Noname manuscript No.
(will be inserted by the editor)

Capacity Analysis for Railway Construction using SAT Modulo Discrete Event Simulation

Bjørnar Luteberget · Koen Claessen · Christian Johansen · Martin Steffen

the date of receipt and acceptance should be inserted later

Abstract Railway capacity is complex to define and analyze, and existing tools and methods used in practice require comprehensive models of the railway network and its timetables. Design engineers working within the limited scope of construction projects report that only ad-hoc, experience-based methods of capacity analysis are available to them. Designs often have subtle capacity pitfalls which are discovered too late, only when network-wide timetables are made – there is a mismatch between the scope of construction projects and the scope of capacity analysis, as currently practiced.

We suggest a language for capacity specifications suited for construction projects, expressing properties such as running time, train frequency, overtaking and crossing. Such specifications can be used as contracts in the interface between construction projects and network-wide capacity analysis.

We show how these properties can be verified fully automatically by building a special-purpose solver which splits the problem into two: an abstracted SAT-based dispatch planning, and a continuous-domain dynamics and timing constraints evaluated using discrete event simulation. The two components communicate in a CEGAR-loop (counterexample-guided abstraction refinement). This architecture is beneficial because it clearly distinguishes the combinatorial choices from continuous calculations, so that the simulation can be extended by relevant details as needed. We describe how loops in the infrastructure can be

The first author was partially supported by the project RailCons, — *Automated Methods and Tools for Ensuring Consistency of Railway Designs,* with number 248714 funded by the Norwegian Research Council.

Bjørnar Luteberget
RailComplete AS, Sandvika, Norway
E-mail: bjornar.luteberget@railcomplete.no

Christian Johansen
Department of Technology Systems, University of Oslo, Norway
E-mail: cristi@ifi.uio.no

Koen Claessen
Chalmers University, Sweden
E-mail: koen@chalmers.se

Martin Steffen
Department of Informatics, University of Oslo, Norway
E-mail: msteffen@ifi.uio.no

handled to eliminate repeating dispatch plans, and use case studies based on data from existing infrastructure and ongoing construction projects to show that our method is fast enough at relevant scales to provide agile verification in a design setting.

1 Introduction

The planning and engineering of a railway control system has safety as primary requirement. Safety is ensured through the so-called signaling principles, and detailed requirements have been put in place for station layouts, controller implementations, and operation procedures.

Secondary to safety, the notion of performance and capacity of a railway control system remains more elusive. The capacity of a railway control system, and thus of railway infrastructure in general, is hard to define precisely (see [15, 1, 22]). Any capacity measure will necessarily make assumptions about the operation of the railway. One can say that the railway infrastructure does not have an inherent capacity, only capacity for specific use cases. A fully accurate assessment of capacity can only be made under a fully specified timetable, meaning that every train's arrival and departure times at all stations in the network must be known. This makes for a highly coupled analysis, as constructing an actual timetable requires bringing together details about infrastructure, rolling stock, transportation demands, and crew schedules. Systematic capacity analysis for railways is typically performed on the scale of national railway networks, using comprehensive input on infrastructure and timetables, and only after planning and engineering has produced a final design. Moreover, the widely used methods and tools for capacity analysis are heavy-duty methods, consisting of complicated simulations, and require specialized knowledge, thus not being suitable for more agile design-time verification of railway stations.

For construction projects and control system engineering, it would not be feasible to use a fully specified timetable for verifying that the control system will be able to provide the required capacity, because (1) detailed timetabling and capacity analysis takes too much effort and specialized knowledge, and is usually saved for later stages of design, and (2) the design of a control system cannot or should not depend too heavily on other parts of the network, as these parts may also change in the future.

Another approach to capacity analysis is the so-called analytical capacity approach, which views the railway network as a network of queues, or a maximum flow problem, abstracting away the low-level discrete behavior while preserving the high-level continuous behavior. These methods can give preliminary or low-precision network-wide results, but fail to account for the critical factors which arise when performance is pushed to the limit. Simplifying assumptions that can be suitable for network-scale capacity analysis, such as instantaneous speed changes, or fixed traveling times between different locations, are usually not suitable for infrastructure design. Specifically, disregarding the discrete allocation logic of the interlocking system, and the position and velocities of individual trains, makes these methods unsuitable for analysis of signalling design. The detailed optimization of signal and detector locations needs to account for a detailed model of train dynamics and control system behavior exactly because higher-level analysis requires this assumption of local optimization to the simplified behaviors used in network-global analysis.

As none of these techniques are particularly well-suited, railway engineers working on construction projects usually rely on informal, vague, or even non-existant capacity specifications, and need to make ad-hoc analyses of how the control system might provide this capacity.

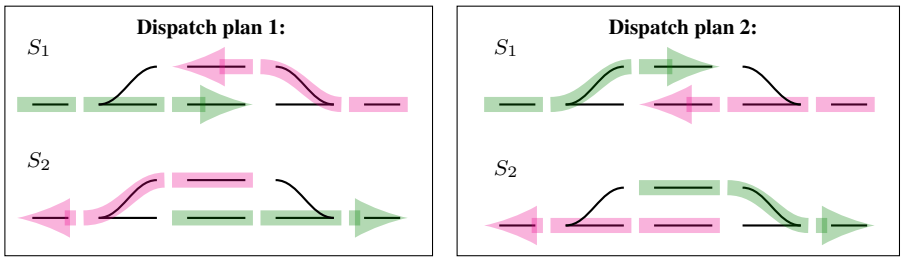

Fig. 1: Two alternative dispatch plans for achieving a crossing of two trains on a two-track station. The green areas show track segments which are currently occupied by a train going from left to right, while the pink areas show track segments which are currently occupied by a train going from right to left.

In consequence, this paper addresses the following problem: in the context of designing the layout and control systems for railway stations, does the station infrastructure have the *capacity* to handle the amount of trains and the desired traveling times to provide adequate service in transportation of goods and passengers?

As an example, consider the question of crossing trains on a railway station. Fig. 1 shows two sequences of movements which result in such a crossing. There are a number of details of the railway design which can cause this scenario to become infeasible (or take an unacceptably long time), such as signal placement, detector placement, correct allocation and freeing of resources, track lengths, train lengths, etc.

Railway design and construction planning is an old engineering discipline with long-standing traditions. Demands for the highest safety, compatibility with existing infrastructure and practices, and high investment costs, make railway engineering a conservative domain. The design process of railways is in practice highly sequential, leading to the known advantages and disadvantages of so-called waterfall process models.

Waterfall-style design processes require that high-level specifications can be written up-front and afterwards implemented without feedback from the implementation process back to the high-level specifications. This also means that verification and validation in waterfall-style design processes is confined to the scope of each separate design activity, or destined to have little hope of improving the design when weaknesses are uncovered.

Unfounded design assumptions which are made early in the early process stages have been known to trickle all the way down to the final stages and require new rounds of design starting from the top, a process which typically takes several years.

These negative effects are typically mitigated by:

a) Re-using proven design concepts, i.e. doing something the same way as somewhere else, where it has already turned out to work well.
b) Allowing sizable margins, e.g. planning the track with more than enough space for safety distances so that it is highly likely that control system engineers will later be able to come up with a safe and performant design.

These mitigations exploit tradition, experience and cross-discipline knowledge in the railway engineers, which in turn contributes to making the engineering community slow-moving and conservative.

However, modern construction practice expects and demands optimization. When space requirements, performance requirements and cost limitations are squeezed to the limits of

the possible, the tradition-based railway engineering approach lacks the methods to accurately reason about the limitations of the finished system from partially finished design plans.

Using *agile verification* of high-level properties from the beginning of a design project, and in every step of the process, allows engineers to better see the consequences of each decision and immediately uncover errors and shortcomings that would otherwise be discovered only months or years later.

Our goal is to develop a verification technique and tool to help engineers specify capacity properties *at design time* and to check these automatically. To be agile, the tool needs to (1) have reasonable running times so that the verification can be run on the fly as the design is being updated by an engineer working in a drafting CAD application, and (2) keep the required input to the minimum of information needed to verify relevant properties. This style of verification gives engineers immediate feedback on their design decisions while requiring small amounts of specification and verification work.

1.1 Problem definition

We consider **the low-level railway infrastructure capacity verification problem**, which we define as follows:

> Given a railway station track plan, including signaling components, rolling stock dynamic characteristics, and a performance/capacity specification, verify whether the specification can be satisfied and find a dispatch plan as a witness to prove it.

Solving this problem subsumes the following railway infrastructure design activities:

a) Low-level **running time** analysis – verify the time required for getting from point A to point B.
b) Low-level **schedulability** analysis – verify frequency of trains arriving at a station, and simultaneous opportunities for crossing, parking, loading, etc.
c) **Combinations** – verify running time requirements on schedulable operations.

1.2 Approach

In this paper we suggest a formalization of capacity requirements as a set of operational scenarios involving a set of trains, a set of locations to visit, and a set of timing constraints.

Verification in this domain can in principle be encoded into the SMT [2,7,31] or PDDL+ [11] languages, essentially resulting in a SAT modulo non-linear real arithmetic problem [12,20]. Many solvers can handle such problems [8,13,9], but we found that the problem size of our test cases, in terms of the number of planned actions and in terms of number of interacting Boolean and non-linear real logic terms, were out of reach for agile verification. Moreover, simplifying the train dynamics to only using constant acceleration $x'' = c$ is too simplistic for real-world engineering. We would like to be able to extend the dynamics equations using, e.g., polynomials of higher order or even numerical integration.

Therefore, we have developed a verification tool chain that uses a CEGAR-loop [6] between a SAT-based planning tool that works on a discrete abstraction of the control system commands, and a discrete event simulation engine (DES) [37] that calculates detailed continuous results for a specific plan, taking the physics of moving trains into account.

The SAT-based planner uses bounded model checking (BMC) [3] where time is reduced to a series of partially ordered actions with unknown durations, and the choice of actions are the available commands in the control system. The DES component verifies the continuous time/space results given the Boolean decisions of control system commands, and adds new SAT constraints excluding unsatisfactory solutions.

The separation of discrete and continuous domains also has the advantage that the simulation component can be extended to handle more complex models, such as engine power curves, tunnel air resistance, curve rolling resistance, train weight distribution, etc., without affecting the planning logic or its computational complexity.

We have tested our method and tool on practical examples from existing infrastructure and ongoing construction projects in collaboration with railway engineers from Railcomplete AS.

This paper extends our work from [26] in the following ways (besides expanding in many places with more clear presentations of the concepts and ideas): *(i)* We provide an expanded introduction to railway control systems and dynamics in Sec. 2, to make the work self-contained. *(ii)* We give more detailed descriptions of the implementation of discrete event simulation in Sec. 4. *(iii)* We have added new results for handling overlaps, overlaps with optional timeout, and swinging overlaps in Sec. 3.2. *(iv)* We have added new results for handling loops in infrastructure and eliminating repetition in dispatch plans in Sec. 3.3.

The paper is organized as follows: Sec. 2 contains an overview of the railway design process and the principles for analysis of these designs. We present a language for capacity specifications, together with examples of how it can be used in construction projects. Sec. 3 describes the tool chain and the solver architecture that we propose to verify performance properties in an agile verification style, integrated in the construction project workflow. We present in detail in Sec. 3.1 how the planner component of our solver is implemented. The simulator component is described in Sec. 4. Sec. 5 contains performance evaluations in a set of relevant case studies. Sec. 6 discusses further related work and presents our conclusions.

2 Problem background

The signaling and interlocking design problem for a railway station takes the track plan as input, typically containing tracks, switches and platforms, and produces the following artifacts:

- Track and trackside component layout, describing the locations of tracks, switches, signals and detectors (see Fig. 2).
- Interlocking specifications, describing the requirements for the logic of the control system (see Fig. 3).

These design artifacts are the subject of verification, i.e., the *model*. Ensuring performance in the context of a construction project consists of verifying *properties* describing a set of trains moving on the tracks and the goals which need to be accomplished by these movements. The design should (1) ensure safe movement while also (2) fulfilling performance requirements. We describe each of these aspects in the sections below.

To verify performance properties, we need to find a sequence of trains and elementary routes for the train dispatcher, i.e., a *dispatch plan*, which when executed under safety and correctness constraints (described in Sec. 2.1 below), demonstrate the properties described in the performance requirements (detailed in Sec. 2.2 below).

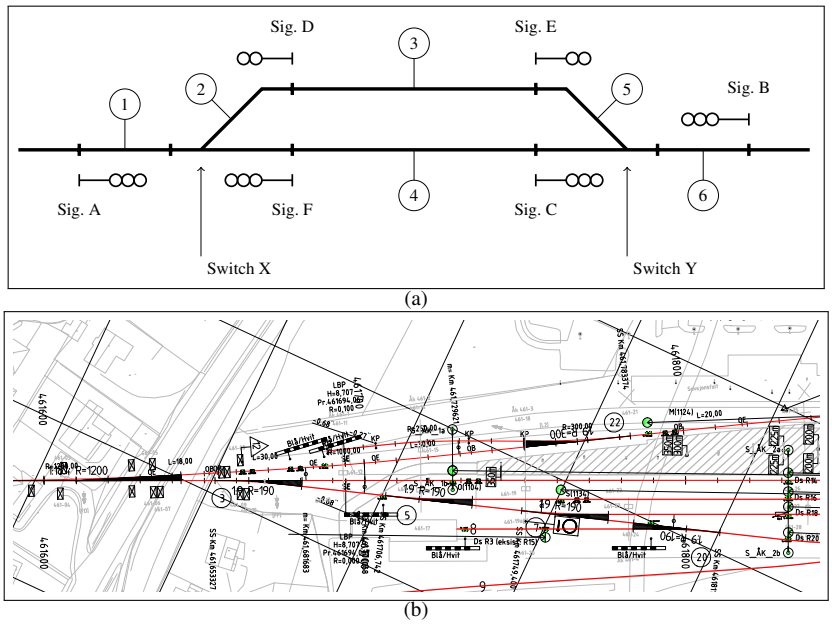

Fig. 2: (a) Example *schematic* construction drawing. (b) Cut-out from a 2D geographical CAD model (construction drawing) of a preliminary design of the Arna station signalling.

Train movements along the railway are coordinated by a train dispatcher, whose task it is to choose which trains go where, and communicate this to the train drivers. The dispatcher uses a control system to perform this task, called the interlocking, which receives input from trackside train detectors and controls movable track elements and signals (see Fig. 4).

2.1 Safe and correct train movements

Low-level analysis of train movements covers a wide range of constraints given by the track layout, the control system, and operational procedures, to be certain that the analysis produces detailed, realistic results. The following subsections give an overview of these constraints, divided into four classes. See [33] for a more in-depth description of railway operation principles.

Elementary route	Start signal	End signal	Switch position	Track segments	Conflicts
AC	A	C	X right	1, 2, 4	AE, BF
AE	A	E	X left	1, 2, 3	AC, BD
BF	B	F	Y left	4, 5, 6	AC, BD
BD	B	D	Y right	3, 5, 6	AE, BF

Fig. 3: Example of a *tabular interlocking*, showing available routes and their conditions.

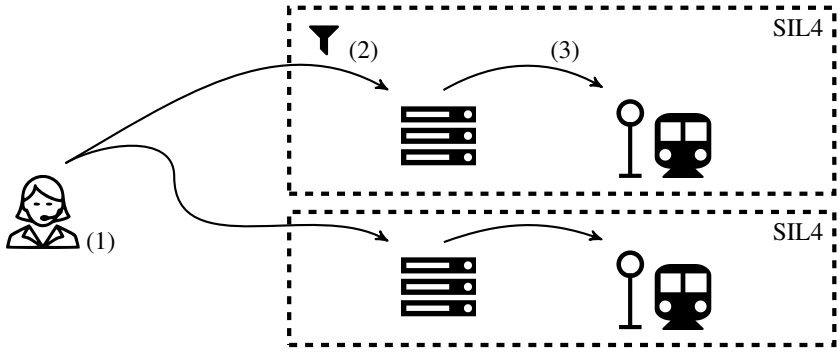

Fig. 4: A dispatcher (1) requests routes from the interlocking control system. The interlocking decides whether to accept the command (2), and signals the resulting movement authority to the train driver (3). The control system itself is responsible for the safety of the resulting movements.

2.1.1 Physical infrastructure

Trains travel on a network of railway tracks which have physical properties such as length, gradient, curvature, etc. Tracks branch off using *switches*, whose *setting* determines where the train goes. Detectors on the track are used by the control system to determine whether track segments are occupied. The physical infrastructure also determines the *sight areas*: the set of locations where a train receives information from a given signal.

2.1.2 Interlocking: allocation of resources

The safety-critical control systems for railway infrastructure are called *interlockings*. An interlocking takes requests for activating routes from a dispatcher. When a route is activated, switches are moved into correct positions and signals are set to show the go aspect. The interlocking is also responsible for assuring that activating the route, i.e., allowing the train to travel the route, is safe. This safety is ensured through the following requirements:

a) Routes require the exclusive allocation of **track segments**, so that two routes which use some of the same track segments cannot be activated at the same time. Routes must be allocated as a unit, i.e., all segments must be free at the time of allocation. However, track segments may be de-allocated to other routes as soon as the train has passed a segment.
b) **Switches** need to be in the correct position for the train to travel along the route. Also, the switches must be locked, so that they cannot accidentally be moved while the train is travelling, and detectors on the switch must report that the switch is actually locked in the correct position.
c) A **safety zone** (also called overlap) beyond the end of the route must be vacant, but not necessarily exclusively allocated, i.e., two safety zones may share track segments. The safety zone is released after a given time which is long enough that it is unlikely that the train is still moving forward. This timeout is calculated based on the length of the route.

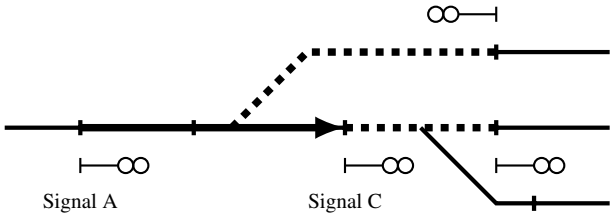

Fig. 5: Elementary route AC from signal A to the adjacent signal C. The thick line indicates track segments on the train's path which are reserved for this movement, and the dashed lines indicate reserved track segments outside the path.

d) Routes which pass through switches require specific track elements to cover any potential movements into the route path. This is known as **flank protection**, and cover can typically be provided by signals, switches or other objects.
e) **Signals** can only show the go aspect when it is the starting point for a currently active route; in all other states, the signal must show the stop aspect. Distant signals, i.e., additional signals showing information about the next upcoming route, must give information consistent with the upcoming signal.

These constraints are explicitly expressed for a given railway station through the interlocking specification, which is an artifact of the design process.

Avoiding collisions by exclusive use of resources is the responsibility of the interlocking, which takes requests from the dispatcher for activating **elementary routes**. An elementary route is the smallest unit of resources that can be allocated to a train, see Fig. 5. Route activation is a process which proceeds as follows:

a) Wait for all **required resources**, such as track segments and switches, to be free. Resources required by a route are typically any resource in the train path (or sometimes outside of it), which ensure that all movements are performed at a safe distance from each other.
b) **Movable elements** (e.g. switches) must be set to correct positions. If they are not, start a sub-process which moves the element into place, and wait for this process to finish before proceeding.
c) **Signals** are then set to show the 'proceed' aspect to the train when the above steps are finished. When the front of the train has passed the signal, it is immediately reset to show the 'stop' aspect.
d) A **release** process is started, which waits for the train to finish using the allocated resources (i.e., to travel over them) and frees them when this has happened.

Influence of safety zones on capacity

The safety zone, as described above, is a set of track sections and switches allocated together with a route to ensure that slightly overrunning a signal showing the stop aspect is safe (see Fig. 6). Different manufacturers and national regulations have various ways of specifying how a safety zone is released and how alternative safety zones are implemented. The main variations are:

a) The safety zone from a route end point persists until a route is allocated from the end point. This can be problematic if the safety zone blocks other traffic or if the train is

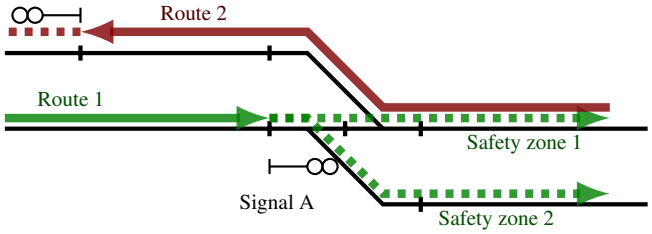

Fig. 6: An elementary route 1 ending in signal A can protect trains from overrunning the signal by allocating one of the safety zones (shown as safety zone 1 and 2). In some situations, safety zone 1 might be preferred so that the switch following signal A is in the correct position for letting the train in route 1 proceed quickly. However, allocating safety zone 1 blocks route 2 from use. So in other situations, safety zone 2 might be preferred, for example for two trains to concurrently enter a station. Some control systems may allow one safety zone to be replaced by another after the route has been allocated.

changing directions and not proceeding past the end point. The following two methods are the usual mitigations for these problems.

b) The safety zone is released after a pre-set time. This time should be long enough so that the probablilty that the train is still running towards the end point is very low.
c) The safety zone it not released, but can be replaced by another safety zone from the same route end point. This method is called *swinging overlap* in the United Kingdom.

2.1.3 Communication constraints

After movement has been allowed by the control system, the driver must be informed of this fact. When a route is activated, a train inside the sight area of the route's entry signal reads the signal's message that movement authority is given. The train driver may then drive the train forward until the next signal. The following types of signalling systems are common in railways:

a) Traditional signaling with trackside lamps. Communication is limited by how many different aspects the lamps can show. To avoid high-speed trains slowing down at every signal, several consecutive elementary routes can be signaled in advance using so-called distant signals.
b) Automatic train protection systems (ATP) work similarly to signals, but may give more information. Many ATP systems communicate information through magnets or short-range radio at specific locations on the track, corresponding to a signal sight area of zero length.
c) The European Rail Traffic Management System (ERTMS) currently being implemented in many European countries replaces lamp signals with trackside marker boards, and uses long-range radio for communication. This effectively removes the communication constraint, as the radio can be used to update any train's movement authority at any time.

The amount of information that can be transmitted to the train drivers through the signaling puts a constraint on how far ahead the routes can be pre-allocated. Traditional signaling can typically show information about either one or two routes, but some countries have extended to information about three consecutive routes. It is also common to extend the information given by signals using track-side electronic communication. See Fig. 7.

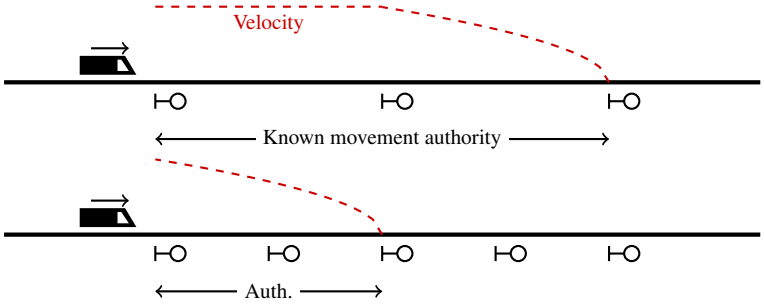

Fig. 7: Signal information only carries across two signals (so-called *distant signals*).

2.1.4 Laws of motion

Trains move within the limits of given maximum acceleration and braking power, so train drivers need to plan ahead for braking so that the train respects its given movement authority and speed restrictions at all times.

The speed increase from v_0 to v over a time interval Δt is limited by the train's maximum acceleration a:

$$v - v_0 \leq a\Delta t \ .$$

However, when there is a more restrictive speed restriction ahead, the driver must start braking in time to meet the restriction. A signal showing the 'stop' aspect can be treated as a speed restriction of zero. Since speed restrictions change with time, the driver must re-evaluate their actions whenever new information is received.

A train has the following constraint on its velocity v for each restriction,

$$v^2 - v_i^2 \leq 2bs_i \ ,$$

where v_i is the maximum allowed speed, s_i is the distance to the location where the restriction starts, and b is the maximum retardation achieved by braking. These restrictions are given as (1) constant maximum velocity restrictions given by signs beside the track, or (2) dynamical velocity restriction given by the distance to the next stop signal (i.e., the length of the movement authority).

2.2 Performance requirements specifications language

To capture typical performance and capacity requirements in construction projects, we define an **operational scenario** $S = (V, M, C)$ as follows:

a) A set of **vehicle types** V, each defined by a length l, a maximum velocity v_{max}, a maximum acceleration a, and a maximum braking retardation b.
b) A set of **movements** M, each defined by a vehicle type and an ordered sequence of visits. Each visit q is a set of alternative locations $\{l_i\}$ and an optional minimum dwelling time t_d.
c) A set of **timing constraints** C, each constraint consisting of two visits q_a, q_b, and an optional numerical constraint t_c on the minimum time between visit q_a and q_b. The two visits can come from different movements. If the time constraint t_c is omitted, the visits are only required to be ordered, so that $t_{q_a} < t_{q_b}$.

To demonstrate how an operational scenario captures requirements of railway construction projects, we give some examples using the syntax of the file format used in our tool[1]. First, we define the following vehicle types:

```
vehicle passengertrain length 220.0
  accel 1.0 brake 0.9 maxspeed 55.0
vehicle goodstrain length 850.0
  accel 0.5 brake 0.5 maxspeed 20.0
```

The following set of **performance specifications** are selected prototypical versions of specifications that railway engineers have suggested as useful for automated verification:

a) **Running time**: expresses an expectation of how long it should take for a train to travel between two locations. To specify this, we simply require that a train visits some location `b1` and later visits some other location `b2`. A timing constraint of 90.0s between these visits sets the running time requirement.

```
movement passengertrain {
  visit #a [b1]; visit #b [b2] }
timing a <90.0 b
```

b) **Train frequency**: a train station processes a set of trains arriving and departing with a fixed frequency. On a two-track station, we exemplify a sequence of four trains and their relative departure times as:

```
movement passengertrain {
  visit [b1]
  visit [platform1,platform2] wait 60.0
  visit #e1 [b2] }
// ...3 more trains with visits e2, e3, e4.
timing e1 <90.0 e2
timing e2 <90.0 e3
timing e3 <90.0 e4
```

c) **Overtaking**: trains traveling in the same direction can be reordered. For example, we specify a passenger train traveling from `b1` to `b2`, and a goods train with the same visits. Timing constraints ensure that the passenger train enters first while the goods train exits first.

```
movement passengertrain {
  visit #p_in [b1]; visit #p_out [b2] }
movement goodstrain {
  visit #g_in [b1]; visit #g_out [b2] }
timing p_in < g_in
timing g_out < p_out
```

d) **Crossing**: trains traveling in *opposite directions* can visit this station simultaneously. This example is similar to the previous one, but the goods train now travels in the opposite direction, and the timing constraints require that the trains are inside the model simultaneously.

```
movement passengertrain {
  visit #p_in [b1]; visit #p_out [b2] }
movement goodstrain {
  visit #g_in [b2]; visit #g_out [b1] }
timing p_in < g_out
timing g_in < p_out
```

[1] For details of the input file formats, see `https://luteberget.github.io/rollingdocs/usage.html`

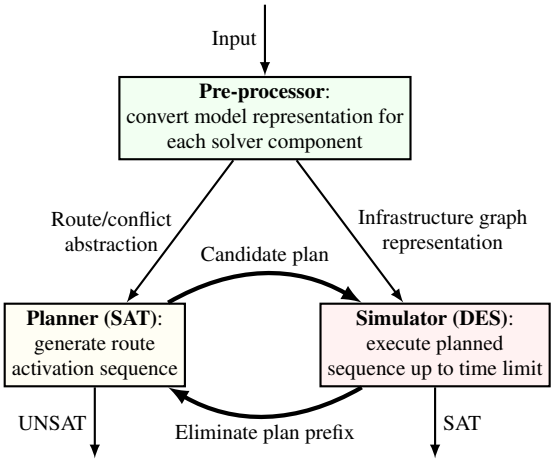

Fig. 8: Conceptual diagram of our CEGAR architecture. Infrastructure, routes, train types, and movement specifications are transformed into (1) the planner's abstract representation, containing only elementary routes and train lengths, and (2) the detailed graph representation used in the simulator component.

Similar specifications, and combinations of such specifications, are relevant in most railway construction projects. Since we typically only need to refer to locations such as model boundaries and loading/unloading locations, these specifications are not tied to a specific design, and can often be re-used even when the design of the station changes drastically.

Stations can be designed to be either intermediate stops, or end-of-line stops. Also, some stations are intermediate stops for some trains while also being the end-of-line stop for other trains. A challenge of planning for end-of-line stops is that the train must usually be allowed to turn around and go back in the direction it came from. Allowing an unbounded number of such turns can in principle lead to an infinite number of dispatch plans.

The turning is also related to the challenge of having loops in the infrastructure. This is uncommon within a single station, but can sometimes occur in construction projects where several stations together form a loop topology. Also here, we must take care not to explore or suggest an infinite number of ways to execute an operational scenario. Such aspects are treated in Sec. 3.2.

3 Tool chain and solver architecture

Being able to do performance verification (i.e., capacity analysis) automatically, using only information typically available in a construction project, is a valuable tool for railway engineering in the disciplines of signalling and interlocking. In this section, we describe our approach to verifying capacity properties using the formalization of capacity given in the previous section.

We have investigated several logic-based approaches for the domain and problem described above. The PDDL+ language has been designed to express planning problems in mixed discrete/continuous domains. As each discrete change is represented by a planning step, our test case problem instances would need at least 50-100 steps to be solvable. We

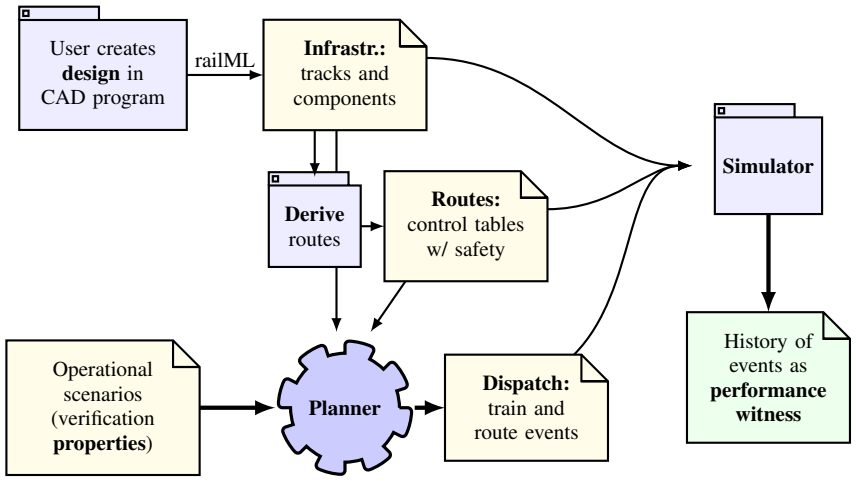

Fig. 9: Capacity verification tool chain overview. Yellow boxes represent input documents. Note that only infrastructure and operational scenarios are strictly required – interlocking tables can be derived, and dispatch plans can be synthesized. Blue boxes represent programs. The green box represents the output document from the simulator, which is a history of events which is the witness that proves the performance requirements.

were able to solve using the SMTPlan+ solver only the most trivial test cases in less than one second, which we consider a reasonable running-time constraint for agile verification.

Encoding into SMT can be done by expressing planning as a BMC problem. This approach suffers from the same problem of having a high number of planning steps (some improvements can be made, s.a. making train driver choices implicit in constraints on the relation between velocity, distance and time).

To address these limitations, we developed a CEGAR-style tool which exploits the limited number of control system commands to make an abstraction of the planning problem, see Fig. 8. A verification tool chain which solves the low-level railway infrastructure capacity verification problem and supports agile verification in railway construction projects is outlined in Fig. 9 along with information flow between the components. The manual, source code and test cases are available online[2]. The tool uses the MiniSAT v2.2.0 solver.

The tool is complementary to other verification techniques in railway design, such as static layout verification [28,27,25], static interlocking verification [17,27], interlocking program verification [4], and timetable analysis [16].

The following input documents are used:

a) **Operational scenarios** defining the performance properties to verify. Examples are given in Sec. 2.2.
b) **Infrastructure** given in the railML format [30,36]. In our case studies, railML was generated using the RailCOMPLETE software, a plugin for the widely used AutoCAD drafting software. In this way the model is taken directly from the engineers' drafting program with no additional model preparation needed.

[2] `https://luteberget.github.io/rollingdocs` and `https://github.com/koengit/trainspotting`

c) **Elementary routes** (*optional*), given in a custom format which is compatible with the upcoming railML interlocking format. Although subject to design, a decent guess of the content can be straight-forwardly derived from the infrastructure by listing resources on paths between adjacent signals, so this input is optional.
d) **Dispatch plans** (*optional*) corresponding to each operational scenario. The verification tool can produce dispatch plans fulfilling the performance specification, so this input is optional.

An advantage of the separation of planner and simulator is that each component can be used separately. *The planner alone* may be used to enumerate different possibilities for train movements, which might be used in an operational testing situation. *The simulator alone* may be used to debug the execution of a specific dispatch plan to examine performance deficiencies, and educationally for demonstrating the workings of the railway system. Put together, the two components provide automated verification, which is the main goal of our efforts. It would also, in principle, be possible to use one of the commercial simulation packages, such as OpenTrack or RailSys, provided that all input and simulation control can be given though a programmable interface (API).

3.1 Dispatch Planning using SAT

The planner solves the abstracted discrete planning problem of finding a dispatch plan, i.e., determining a sequence of trains and elementary routes which make the trains end up visiting locations according to the movements specification.

We encode an instance of the abstracted planning problem into an instance of the Boolean satisfiability problem (SAT). We consider the problem a model checking problem, and use the technique of bounded model checking (BMC) to unroll the transition relation of the system for a number of k steps, expressing state and transitions in propositional logic.

Using BMC for planning works by asserting the existence of a plan, so that when the corresponding SAT instance is satisfiable, it proves the fulfillment of the performance requirements and gives an example plan for it. When unsatisfiable, we are ensured that there is no plan within the number of steps k. In practice plans with higher number of steps are not of interest; i.e., the bound k is chosen based on practical considerations (e.g., twice the number of trains was sufficient in our case studies). The SAT instance is built incrementally by solving with $k-1$ steps and then adding the k^{th} step if necessary.

The abstracted planning problem is encoded as a SAT instance by representing states, constraints on each state, and constraints on consecutive states. *State* i of the system in the planner component is represented as:

a) Each route r_j has an **occupancy status** $o^i_{r_j}$ which is either free ($o^i_{r_j} = \text{Free}$) or occupied by a specific train t_k ($o^j_{r_i} = t_k$). Each combination of route and train is represented by a Boolean variable, but we will write constraints with $o^i_{r_j}$ as a variable from the set of trains.
b) Each route also has a choice from its associated **safety zones** $z^j_{r_j} \in \{1 \dots n\}$ which determines which other routes are considered to be in conflict (see Fig. 10).
c) Each train t_k has a Boolean representing **appearance status** b^i_k, used to propagate to future states that a train has started (used in constraint C2 below).
d) Each visit l has a Boolean representing **required visits** v^i_l, which is used to propagate to future states that a visit requirement has been fulfilled (used in constraint C5).

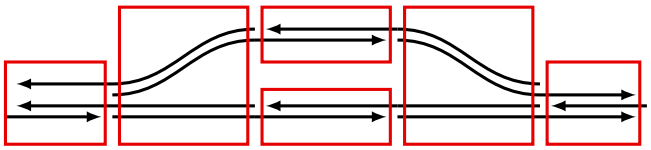

Fig. 10: The planner component takes an abstracted view of the railway infrastructure. Lines represent elementary routes with traveling direction given by the arrows. Boxes indicate routes in conflict, i.e. only one of them can be in use at a time.

e) Each combination of route r_j and train t_k has a Boolean representing **deferred progress** $p^i_{j,k}$, used to propagate to future states that a train is not progressing, and must resolve the conflict in the future (used in constraint C8).

A dispatch plan is produced directly from the occupancy status $o^i_{r_j}$ and safety zone choices $z^j_{r_j}$ of states by taking the difference between consecutive states and then dispatching any trains and routes which become active from one state to the next. If swinging safety zones (also known as *swinging overlaps*) or safety zone timeouts are enabled, then consecutive steps can have different safety zones, and when this happens, a swing command is also added to the dispatch plan.

Constraints are applied to each state and each pair of consecutive states to ensure that:

– The plan is viable for execution (i.e., **correctness**):
(C1) Conflicting routes are not activated simultaneously.
(C2) Each train can only take one contiguous path.
(C3) An elementary route must be allocated as a unit, but its parts may be deallocated separately.
(C4) (Partial) routes are deallocated only after a train has fully passed over them.
– The **plan fulfills capacity specifications**:
(C5) Trains perform their specified visits.
(C6) Visits happen in specified order.
– Equivalent solutions are eliminated (for **performance**):
(C7) Routes are deallocated immediately after the train has fully passed over them.
(C8) A train's path is extended as far as possible in the current time step, unless hindered by a conflicting train (i.e., **maximal progress**).

Equivalent plans, which result in the same trains traversing the same paths and conflicting in the same locations, should have the same representation so that enumeration of different plans produces meaningful alternatives. For example, the two dispatch plans for crossing shown in Fig. 1 are the only two alternatives given by the planner for this operational scenario. See Fig. 11 for other dispatch plans which fulfil the correctness constraints (C1-6) but which do not have maximal progress in each state.

The simulator component, which evaluates the time consumption of plans, reports which parts of the plan fail the timing constraints, and the negation of this partial plan is added to the SAT instance. Since the timing calculations are path dependent, we use the part of the plan starting from the beginning and going up to the step where the timing specification violation occurs. This way of refining the abstraction can cause performance problems when many different choices are possible early in the plan, and the timing violation can only be found near the end of the plan, as demonstrated in Sec. 5. Finding a way to make more precise refinements could be necessary for larger problem instances.

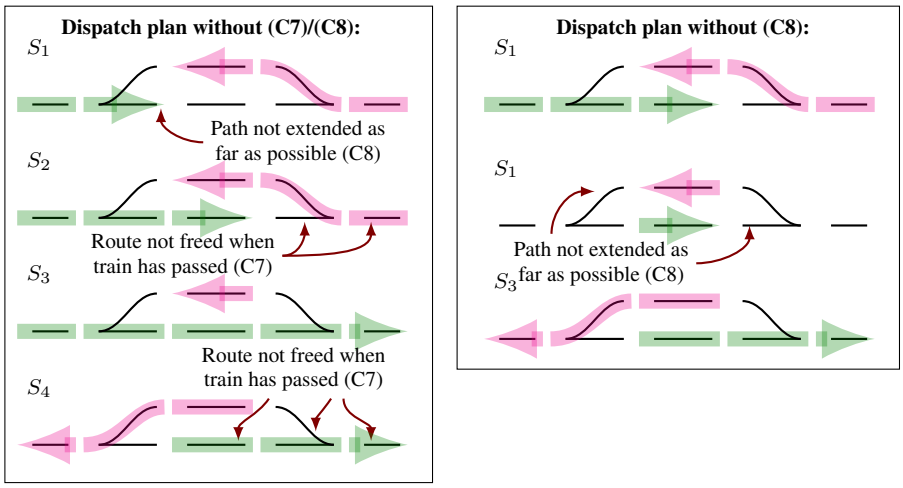

Fig. 11: Examples of dispatch plans which are correct plans (constraints (C1-6)), but which have better equivalent descriptions that allocate and deallocate as soon as possible. These plans do not fulfil constraints (C7) and (C8). Compare with plan 1 in Fig. 1.

The implementation of each of these constraints as propositional logic statements is described below. Constraints apply separately to all states i unless noted otherwise.

3.1.1 Resource conflicts (C1)

Any two routes which require the same resources cannot both be allocated in the same state.

$$\forall r_a \in \text{Routes} : \forall r_b \in \text{conflict}(r_a) : o^i_{r_a} = \text{Free} \vee o^i_{r_b} = \text{Free}.$$

3.1.2 Train path (C2)

At most one alternative route is taken by a train in a single state. First, ensure that only one route from a given start signal may be taken at any time.

$$\forall t \in \text{Trains} : \forall s \in \text{Signal} : \text{atMostOne}(\left\{ o^i_r = t \mid \text{entry}(r) = s \right\}) .$$

We use a standard sequential encoding to encode atMostOne and other similar constraints, as explained in e.g. [38]. Note that entry signals for all routes entering from a model boundary share the same null value, so that this constraint also excludes plans where a single train appears in several positions at once. Each train should only enter the plan once, thus the appearance Boolean changes to true in exactly one transition.

$$\forall t \in \text{Trains} : b^i_t \Rightarrow b^{i+1}_t.$$

$$\forall t \in \text{Trains} : \text{exactlyOne}\left(\left\{ \neg b^j_t \wedge b^{j+1}_t \mid j \in \text{States} \right\}\right).$$

A train appears when an entry boundary route is allocated:

$$\forall t \in \text{Trains} : \forall r \in \{ r \in \text{Routes} \mid \text{entry}(r) = \text{null} \} : \left(o^i_r \neq t \wedge o^{i+1}_r = t \right) \Rightarrow b^{i+1}_t.$$

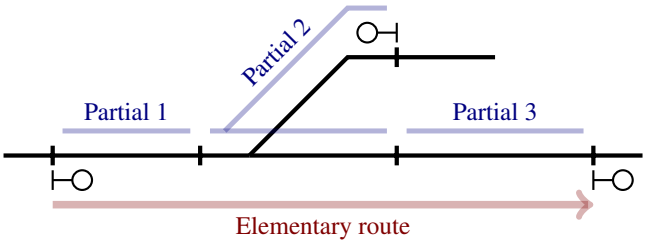

Fig. 12: The planning abstraction of the train dispatch allocates a set of partial routes to each train. Elementary routes are sets of partial routes which must always be allocated together.

Routes which are not entry routes can only be allocated to a train when they extend some other route which was already allocated to the same train, i.e., consecutive routes must match so that the exit signal of one is the entry signal of the next:

$$\forall t \in \text{Trains} : \forall r \in \{r \in \text{Routes} \mid \text{entry}(r) \neq \text{null}\} :$$
$$\left(o_r^i \neq t \wedge o_r^{i+1} = t\right) \Rightarrow \bigvee \left\{o_{r_x}^{i+1} = t \mid r_x \in \text{Routes}, \text{entry}(r) = \text{exit}(r_x)\right\} .$$

Note that this constraint ensures that the trains' allocation to routes *locally* forms a path in the graph of routes. In the presence of cycles, this constraint does not rule out cyclic allocations disjoint from the rest of the train's path. This problem is handled separately in Sec. 3.2 below.

3.1.3 Partial release (C3)

Partial release is represented by splitting each elementary route into separate routes for each component which is released separately. The set *Partial* contains such sets of routes. Partial routes are allocated together (see Fig. 12):

$$\forall t \in \text{Trains} : \forall q \in \text{Partial} : \text{allEqual}(\left\{o_r^i \neq t \wedge o_r^{i+1} = t \mid r \in q\right\})$$

3.1.4 Deallocation (C4, C7)

Routes are freed when sufficient length has been allocated ahead to fully contain the train.

$$\forall t \in \text{Trains} : \forall r \in \text{Routes} :$$
$$o_r^i = t \Rightarrow \left((o_r^{i+1} = t) \Leftrightarrow \text{freeable}_{r,t}(\left\{o^i\right\})\right) .$$

Note that the equality sign on the right hand side implies that deallocation is both allowed (C4) and required (C7). The freeable predicate is a disjunction of paths (conjunction of routes) ahead which are long enough to contain the train. For example, on the routes shown in Fig. 13, if route A holds a train t of length 200.0 m, freeing A is constrained by:

$$A^i \Rightarrow \left(A^{i+1} \vee (B^i \wedge C^i) \vee (D^i \wedge E^i)\right) .$$

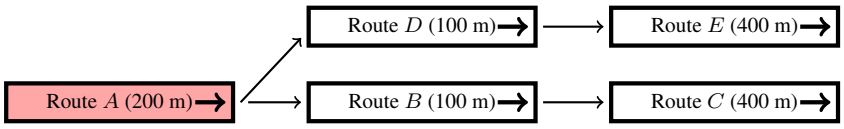

Fig. 13: When a train of length 200.0 m has been allocated to route A, that route can only be freed when the train has been allocated to either both B and C or both D and E.

3.1.5 Visits (C5, C6)

Visits and their order are given by the set VisitOrder, which contains pairs of (t, v), where t is a train and v is a set of alternative routes. Visits must happen using any of the alternative routes, and must be in an order such that the visit (t_1, v_1) comes before (t_2, v_2):

$$\forall((t_1, v_1), (t_2, v_2)) \in \text{VisitOrder} :$$
$$\bigvee \left\{ o^i_{r_a} = t_1 \wedge o^j_{r_b} = t_2 \wedge i \leq j \;\mid r_a \in (v_1), r_b \in (v_2), i, j \in \text{States} \right\} .$$

3.1.6 Forced progress (C8)

In addition to the constraints on allocation and freeing required to produce a valid plan, we also add constraints which force each train to get allocated routes further along a path forward unless there is a conflict. Routes ahead are either allocated, or the train is deferred p:

$$\forall t \in \text{Trains} : \forall r \in \text{Routes} : o^i_r \Rightarrow p^i_{t,r} \vee \bigvee \left\{ o^i_{r_x} \mid r_x \in \text{Routes}, \text{entry}(r_x) = \text{exit}(r) \right\}$$

Deferred progress must be resolved by freeing a conflicting route, and then allocating it to the train in the following step:

$$\forall t \in \text{Trains} : \forall r \in \text{Routes} :$$
$$p^i_{t,r} \Rightarrow p^{i+1}_{t,r} \vee \bigvee \left\{ o^i_{r_c} \neq \text{Free} \wedge o^i_{r_x} \neq t \wedge o^{i+1}_{r_x} = t \right.$$
$$\left. \mid r_c, r_x \in \text{Routes}, \text{exit}(r) = \text{entry}(r_x), r_c \in \text{conflict}(r) \right\}$$

When i is the last state, $p^{i+1}_{t,r}$ is considered to be false, which forces the deferred progress to be resolved eventually. Note that it is not required that the conflicting trains are distinct.

3.2 Handling turning and loops

Many railway construction projects have only *acyclic* infrastructure, in the sense that trains enter from one side of the station and exit on the other side, and all paths from one side to the other are acyclic. However, if the infrastructure has a same-directed cycle which can be allocated without conflicting with other routes, the constraints C2 above are insufficient to ensure train path consistency, see Fig. 14. The train path consistency constraints described in the previous section require each active route to have a route before it already being active. This works in the acyclic case, because the chain of routes always leads back to either a model boundary or a route already allocated in the previous step. With cyclic infrastructure,

however, a sequence of routes can justify each other, which would lead to a train appearing out of nowhere. It is a known problem that expressing this kind of constraints in SAT can be very inefficient (see e.g. [23, 14]), and to handle same-directed cycles in the infrastructure, we add instead a refinement step around the SAT solver which searches each state for this kind of circular reasoning and adds a single constraint each time this situation appears.

The loop check procedure checks for each train t_i and for each state s_j, whether the set of routes R_i^j allocated to the train has any strongly connected components $\text{scc}_i^j \subseteq R_i^j$ with $|\text{scc}_i^j| > 1$, and in that case adds a new constraint to the SAT problem:

$$\bigvee \left\{ \neg(o_r^j = t_i^j) \mid r \in \text{scc}_i^j \right\} .$$

Fixing these consistency errors gives valid plans even in the presence of same-directed infrastructure cycles, but even planning on infrastructure without cycles may cause repetition to appear in the dispatch plans. For example, at the end of a railway corridor, trains must be able to switch directions and go back to where they came from. In the description of dispatch planning above, if trains are allowed to stop and reverse their direction, the directed graph of routes becomes cyclic, and there is in principle an infinite number of different possible dispatch plans for any train movement.

Allowing trains to turn, and allowing loops in the infrastructure, will lead to the bounded model checking planning method finding more and more solutions when increasing the number of steps. Most of these solutions will exhibit some amount of repetition in the movement of trains, and this makes them of little value to the railway engineer. We suggest some different solutions to this challenge below, roughly ordered by how complex the implementation would be and how much quality would be improved:

– **Unlimited:** it could be feasible to have no limit on turning of trains, and no limit on the use of loops in the infrastructure. Since the bounded model checking technique will find the shortest plans first, they will often be the most valuable plans for the engineer, and the planner can be aborted when plans get too long and repetitive and as such are no longer valuable for the verification of the design. However, the fully automated verification tool would have to set a carefully considered upper bound on the number of plan steps.

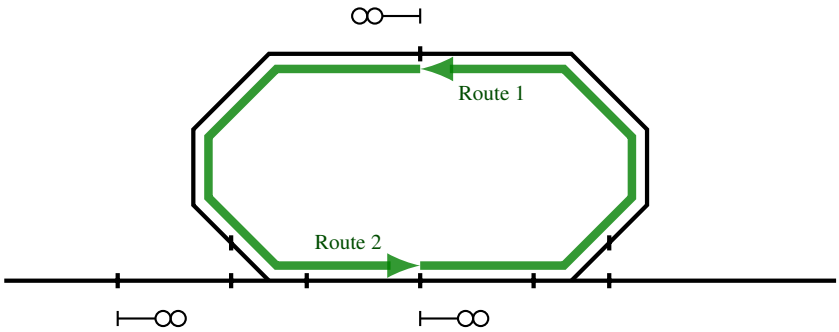

Fig. 14: Example of cyclic infrastructure. Here, to ensure train path consistency (C2), additional constraints are needed over the acyclic case. Route 1 and route 2 both provide each other's justification for a train appearing there, possibly making an error of circular reasoning.

- **Specified turning:** the specifications of the operational scenarios can be extended to include turning explicitly at visits. This increases the specification burden on the engineer, but ensures that there cannot be an unbounded number of distinct plans. However, it could also cause some plans to stay undetected if they require turning and the engineer did not think of it. Also, this method does not help the situation with loops in the infrastructure.
- **Bounded number of turns:** instead of writing out each turn explicitly, the capacity specifications could be extended to include an upper bound on the number of turns. The bound would have to be adjusted to balance running time and plan quality (low bound) with the possibility of detecting more complex plans (high bound).
- **State space repetition constraint:** to ensure that the whole state of the system does not repeat from one stage to another. This requires adding a constraint on each pair of states, which could make the SAT instance significantly larger.

$$\bigwedge_{0 \leq i < j < k} S_i \neq S_j.$$

  Such constraints may also be added lazily, i.e. by incrementally adding the constraints only when they are violated in a SAT solution (see [10]). This constraint would eliminate the possibility for an infinite number of distinct plans, but could still cause unnecessary repetition locally, since repetition in one part of the model could be accompanied by progress in another part of the model.
- **Repetition filtering:** even when the state as a whole does not repeat, there may be sequences of allocation to a subset of trains which can be considered repeating. We would like a more domain-specific definition of repetition, based on a graph analysis of the dispatch plans produced. This can be implemented by rejecting solutions which exhibit such repetition. We define this more carefully in the section below.

As we find the last option to be the most complete solution requiring no change to the specifications, we describe its implementation here in more detail.

3.3 Filtering out unnecessary repetitions

We now define the notion of *unnecessary repetitions* and show how to identify them on a given dispatch plan. First, we define the notions of *yield* and *repetition*.

A train t_1 *yields* to another train t_2 if t_2 is occupying a route whose resources are needed for t_1 to proceed (thereby allowing t_1 to defer its progress as defined in constraint (C8), Sec. 3.1.6). More precisely, if t_2 occupies some route r_2 in state s, and t_1 allocates a route r_1 in state $s+1$, where r_1 conflicts with r_2, we say that t_1 yielded to t_2 in state s.

Now, consider a train t that enters the model from some model boundary and exits through another boundary by traveling a sequence of routes $r_1, \ldots, r_{m+1}$, which we call the train's *path*. For each pair of consecutive routes r_i, r_{i+1}, the exit signal of r_i is the same as the entry signal for r_{i+1} (described as constraint C2 in Sec. 3.1.2), which we call the delimiting signal $u_i = \mathrm{delim}(r_i, r_{i+1})$ between the routes r_i and r_{i+1}. We say that the train visits the sequence of signals $u_1, \ldots, u_m$ defined in this way.

A signal appearing several times in this sequence ($u_i = u_j$ with $i < j$) indicates a cycle in the train path. Let $s_a = \mathrm{alloc\_state}^t(r_i)$ be the state where route r_i starting in u_i is allocated to t, and let $s_b = \mathrm{alloc\_state}^t(r_{j-1})$ be the state where route r_{j-1} ending

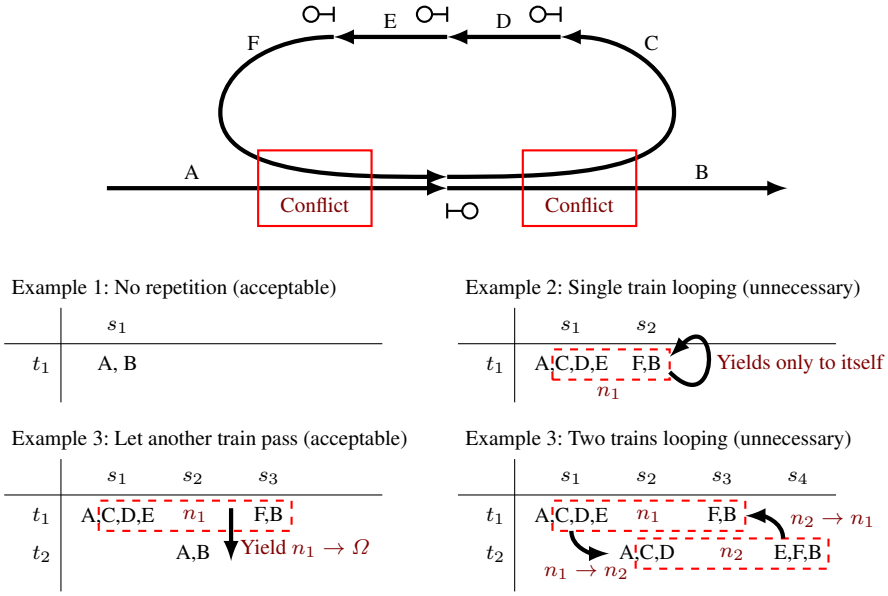

Fig. 15: Examples of repetition justification using yields, demonstrating acceptable and unnecessary repetitions. Each of the routes A, B, C, D, E, and F shown in the infrastructure route graph is long enough to contain each train completely. Examples use trains t_1, t_2 and states s_1, s_2, s_3, s_4. Repetitions are shown as red dashed boxes, and yields are shown as arrows between repetitions.

in u_j is allocated to t. We say that the train t *repeats* on the interval s_a to s_b and write $\mathrm{repeat}(t, s_a, s_b)$.

In most cases, we would like to disallow such repetitions, but there are two exceptions. Firstly, if the train fulfils a specified visit on the state interval s_a to s_b (see constraint (C5), Sec. 3.1.5), the repetition is acceptable. Secondly, if the train yields to another train in a state s_y such that $s_a \leq s_y \leq s_b$, we say that the yield *justifies* the repetition. For example, if a train goes into a siding track to allow another train to pass by, the first train could reverse into the main track again to proceed, thereby performing a repetition that is acceptable. See Fig. 15 for a few examples. However, if one repetition is justified by yielding to another train in a state which also has a repetition that is justified by yielding back to the first train, this does not make these repetitions acceptable. We would like to disallow such circular justifications, and we formalize this using the *yield justification graph*, $G = (V, E)$, defined as the directed graph where:

- The set of nodes N contains each repetition, $\mathrm{repeat}(t, s_a, s_b)$, and a special non-repetition node Ω.
- The set of edges E contains the edge $n_1 \to n_2$, where $n_1 = \mathrm{repeat}(t_1, s_a, s_b)$ and $n_2 = \mathrm{repeat}(t_2, s_c, s_d)$, whenever these nodes n_1, n_2 exist in N and t_1 yields to t_2 in a state s where $a \leq s \leq b$ and $c \leq s \leq d$.
  However, if $\mathrm{repeat}(\mathrm{t_1}, \mathrm{s_a}, \mathrm{s_b})$ exists, and t_1 yields to t_2 in state $s_a \leq s \leq s_b$, but there are no matching repetitions n_2, then the edge $n_1 \to \Omega$ is included instead.

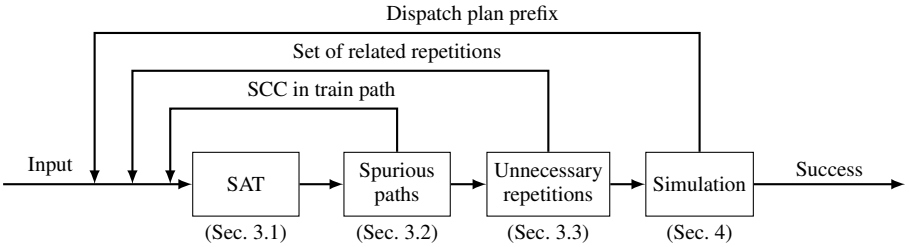

Fig. 16: Main algorithm for local capacity verification (extended from Fig. 8) with two more tests for handling loops and repetitions.

We say that a repetition is acceptable if Ω is reachable from the repetition's corresponding node in the yield justification graph. A repetition that is not acceptable by these two criteria, is an *unnecessary repetition*, and we discard the candidate dispatch plan and add a new constraint to the SAT problem to disallow it using the relevant component of the yield justification graph. This adds another kind of abstraction refinement to our algorithm, see Fig. 16.

The methods for handling both loops and repetitions described here may cause performance problems on certain inputs. However, we have not encountered any real-world examples where this dominates the solver's performance.

4 Timing Evaluation using Simulation

For evaluating the behavior of a railway system in full detail, there are various well-known simulation approaches which are routinely successfully used to analyze railway capacity. Because a simulation works by starting in a known state and applying known input to the system, it proceeds by executing imperative code to change the system state and to register event handlers to processes. Deterministic simulation models can handle very complex models in a short amount of time, but unlike a planning model, one cannot prescribe which state the simulation will end up in, only measure the outcome. Simulation methods are commonly used to develop and assess time tables, and by introducing stochastic elements in the model and repeating the simulation a large number of times, the robustness of a time table can be analysed (e.g., see [32]).

Discrete event simulation (DES) is a simulation technique based on assuming that changes to system state happen only at a set of discrete points in time, so that the simulation can progress efficiently by jumping from one point in time to the next point in time where an event is scheduled. This simulation assumption can be made to work even for the continuous dynamics of train movements, because we assume that each train's dynamics do not interact directly with other train's dynamics. Trains exist in separate worlds which are only connected to each other through the control system, and the control system has only discrete state changes. Each train acts separately on the information it has received from signals so far, and needs only to predict how long it will take to reach the next signal or sensor where it interacts with the control system.

In our tool architecture, the planner component works on an abstraction of the simulation problem that is just detailed enough to ensure that trains end up where they are specified to go, and that the system does not enter a dead-lock state. This is the reason why the

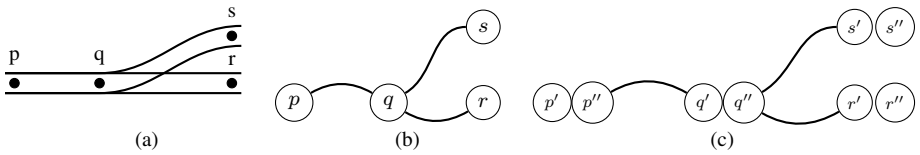

Fig. 17: (a) On the railway network, paths p-q-r and p-q-s exist, in both directions. (b) In a conventional undirected graph representation, there would also be a path r-q-s. (c) When the graph is extended to include two sides of each node, there is no longer a path r-q-s.

planning model must include safety zones, partial release and the lengths of routes and trains – the sequences of routes and trains are represented precisely so that we know what to expect during the simulation. If it turns out that the planner's assumptions about where the trains end up does not work out correctly in the simulator, then the correspondence between planning and simulation is broken, which may be a modeling error in the simulator or errors in the route specifications, for example if the switches are configured to turn in the wrong direction. Runnning the capacity verification assumes that the route specifications are correct, and this may be verified through other means (e.g., see [39]).

4.1 Implementation

For our capacity verification tool for railway construction projects, we have implemented a simulation program using techniques described in [18]. We provide a brief overview here of the main components of this simulation program.

4.1.1 Infrastructure and interlocking specifications

For the work in this paper, we have implemented a simulation system for railways containing main signals, detector, switches, routes, trains, partial release, safety zones and more.

The input of railway infrastructure consists of *nodes*, representing locations on the tracks where transfer of information between the infrastructure and the train can happen. Objects include switches, detectors, signal sighting locations, and points of discrete changes in track properties such as radius and gradient. Nodes are connected by *edges*, which have a specified length. Edges are not directed, because tracks can be traversed in both directions, so each train refers to the edge it is travelling on as an ordered tuple of nodes. For example, a train travelling between node a and node b will store its current location as either (a, b) or (b, a), depending on the direction of travel.

However, a simple undirected graph model also lacks the information that the train needs to figure out which edges can be followed while still travelling in the same direction. Representing this as a directed graph would require deciding on a *global* notion of direction, which is not compatible with cyclic infrastructure graphs where a train can travel back into the same track in the opposite direction.

A more suitable data structure for simulating railway networks is the *double node graph* described in [29] where each node of a conventional graph is represented as a two linked nodes representing each of the two sides for approaching each track location, see Fig. 17. A train reaching a node may only proceed by traveling on edges starting in the opposite node. Also, signals typically only apply in one of the travelling directions, so a train passing a pair

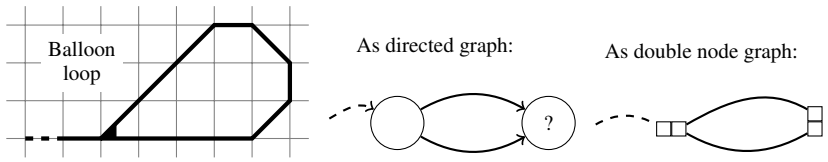

Fig. 18: The balloon loop infrastructure is an example where directionality of travel cannot be suitably captured as a directed graph.

of nodes only reacts to the objects that are located on the *exit side* of the node pair. This model allows for a *local* notion of directedness, and avoids deciding on a global direction concept such as up/down or outgoing/incoming often used in railway engineering. A global directionality requires considering special cases to handle railway networks where a train's up/down direction may change without the train reversing its direction, such as the balloon loop example which is commonly seen on tram lines, see Fig. 18.

4.1.2 Dispatching trains

From the planner we extract the following **dispatch plan**, which serves as the external events input to the simulation.

- **Start train**: start a new train process with given train parameters, initial velocity, and a route entering from a model boundary.
- **Activate route**: start a new route activation process for the given route.
- **Swing safety zone**: replace one active safety zone with another.

Note that the times at which these events happen are not given by the planner, only their order in time.

All the rest of the simulation output is determined from these inputs. The inputs start processes in the simulator, which may in turn start other processes. When all processes have finished, the simulation is done.

4.1.3 Dynamic infrastructure data

Our DES for railway simulation uses the following observable state:

- **Switches**: objects that fire events when they enter their left or right traversable state, and which can be called from the route process to start switching.
- **Detection sections**, objects that have an allocation and an occupancy status, which are observable through events.
- **Signals**: objects that have a movement authority length which can be observed by a train.

4.1.4 Processes

The core of the discrete event simulation technique can be implemented in a mainstream programming language as coroutine-like processes which can manipulate the state of the world at the current point in time or choose to wait for events which will be caused by other processes in the future. We used the Rust programming language, where coroutines are

only experimentally available, so we used an explicit state machine model to represent the progress of each process. The overall simulation process maintains a global clock, and when all processes are waiting for events, the global clock is advanced until the next scheduled event.

The main processes in railway simulation are:

- **Elementary route activation** waits for resources, allocates them, sets switches to given positions and starts the following sub-processes:
  - **Release trigger**: listens to a *trigger* detection section which is designated as the release trigger for a partial route. After the detection section has first been occupied, and later freed, resources are released for use in other elementary routes.
  - **Signal catcher**: sets the route entry signal to the 'proceed' aspect, then waits for a given trigger section to become occupied before setting the signal back to 'stop'.
  - **Overlap timeout**: releases some resources after a given timeout. The timeout is started on the allocation of a specific track section (the *trigger*).
- **Swing safety zone**: replace one active safety zone with another. First, wait to allocate the additional required resources. Then release the resources which are no longer required.
- **Train** evaluates movement authority using information from signals currently in sight, and takes one of the following actions: accelerate, brake, or coast/wait. Braking curves from velocity limitations are calculated, representing the train driver's plan for when to start braking. We calculate a guaranteed minimum time until further action is required from the driver by taking the minimum time until one of the following happen (see also Fig. 19):
  - train arrives at a new node
  - train reaches maximum velocity
  - train enters the area of a new velocity restriction
  - acceleration/coasting curve intersects the braking curve

  After this minimum time has passed, or any signals currently in sight have changed state, the train updates its position and velocity according to the chosen driver action and the laws of motion.

Note that since we assume a constant maximum acceleration and braking, the equations of motion can be solved analytically, and there is no need for discretizing the time or space domains, except for the re-evaluation of the equations of motion at discrete events. This ensures that the train starts braking in time using only the information available to the driver at any given time.

4.2 Extensions and alternative simulators

In our simluation model, trains re-calculate braking curves analytically on every possibly relevant event. This makes for a high-performance system, but in real-world engineering there are other complexities that we do not yet handle in this system, such as:

- More complex signaling and automated train protection systems.
- Local variations and details of infrastructure, such as the inner workings of components from different vendors performing various tasks like route allocation, de-allocation, safety zones, partial release, level crossings, etc.
- Train dynamics models using curve radius, gradient, air/tunnel resistance, weight distribution, etc.

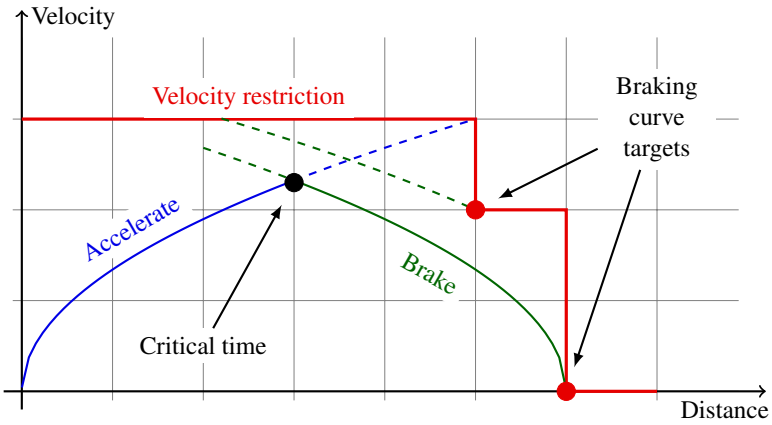

Fig. 19: The train driver's decision about when to accelerate/brake/coast happens at intersections between acceleration curves, braking curves and velocity restriction curves. In this example, the train can accelerate until the critical time where the acceleration intersects with the braking curve towards the second velocity restriction ahead (the first one is not critical).

– Stochastic variation in simulation output.

Our system can be extended with these features, or it would also be possible to swap out our simulation module with a more comprehensive solution or a commercially available offering (see [32,24]), as long as this simulation program can be run in batch mode using the range of input described above. Also, implementing a discrete event simulation is most elegantly done through co-routines, such as in the SimPy[3] Python library, or though specialized languages for simulation such as ABS[4]. However, for the simple simulation system we have implemented, the number of distinct states in each type of process is so low that it can easily be managed by explict state machine logic.

5 Case studies and performance

This section presents running times for different typical performance specifications on different types of railway infrastructure where the size and complexity of the model is typical for the scope of railway construction projects. Verification performance on various test examples as well as real stations is presented in Table 1. The table shows the time spent in each solver component, and also the number of invocations n_{DES} of the simulator, which is very low in most of the practical cases. This supports our hypothesis that the chosen abstraction and CEGAR loop is efficient. The two-track station used in Fig. 1 is not too complex, having only 6 elementary routes. Even so, this scale is still interesting for verification in practice, since there are many possible mistakes to uncover.

The Norwegian railway infrastructure manager Bane NOR has supplied a railML infrastructure model of the whole national railway network [35] from which we have extracted some more complex examples. Fig. 20 shows cut-outs from the visual representation of

[3] See `https://en.wikipedia.org/wiki/SimPy`
[4] See `http://abs-models.org/`

Infrastructure	**Property**	**Result**	n_{DES}	t_{SAT}	t_{DES}	t_{total}
Simple (3 elem.)	Run.time	Sat.	1	0.00	0.00	0.00
	Crossing	Unsat.	0	0.00	0.00	0.00
Two track (14 elem.)	Run.time	Sat.	1	0.01	0.00	0.01
	Frequency	Sat.	1	0.01	0.00	0.01
	Overtaking 2	Sat.	1	0.00	0.00	0.01
	Overtaking 3	Unsat.	0	0.01	0.00	0.01
	Crossing 3	Unsat.	0	0.01	0.00	0.01
Kolbotn (BN) (56 elem.)	Run. time	Sat.	2	0.01	0.00	0.02
	Overtake 4	Sat.	1	0.05	0.00	0.06
	Overtake 3	Unsat.	0	0.05	0.00	0.06
Eidsvoll (BN) (64 elem.)	Run. time	Sat.	2	0.01	0.00	0.02
	Overtake 2	Sat.	1	0.08	0.00	0.08
	Crossing 3	Sat.	1	0.04	0.00	0.04
	Crossing 4	Unsat.	0	0.21	0.00	0.21
Asker (BN) (170 elem.)	Overtaking 2	Sat.	1	0.20	0.00	0.21
	Overtaking 3	Unsat.	1	0.73	0.00	0.74
	Crossing 4	Sat.	0	0.75	0.00	0.77
Arna (CAD) (258 elem.)	Run. time	Sat.	1	0.02	0.00	0.04
	Overtaking 2	Sat.	1	0.50	0.00	0.51
	Overtaking 3	Sat.	1	1.43	0.00	1.45
	Crossing 4	Sat.	1	1.73	0.00	1.74
Gen. 3x3 (74 elem.)	High time	Sat.	1	0.01	0.00	0.01
	Low time	Unsat.	27	0.18	0.01	0.19
Gen. 4x4 (196 elem.)	High time	Sat.	1	0.01	0.00	0.03
	Low time	Unsat.	256	2.08	0.26	2.34
Gen. 5x5 (437 elem.)	High time	Sat.	1	0.06	0.00	0.09
	Low time	Unsat.	3125	38.89	4.35	43.24

Table 1: Verification performance on test cases, including Bane NOR (BN) and RailCOMPLETE (CAD) infrastructure models. The number of elementary routes (*elem.*) is shown for each infrastructure to indicate the model's size. n_{DES} is the number simulator runs, t_{SAT} the time in seconds spent in SAT solver, t_{DES} the time in seconds spent in DES, and t_{total} the total calculation time in seconds.

these models, i.e., the stations Kolbotn, Eidsvoll, and Asker were converted from the railML models.

We have also tested against an infrastructure model from the Arna construction project that uses the RailCOMPLETE CAD design software, a realistic use case for agile verification.

Finally, to test the limitations of scalability in our method, we construct a set of examples where m stations each with n parallel tracks each are serially connected by a single track. In this case, when a timing bound is slightly too small to be satisfiable, the planner will have to come up with n^m plans for timing evaluation. This scenario is outside the intended use case for our method: path selection can on this scale instead be based on static speed profiles. Capacity over many stations is better suited for the established timetabling tooling.

We attempted an alternative implementation using the PDDL+ solver SMTPlan+, but found that even for greatly simplified models, the required number of steps and numerical constraints put all our case studies out of reach for sub-second verification times.

6 Conclusions and Related Work

Railway timetabling and capacity analysis has often been posed as a planning problem and solved using mixed integer programming and similar approaches. Zwaneveld et al. [40] use integer programming on a problem closely related to our low-level railway infrastructure capacity verification problem. Isobe et al. [19] formulate a similar model in timed CSP, representing train locations, velocities, and control logic. Our definition of the problem in this paper includes non-linear constraints on train dynamics (acceleration/braking power) and communication constraints (trains must slow down if they have not been informed of movement authority), which are relevant in construction projects but less relevant in timetabling.

Many variations on discrete event simulation are used in railway dynamic analysis. A comprehensive account of object-oriented modeling and simulation of railway infrastructure is given in D. Hürlimann's Ph.D. thesis [18] (also based on M. Montigel's thesis [29]), which was later developed into the commerical simulation software OpenTrack. A similar

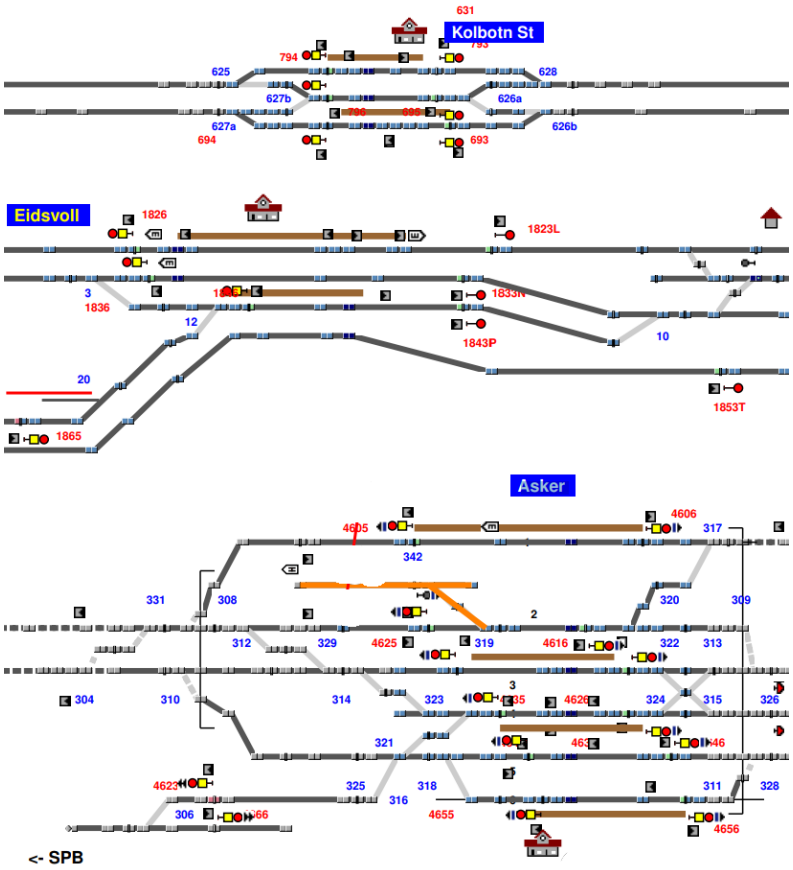

Fig. 20: Stations Kolbotn, Eidsvoll, and Asker from Bane NOR's model of the Norwegian national network [35].

approach presented in [21] uses futures and resource analysis support in the ABS programming language to simulate operational procedures.

In the planning literature, the PDDL+ language [11] has been introduced to capture mixed discrete/continuous planning problems such as the one studied in this paper. General-purpose solvers have recently been developed, using time domain discretization (DiNo [34]) or the SMT theory of non-linear real arithmetic (SMTPlan+ [5]).

The goal of our suggested tool chain for railway engineering is (1) to allow fully automated performance verification and (2) use minimal input documentation for the verification. Both of these aspects encourage bringing performance verification into the frequently changing early-stage design projects, avoiding the costly and time-consuming backtracking required when later-stage analysis reveals unacceptable performance.

The control system design phase is lacking tools for rapid prototyping by anticipating the verification which is to be performed in later stages.

In this paper, we have demonstrated a control system design tool that can verify performance properties in the scope of a single project from high-level specification by synthesizing schedules. Our work thus automates the following activities:

- Detailed **running time** analysis – verify the time required for getting from point A to point B, taking into account train dynamic characteristics, communication constraints, and control system logic and latency.
- Detailed **schedulability** analysis – verify frequency of trains arriving at a station, and simultaneous opportunities for crossing, parking, loading, etc.

Our approach carves a new niche in railway design automation in the following sense: the level of detail supported by this tool is much greater than the traditional by-hand approaches for running time and schedulability analysis – and the amount of background data and work is much less than the whole-network stochastic operational analysis typically used for later-stage verification. To make the method approachable for engineers, the required input is the minimum of information needed to verify the relevant properties. For example, the specific paths each train takes through the station is not an input, but different possibilities for realizing paths are explored by the verification procedure. This also makes the method more appropriate for early-stage design, where track lengths, topology, and component placement might be adjusted to achieve design goals, and engineers can in this way get feedback on design choices without requiring large efforts to repeat the verification.

Acknowledgments

We thank the engineers at Railcomplete AS, especially senior engineer Claus Feyling, for guidance on railway operations and design methodology. We plan to integrate the current prototype in their RailCOMPLETE tool and test the usability with the engineers using this tool in their design work.

References

1. Montserrat Abril, Federico Barber, Laura Paola Ingolotti, Miguel A. Salido, Pilar Tormos, and Antonio Luis Lova. An assessment of railway capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(5):774 – 806, 2008.
2. Clark Barrett and Cesare Tinelli. Satisfiability modulo theories. In *Handbook of Model Checking*, pages 305–343. Springer, 2018.

3. Armin Biere, Alessandro Cimatti, Edmund M Clarke, Ofer Strichman, Yunshan Zhu, et al. Bounded model checking. *Advances in computers*, 58(11):117–148, 2003.
4. Arne Borälv and Gunnar Stålmarck. Formal verification in railways. In Michael G. Hinchey and Jonathan P. Bowen, editors, *Industrial-Strength Formal Methods in Practice*, pages 329–350. Springer, 1999.
5. Michael Cashmore, Maria Fox, Derek Long, and Daniele Magazzeni. A compilation of the full PDDL+ language into SMT. In Amanda Jane Coles, Andrew Coles, Stefan Edelkamp, Daniele Magazzeni, and Scott Sanner, editors, *International Conference on Automated Planning and Scheduling, ICAPS 2016*, pages 79–87. AAAI Press, 2016.
6. Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. Counterexample-guided abstraction refinement. In E. A. Emerson and A. P. Sistla, editors, *Proceedings of the 12th International Conference on Computer-Aided Verification (CAV '00)*, volume 1855 of *Lecture Notes in Computer Science*, pages 154–169. Springer Verlag, 2000.
7. Leonardo De Moura and Nikolaj Bjørner. Satisfiability modulo theories: introduction and applications. *Communications of the ACM*, 54(9):69–77, 2011.
8. Leonardo Mendonça de Moura and Nikolaj Bjørner. Z3: an efficient SMT solver. In C. R. Ramakrishnan and Jakob Rehof, editors, *Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2008)*, volume 4963 of *Lecture Notes in Computer Science*, pages 337–340. Springer, 2008.
9. Bruno Dutertre. Yices 2.2. In Armin Biere and Roderick Bloem, editors, *26th Computer Aided Verification (CAV)*, volume 8559 of *Lecture Notes in Computer Science*, pages 737–744. Springer, 2014.
10. Niklas Eén and Niklas Sörensson. Temporal induction by incremental SAT solving. *Electronic Notes in Theoretical Computer Science*, 89(4):543–560, 2003.
11. Maria Fox and Derek Long. Modelling mixed discrete-continuous domains for planning. *Journal of Artificial Intelligence Research*, 27:235–297, 2006.
12. Martin Franzle, Christian Herde, Tino Teige, Stefan Ratschan, and Tobias Schubert. Efficient solving of large non-linear arithmetic constraint systems with complex boolean structure. *Journal on Satisfiability, Boolean Modeling and Computation*, 1:209–236, 2007.
13. Sicun Gao, Soonho Kong, and Edmund M. Clarke. dReal: An SMT solver for nonlinear theories over the reals. In Maria Paola Bonacina, editor, *24th International Conference on Automated Deduction (CADE)*, volume 7898 of *Lecture Notes in Computer Science*, pages 208–214. Springer, 2013.
14. Martin Gebser, Tomi Janhunen, and Jussi Rintanen. SAT modulo graphs: Acyclicity. In Eduardo Fermé and João Leite, editors, *14th European Conference on Logics in Artificial Intelligence (JELIA)*, volume 8761 of *Lecture Notes in Computer Science*, pages 137–151. Springer, 2014.
15. Ingo Arne Hansen and Jörn Pachl. *Railway Timetabling and Operations*. Eurailpress, 2014.
16. Steven S. Harrod. A tutorial on fundamental model structures for railway timetable optimization. *Surveys in Operations Research and Management Science*, 17(2):85 – 96, 2012.
17. Anne E. Haxthausen and Peter H. Østergaard. On the Use of Static Checking in the Verification of Interlocking Systems. In Tiziana Margaria and Bernhard Steffen, editors, *Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination, Applications*, pages 266–278. Springer, 2016.
18. Daniel Hürlimann. *Objektorientierte Modellierung von Infrastrukturelementen und Betriebsvorgängen im Eisenbahnwesen*. PhD thesis, ETH Zurich, 2002.
19. Yoshinao Isobe, Faron Moller, Hoang Nga Nguyen, and Markus Roggenbach. Safety and Line Capacity in Railways – An Approach in Timed CSP. In John Derrick, Stefania Gnesi, Diego Latella, and Helen Treharne, editors, *9th International Conference on Integrated Formal Methods (iFM)*, Lecture Notes in Computer Science, pages 54–68. Springer, 2012.
20. Dejan Jovanovic and Leonardo de Moura. Solving non-linear arithmetic. *ACM Comm. Computer Algebra*, 46(3/4):104–105, 2012.
21. Eduard Kamburjan and Reiner Hähnle. Uniform modeling of railway operations. In *Formal Techniques for Safety-Critical Systems FTSCS 2016*, volume 694 of *Communications in Computer and Information Science*, pages 55–71. Springer, 2016.
22. Alex Landex. *Methods to estimate railway capacity and passenger delays*. PhD thesis, 2008.
23. Fangzhen Lin and Yuting Zhao. ASSAT: computing answer sets of a logic program by SAT solvers. *Artificial Intelligence*, 157(1-2):115–137, 2004.
24. LUKS: Analysis of lines and junctions, 2018.
25. Bjørnar Luteberget, John J. Camilleri, Christian Johansen, and Gerardo Schneider. Participatory Verification of Railway Infrastructure by Representing Regulations in RailCNL. In Alessandro Cimatti and Marjan Sirjani, editors, *International Conference on Software Engineering and Formal Methods (SEFM)*, volume 10469 of *Lecture Notes in Computer Science*, pages 87–103. Springer International Publishing, 2017.

26. Bjørnar Luteberget, Koen Claessen, and Christian Johansen. Design-time railway capacity verification using SAT modulo discrete event simulation. In Nikolaj Bjørner and Arie Gurfinkel, editors, *Formal Methods in Computer Aided Design (FMCAD)*, pages 1–9. IEEE, 2018.
27. Bjørnar Luteberget and Christian Johansen. Efficient verification of railway infrastructure designs against standard regulations. *Formal Methods in System Design*, 52(1):1–32, Feb 2018.
28. Bjørnar Luteberget, Christian Johansen, and Martin Steffen. Rule-based consistency checking of railway infrastructure designs. In Erika Ábrahám and Marieke Huisman, editors, *Integrated Formal Methods 2016*, volume 9681 of *Lecture Notes in Computer Science*, pages 491–507. Springer, 2016.
29. Markus Montigel. *Modellierung und Gewährleistung von Abhängigkeiten in Eisenbahnsicherungsanlagen*. PhD thesis, ETH Zurich, 1994.
30. Andrew Nash, Daniel Huerlimann, Jörg Schütte, and Vasco Paul Krauss. RailML — a standard data interface for railroad applications. In *Computers in Railways IX*, pages 233–240. WIT Press, 2004.
31. Robert Nieuwenhuis, Albert Oliveras, and Cesare Tinelli. Solving SAT and SAT modulo theories: From an abstract Davis–Putnam–Logemann–Loveland procedure to DPLL(T). *Journal of the ACM*, 53(6):937–977, 2006.
32. OpenTrack: Simulation of railway networks, 2018.
33. Jörn Pachl. *Railway Operation and Control*. VTD Rail Publishing, 2015.
34. Wiktor Mateusz Piotrowski, Maria Fox, Derek Long, Daniele Magazzeni, and Fabio Mercorio. Heuristic planning for PDDL+ domains. In Subbarao Kambhampati, editor, *International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 3213–3219. IJCAI/AAAI Press, 2016.
35. Bane NOR: Model of the Norwegian rail network, 2016.
36. railML. The XML interface for railway applications, 2018.
37. Stewart Robinson. *Simulation: The Practice of Model Development and Use*. John Wiley & Sons, Inc., USA, 2004.
38. Carsten Sinz. Towards an optimal CNF encoding of boolean cardinality constraints. In Peter van Beek, editor, *Principles and Practice of Constraint Programming*, volume 3709 of *Lecture Notes in Computer Science*, pages 827–831. Springer, 2005.
39. Linh Hong Vu, Anne Elisabeth Haxthausen, and Jan Peleska. A domain-specific language for railway interlocking systems. In *10th Symposium on Formal Methods for Automation and Safety in Railway and Automotive Systems*, pages 200–209. TU Braunschweig, 2014.
40. Peter J. Zwaneveld, Leo G. Kroon, and Stan P.M. van Hoesel. Routing trains through a railway station based on a node packing model. *European Journal of Operational Research*, 128(1):14 – 33, 2001.