $\mathcal{T}$. For each $h \in Hyp$ we create a single clause assuming h to be true. Additionally, we generate a disjunction containing the negated observations. The resulting unsatisfiable formula is referred to as ϕ. $\Delta - Set$ is the set of diagnoses obtained from the PHCAP.

The diagnostic task consists in computing the sets of hypotheses which are responsible for the unsatisfiability of ϕ, i.e. $MUSes_{Hyp}(\phi)$. Since finding satisfiable subsets is an NP-hard problem whereas UNSAT resides in Co-NP, we employ an MCSes enumeration algorithm on the unsatisfiable formula and then derive the diagnoses via hitting set computation [25]. As we are only

$C_1 : \neg mode(Fan, Corrosion) \vee T\_cabinet$	$C_2 : \neg mode(Fan, Corrosion) \vee P\_turbine$
$C_3 : \neg mode(Fan, TMF) \vee T\_cabinet$	$C_4 : \neg mode(Fan, TMF) \vee P\_turbine$
$C_5 : \neg mode(IGBT, HCF) \vee T\_inverter\_cabinet$	$C_6 : \neg mode(IGBT, HCF) \vee T\_nacelle$
$C_7 : \neg mode(IGBT, HCF) \vee P\_turbine$	$C_8 : mode(Fan, Corrosion)$
$C_9 : mode(Fan, TMF)$	$C_{10} : mode(IGBT, HCF)$
$C_{11} : \neg P\_turbine \vee \neg T\_cabinet$	

Figure 1: SAT encoding of the running example

Algorithm 3 satAB

procedure SATAB (A, Hyp, Th, Obs)
  $MCSes \leftarrow \emptyset$
  $MCSes_{Hyp} \leftarrow \emptyset$
  $\mathcal{T} \leftarrow$ `CNF(`Th`)` ▷ CNF representation of Th
  $\phi \leftarrow \mathcal{T} \cup Hyp \cup \bigvee_{o \in Obs} \neg o$
  $MCSes \leftarrow$ `MCSes(`ϕ`)` ▷ MCS enumeration algorithm
  for all $m \in MCSes$ **do**
    if $m \subseteq Hyp$ **and** $m \cup Th$ is consistent **then**
      $MCSes_{Hyp} \leftarrow m \cup MCSes_{Hyp}$
    end if
  end for
  $\Delta - Set \leftarrow$ `MHS(`$MCSes_{Hyp}$`)` ▷ Minimal hitting set algorithm
  return $\Delta - Set$
end procedure

interested in the conflicts stemming from the assumptions that all hypotheses are true, we select each MCS only containing clauses referring to explanations. For this reason, we create the set $MCSes_{Hyp}$ such that $\forall m \in MCSes_{Hyp} : m \subseteq Hyp$. This has one practical rational: it diminishes the number of sets to be considered by the hitting set algorithm. The corresponding MUSes derived via hitting set computation of $MCSes_{Hyp}$ already constitute the abductive diagnoses.

Consider again our running example of the converter. We already obtained the KB via the mapping function $\mathfrak{M}$. Let us assume that the condition monitoring system of the wind turbine encountered that the turbine's power output is lower than expected ($P\_turbine$) and that the cabinet temperature exceeds a certain threshold ($T\_cabinet$), i.e. $Obs = \{P\_turbine, T\_cabinet\}$. In Figure 1 we depict the CNF representation ϕ of the abduction problem. Clauses C_1 to C_7 refer to $\mathcal{T}$, C_8 to C_{10} to the set Hyp and clause C_{11} contains the negation of the set of observations.

Computing the $MCSes$ of ϕ we obtain: $MCSes =$

$$\left\{ \begin{array}{c} \{C_{11}\}, \{C_1, C_3\}, \{C_1, C_9\}, \{C_3, C_8\}, \{C_9, C_8\}, \\ \{C_4, C_7, C_2\}, \{C_4, C_{10}, C_2\}, \{C_4, C_7, C_8\}, \\ \{C_4, C_{10}, C_8\}, \{C_2, C_9, C_7\}, \{C_2, C_9, C_{10}\} \end{array} \right\}.$$

Extracting the MCSes, which only contain clauses from Hyp and are consistent with regard to the theory, results in

$$MCSes_{Hyp} = \{\{C_9, C_8\}\}.$$

By computing the hitting set of $MCSes_{Hyp}$, we obtain the set of MUSes solely referring to explanations, which is in fact the set of diagnoses:

$$\Delta - Set = \{\{C_9\}, \{C_8\}\}.$$

Hence the abductive diagnoses are $\Delta_1 = \{mode(Fan, Corrosion)\}$ and $\Delta_2 = \{mode(Fan, TMF)\}$.

6 Empirical Evaluation

To determine whether computing abductive diagnoses via SAT yields any computational advantages in the case of our models, we conducted an empirical evaluation, comparing `abductiveExplanations` to `satAB` on several instances of FMEAs. In case of the former we employed a Java implementation of an unfocused ATMS. The algorithm `satAB` exploits on the one hand an MCS enumeration procedure and on the other hand an implementation of a hitting set algorithm. We utilized the MCS_{LS} tool by [19] to compute the MCSes. MCS_{LS} is written in C++, employs Minsat 2.2 as the SAT solver, and provides the possibility to apply several MCS enumeration algorithms. We decided for the CLD approach of MCS_{LS}, which takes advantage of disjoint unsatisfiable cores and showed the best overall performance in a preliminary experimental set-up. Regarding the hitting set computation, we engaged a Java implementation of the Binary Hitting Set Tree algorithm [29] which performed well in a comparison of minimal hitting set algorithms [30]. All the numbers presented in this section were obtained from a Lenovo ThinkPad T540p Intel Core i7-4700MQ processor (2.60 GHz) with 8 GB RAM running Ubunutu 14.04 (64-bit).

Several publicly available as well as project internal FMEAs provide the basis for our evaluation. They cover various technical systems and subsystems with different underlying structures. In particular they describe faults in electrical circuits, a connector system by Ford (FCS), the Focal Plane Unit (FPU) of the Heterodyne Instrument for the Far Infrared (HIFI) built for the Herschel Space Observatory, printed circuit boards (PCB), the Anticoincidence Detector (ACD) mounted on the Large Area Telescope of the Fermi Gamma-ray Space Telescope, the Maritim ITStandard (MiTS), and rectifier, inverter, transformer, backup components, as well as main bearing of an industrial wind turbine. By applying the mapping function $\mathfrak{M}$, we generated the corresponding abductive knowledge bases KB for each FMEA. Table 2 provides an overview of the FMEAs' structure and the evaluation results. It is worth noting that the FMEAs vary in the number of hypotheses, i.e. component-fault mode couples, the number of effects, and the number of rules, i.e. the links between faults and symptoms. Due to Th of an abductive KB comprising Horn clauses, a conversion into a CNF representation, suitable for the MCS_{LS} tool, is straightforward. We do not address the model compilation times, since the system description would be compiled offline and

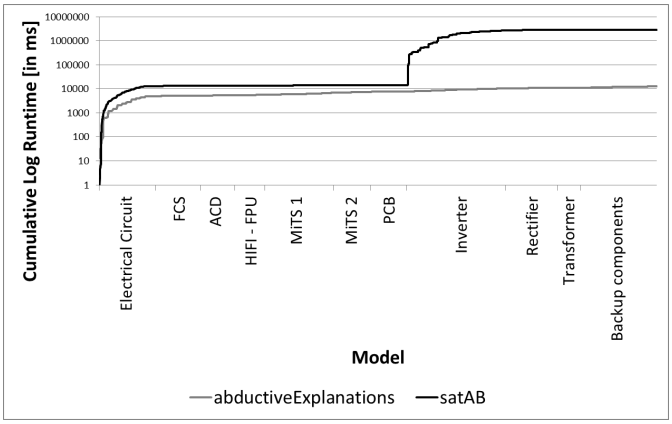

Figure 2: Cumulative runtimes of `abductiveExplanations` and `satAB` for the FMEA instances

the mapping execution consumed less than one second for the examples we utilized so far.

Table 2 shows that none, except of the model resulting from the transformer's FMEA, of the original models satisfy the OSFDP. Therefore, we compiled a second set of models fulfilling the property by exchanging each set of indistinguishable hypotheses with a new single hypothesis representing said set. For example, Algorithm `distinguishHyp` ensures that the resulting KB satisfies the OSFDP. In Table 2 the original models are identified accordingly, and the adapted models are provided with the label *OSFDP*. Note that the number of hypotheses and rules diminishes for the adapted models.

In the experiments, we computed the abductive explanations for $|Obs|$ from one to the maximum number of effects possible. The observations were generated randomly; however, the same set was used for `satAB` and `abductiveExplanations` on the original as well as adapted model. The results reported in Table 2 have been obtained from ten trials and both algorithms faced a 200 seconds runtime limit. Whereas some of the small runtimes are arguable due to the measurement in the milliseconds range, Table 2 reveals that `satAB` (Mean = 703.73 ms, SD = 8432.07 ms, Median = 0.59 ms, Skewness = 18.61) does not outperform `abductiveExplanations` (Mean = 3.08 ms, SD = 16.38 ms, Median = 1 ms, Skewness = 12.68) in general. From the statistical data we can infer that the underlying distribution of both algorithms is highly right skewed, thus the bulk of values is located towards the lower runtimes. We can even observe that for certain instances, the SAT-based approach performs rather poorly. Amongst these are the model of an inverter and a rectifier of an industrial wind turbine. `satAB` exceeded the given timeout four times for the former. Notice that in all these cases the MCSes generation already reached the time threshold. According to [19] CLD requires $|\phi| - p + 1$ SAT solver calls, where p refers to the size of the smallest MCS of ϕ. In our case $p = 1$, as the clause representing the set of negated observations always constitutes an MCS. Thus, $|\phi|$ SAT solver calls are necessary, where $|\phi|$ is determined by $|Th| + |Hyp| + 1$, with 1 referring to the clause containing the observations. Unsurprisingly, the larger FMEAs are more computationally demanding. It is worth mentioning that in the majority of cases the hitting set computation accounted for a negligible fraction of the total runtime.

Figure 2 illustrates the cumulative log runtimes for `satAB` and `abductiveExplanations` on the FMEA models generated. Although `abductiveExplanations` performs on average better, the first model requires a longer computation time for both algorithms. Moreover, the illustration reveals the high computational effort necessary for `satAB` to compute the diagnoses for the model of the inverter. As expected we observe particularly high runtimes when the set of observations contains effects corresponding to different hypotheses. This has a greater impact on `satAB` than on the ATMS implementation. For the section from the models FCS to PCB in Figure 2, however, we can see that the cumulative runtime for `abductiveExplanations` rises at a steeper angle. Generally, the data gathered in the experiment do not suggest a performance benefit of the SAT-based approach over an ATMS implementation.

7 Conclusion and Future Work

In the course of the paper, we presented a mapping from failure assessments available to propositional Horn clause models. The modeling methodology relies on FMEAs as they comprise information on faults and their symptoms. Hence, they provide a suitable source for model compilation. Although in our case an ATMS can be used to compute abductive diagnoses, it is limited to propositional Horn theories. We proposed a SAT-based approach to abductive model-based diagnosis which allows us to reason on more expressive representations. Our method is based on computing conflict sets, i.e. MUSes, resulting from a rewritten, unsatisfiable system description. Subsets of these unsatisfiable cores constitute the minimal abductive explanations. Since the computation of MUSes is computationally demanding our proposed algorithm exploits its hitting set dual, MCSes, in order to derive minimal diagnoses.

We empirically compared an implementation of a diagnosis engine employing an ATMS to our SAT-based algorithm. The results indicate that while for some of the models, the algorithm performs well, in general we could not observe a performance advantage. Particular examples led to even longer computation times than the ATMS-based implementation. Despite the fact that the data provided no evidence of a computational benefit in employing a SAT-based approach, we believe that the possibility to utilize more expressive models provides an interesting incentive for future research in this area.

Since the evaluation results, did not indicate a superiority of the SAT-based approach on grounds of MCSes enumeration, we currently investigate direct conflict generation methods. Additionally, due to the model structure and the experiment data we are planning on employing compilation methods [31, 32], in order to divert some of the computational inefficiency to the model generation process.

Acknowledgments

The work presented in this paper has been supported by the FFG project Applied Model Based Reasoning (AMOR) under grant 842407. We would further like to express our gratitude to our industrial partner, Uptime Engineering GmbH.

	Model Structure			#Diagnoses					Runtime [in ms]			
Component	**#Hyp**	**#Effects**	**#Rules**	**MAX**	**AVG**	**SF**	**DF**	**TF**	**Algorithm**	**MIN**	**MAX**	**AVG**
Electrical circuit												
Original	32	17	52	792	197.15	11	11	66	**abductive Explanations**	< 1	425	27.87
									satAB	< 1	181.33	76.05
OSFDP	15	17	35	1	1	1	1	1	**abductive Explanations**	< 1	8	0.33
									satAB	< 1	1.91	0.16
FCS												
Original	17	17	51	18	2.93	3	6	18	**abductive Explanations**	< 1	1	0.42
									satAB	< 1	6.41	1.28
OSFDP	15	17	49	18	2.75	3	6	18	**abductive Explanations**	< 1	61	2.04
									satAB	< 1	4.73	0.56
ACD												
Original	13	16	41	15	2.89	5	15	15	**abductive Explanations**	< 1	84	1.38
									satAB	< 1	2.89	0.35
OSFDP	12	16	39	10	2.04	5	10	10	**abductive Explanations**	< 1	1	0.29
									satAB	< 1	2.435	0.28
Main bearing												
Original	3	5	20	3	2.54	3	0	0	**abductive Explanations**	< 1	1	0.16
									satAB	< 1	1	0.09
OSFDP	2	5	15	2	1.54	2	0	0	**abductive Explanations**	< 1	1	0.12
									satAB	< 1	0.61	0.03
HIFI - FPU												
Original	17	11	36	63	8.64	3	7	21	**abductive Explanations**	< 1	86	2.54
									satAB	< 1	8.33	3
OSFDP	9	11	27	6	1.55	2	2	3	**abductive Explanations**	< 1	1	0.15
									satAB	< 1	1	0.09
MiTS 1												
Original	18	21	48	24	8.40	3	2	6	**abductive Explanations**	< 1	94	3.40
									satAB	< 1	3.02	0.39
OSFDP	13	21	43	1	1	1	1	1	**abductive Explanations**	< 1	100	1.54
									satAB	< 1	2.15	0.16
MiTS 2												
Original	22	15	48	288	39.98	4	8	18	**abductive Explanations**	< 1	109	4.49
									satAB	< 1	15.16	3.43
OSFDP	14	15	37	5	2.02	1	5	2	**abductive Explanations**	< 1	1	0.33
									satAB	< 1	1.68	0.20
PCB												
Original	10	11	24	2	1.49	2	2	2	**abductive Explanations**	< 1	1	0.21
									satAB	< 1	1.49	0.1
OSFDP	9	11	23	1	1	1	1	1	**abductive Explanations**	< 1	1	0.11
									satAB	< 1	1	0.1
Inverter												
Original	30	38	144	450	23.73	19	5	50	**abductive Explanations**	< 1	107	6.15
									satAB	< 1	166593	5007.37
OSFDP	23	38	124	66	5.89	14	3	6	**abductive Explanations**	< 1	94	1.67
									satAB	< 1	1110.82	38.23
Rectifier												
Original	20	17	93	88	10.83	8	24	32	**abductive Explanations**	< 1	6	1.07
									satAB	< 1	24236.9	1070.88
OSFDP	14	17	66	22	3.06	5	18	8	**abductive Explanations**	< 1	1	0.63
									satAB	< 1	44.74	4.88
Transformer												
Original	5	8	22	2	1.06	2	2	1	**abductive Explanations**	< 1	1	0.16
									satAB	< 1	1.69	0.06
OSFDP	5	8	22	2	1.06	2	2	1	**abductive Explanations**	< 1	1	0.13
									satAB	< 1	1.91	0.08
Backup components												
Original	25	30	114	252	23.06	8	12	21	**abductive Explanations**	< 1	138	5.24
									satAB	< 1	41.98	12.89
OSFDP	19	30	95	48	3.29	7	7	10	**abductive Explanations**	< 1	4	0.79
									satAB	< 1	10.06	3.09

Table 2: Features of the FMEAs and experimental results. For each component we conducted the experiment using an implementation of `abductiveExplanations` and `satAB`. The columns *SF*, *DF*, *TF* display the maximum number of single faults, double faults, and triple faults, respectively.

References

[1] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.

[2] Johan de Kleer and Brian C Williams. Diagnosing Multiple Faults. *Artificial Intelligence*, 32(1):97–130, 1987.

[3] Brian C Williams and P Pandurang Nayak. A model-based approach to reactive self-configuring systems. In *Proceedings of the National Conference on Artificial Intelligence*, pages 971–978, 1996.

[4] Peter Struss, Andreas Malik, and Martin Sachenbacher. Case studies in model-based diagnosis and fault analysis of car-subsystems. In *Proc. 1st Int'l Workshop Model-Based Systems and Qualitative Reasoning*, pages 17–25, 1996.

[5] Luca Console, Daniele Theseider Dupre, and Pietro Torasso. On the Relationship Between Abduction and Deduction. *Journal of Logic and Computation*, 1(5):661–690, 1991.

[6] Peter Zoeteweij, Jurryt Pietersma, Rui Abreu, Alexander Feldman, and Arjan JC Van Gemund. Automated fault diagnosis in embedded systems. In *Secure System Integration and Reliability Improvement, 2008. SSIRI'08. Second International Conference on*, pages 103–110. IEEE, 2008.

[7] Franz Wotawa. Failure mode and effect analysis for abductive diagnosis. In *Proceedings of the International Workshop on Defeasible and Ampliative Reasoning (DARe-14)*, volume 1212. CEUR Workshop Proceedings, ISSN 1613-0073, 2014. http://ceur-ws.org/Vol-1212/.

[8] Peter G. Hawkins and Davis J. Woollons. Failure modes and effects analysis of complex engineering systems using functional models. *Artificial Intelligence in Engineering*, 12:375–397, 1998.

[9] Chris Price and Neil Taylor. Automated multiple failure fmea. *Reliability Engineering & System Safety*, 76:1–10, 2002.

[10] Sheila A McIlraith. Logic-based abductive inference. *Knowledge Systems Laboratory, Technical Report KSL-98-19*, 1998.

[11] Pierre Marquis. Consequence finding algorithms. In *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, pages 41–145. Springer, 2000.

[12] Katsumi Inoue. Linear resolution for consequence finding. *Artificial Intelligence*, 56(2):301–353, 1992.

[13] Franz Wotawa, Ignasi Rodriguez-Roda, and Joaquim Comas. Environmental decision support systems based on models and model-based reasoning. *Environmental Engineering and Management Journal*, 9(2):189–195, 2010.

[14] Amit Metodi, Roni Stern, Meir Kalech, and Michael Codish. A novel SAT-based approach to model based diagnosis. *Journal of Artificial Intelligence Research*, pages 377–411, 2014.

[15] Alexander Feldman, Gregory Provan, Johan de Kleer, Stephan Robert, and Arjan van Gemund. Solving model-based diagnosis problems with Max-SAT solvers and vice versa. In *DX-10, International Workshop on the Principles of Diagnosis*, 2010.

[16] Alexander Feldman, Gregory M Provan, and Arjan JC van Gemund. Computing minimal diagnoses by greedy stochastic search. In *AAAI*, pages 911–918, 2008.

[17] Iulia Nica and Franz Wotawa. ConDiag-computing minimal diagnoses using a constraint solver. In *International Workshop on Principles of Diagnosis*, pages 185–191, 2012.

[18] Alexander Felfernig and Monika Schubert. Fastdiag: A diagnosis algorithm for inconsistent constraint sets. In *Proceedings of the 21st International Workshop on the Principles of Diagnosis (DX 2010), Portland, OR, USA*, pages 31–38, 2010.

[19] Joao Marques-Silva, Federico Heras, Mikolás Janota, Alessandro Previti, and Anton Belov. On computing minimal correction subsets. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 615–622. AAAI Press, 2013.

[20] Andreas Pfandler, Stefan Rümmele, and Stefan Szeider. Backdoors to abduction. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1046–1052. AAAI Press, 2013.

[21] Gustav Nordh and Bruno Zanuttini. What makes propositional abduction tractable. *Artificial Intelligence*, 172:1245–1284, 2008.

[22] Gerhard Friedrich, Georg Gottlob, and Wolfgang Nejdl. Hypothesis classification, abductive diagnosis and therapy. In *Expert Systems in Engineering Principles and Applications*, pages 69–78. Springer, 1990.

[23] Franz Wotawa, Ignasi Rodriguez-Roda, and Joaquim Comas. Abductive Reasoning in Environmental Decision Support Systems. In *AIAI Workshops*, pages 270–279, 2009.

[24] Chin-Liang Chang and Richard Char-Tung Lee. *Symbolic logic and mechanical theorem proving*. Academic press, 1973.

[25] Mark H Liffiton and Karem A Sakallah. Algorithms for computing minimal unsatisfiable subsets of constraints. *Journal of Automated Reasoning*, 40(1):1–33, 2008.

[26] Elazar Birnbaum and Eliezer L Lozinskii. Consistent subsets of inconsistent systems: structure and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, 15(1):25–46, 2003.

[27] Christopher S Gray, Roxane Koitz, Siegfried Psutka, and Franz Wotawa. An abductive diagnosis and modeling concept for wind power plants. In *International Workshop on Principles of Diagnosis*, 2014.

[28] Gregory Tseitin. On the complexity of proofs in propositional logics. In *Seminars in Mathematics*, volume 8, pages 466–483, 1970.

[29] Li Lin and Yunfei Jiang. The computation of hitting sets: review and new algorithms. *Information Processing Letters*, 86(4):177–184, 2003.

[30] Ingo Pill, Thomas Quaritsch, and Franz Wotawa. From conflicts to diagnoses: An empirical evaluation of minimal hitting set algorithms. In *22nd Int. Workshop on the Principles of Diagnosis*, pages 203–210, 2011.

[31] Adnan Darwiche. Decomposable negation normal form. *Journal of the ACM (JACM)*, 48(4):608–647, 2001.

[32] Pietro Torasso and Gianluca Torta. Computing minimum-cardinality diagnoses using OBDDs. In *KI 2003: Advances in Artificial Intelligence*, pages 224–238. Springer, 2003.