Ferromone Trails Concept

Using ant-colony algorithms to increase rail capacity on transitional infrastructure

Transport Technology Research Innovation Grant (DfT T-TRIG OC068)
Final Report, June 2017

Anthony J H Simons and Sina Shamshiri

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield S1 4DP, UK.

Executive Summary

The UK has an ageing rail network with old fashioned signalling systems, which is restricting the growth of rail capacity. The most cost-effective way to increase capacity, without extensive junction alteration, is by altering the rules of train movement. The UK is partway through adopting an ambitious European standard for centralized radio control of all trains, which will replace traditional signals and allow a higher throughput of rail traffic; but it is unclear if the UK will be able to afford the complete system.

This project has prototyped a multi-agent simulation, which explores how to deliver the same degree of flexibility more cheaply, on transitional infrastructure. The simulation allows trains to self-organise and dynamically re-plan their headway, speed, and reactions under adverse circumstances (line gradients, wet weather, ice or leaves on the line), using a system of decentralized control based on ant-colony algorithms. Similar to ant pheromone trails, trains deposit and sense signals left by other trains (dubbed "ferromone" trails), using two layers of control that exploit the existing Eurobalise transponder beacons and the Global System for Mobile Communications – Railway (GSM-R) cellular radio network.

The simulator allows the construction of railway landscapes for single- and dual-line traffic, with intermediary stations and junctions controlled by signals. Routes are travelled end-to-end by an increasingly compressed schedule of trains. A railway landscape is evaluated first using traditional fixed-block signalling, to establish the maximum capacity under current rules of operation. Then, the same landscape is evaluated using trains equipped to read/write ferromone trails, which self-organise according to dynamic-block control. Speed and headway planning is greatly improved on open stretches and on approach to stations, where closely-spaced trains decelerate as a cohort. With bi-directional ferromone trail sensing, only slight further gains in planning are made, unless this mode used with heavy traffic at 1–2 minute departure cadences.

At a departure cadence of 3 minutes, capacity can be increased by five-fold on High Speed lines, and by three-fold on Inter-City lines with intermediate stations, compared to 10 km fixed blocks. Routes can also be run at over-capacity (with 1–2 minute departure cadences), subject to trains running at less than their best line speeds. Departure cadences of 3–5 minutes give rise to stable traffic throughput, even when lines are fully occupied. Trains behave safely, unless advisory line-speeds are sharply discontinuous and trains are closely spaced, where a queue of braking trains will eventually trigger emergency braking. The fail-safe emergency radio relay causes following trains to brake early and avoid collision. Ferromone trail balises should be spaced at intervals not exceeding 100 m. Ferromone trails are most needed where lines run with mixed-class (two-speed) traffic, which requires more than static advisory line speeds for safe control.

Table of Contents

1. Introduction

According to current estimates, the capacity of the UK national railway network must at least double over the next thirty years to 2040 [1]. Different approaches to increasing capacity and reducing congestion include operational solutions (altering the rules for train movement), engineering solutions (re-configuring the track at junctions and stations) and technological solutions (introducing in-cab radio signalling with central traffic management). The likely effect of these combined measures is not well-understood. Solving one local capacity-related problem may have unpredictable global effects: bigger trains may increase station dwell-times and cause cascading effects in the timetable; or removing a pinch-point at one junction may only succeed in congesting another, possibly distant, junction. A better global solution might even require counter-intuitive measures, such as slowing trains, removing services or disaggregating line usage (separating freight and passenger traffic) to increase throughput. The solutions to these chaotic problems are likely to be dynamic, self-organising and not predictable.

The UK is partway through adopting the European Rail Traffic Management System (ERTMS) [2] using trackside balises (RFID-accessed data beacons) interrogated by trains as they pass over. ERTMS-1 focused on uploading line speed and signal profiles, but kept traditional *fixed block* signals. ERTMS-2 is adding *dynamic block* radio signalling from the lineside to trains, optimizing blocks for train speeds and removing the need for lineside signals. ERTMS-3 will adopt a centralized European Traffic Management Layer (ETML) with centralized radio control of all train movement. However, it is unclear whether the UK will finally adopt ERTMS-3: the ETML technology has not yet been developed; the data capacity of the GSM-R radio network (Global System for Mobile Communications – Railway, a 2G mobile network) has not been tested under the full centralized control of all trains; and the cost/benefit returns are uncertain. A careful 2010 study by the Transport Research Laboratory [3] concluded that, while ERTMS-2 could deliver operational cost reductions of 40%, ERTMS-3 would only deliver a further 25% reduction, for considerable expense of implementation.

The challenge is to increase rail capacity, through smart operational changes, using the existing transitional infrastructure, while the uncertain projected longer-term technological solutions are still being phased in. We think it may be possible to achieve the benefits of fully-autonomous self-organising control of trains using only small modifications to ERTMS-2 technology.

2. Ferromone Trails Concept

The novel idea explored in this project is to simulate the decentralized control of trains, so that they may select their best headways, line speeds and respond to junction signals, rail accidents and emergencies using only local information transmitted to and from the trackside. The simulation uses balises in a new configuration to monitor dynamic speeds and headway, and uses GSM-R trackside equipment for failsafe operation, junction control and signalling. The idea is modelled after the behaviour of eusocial insects, in particular ant-colonies, which exercise no global top-down control, but exhibit optimal self-organising behaviour using only local environmental constraints. We have previously modelled ant-colony foraging behaviour at the level of individual ants, as they lay down pheromone trails to encourage (or inhibit) other ants on the same trails, and have replicated the robust behaviour found in nature, honed by millions of years of evolution, whereby ants always recover the shortest path to food sources, when these are unreliable, disrupted, or change location [4].

2.1 Traditional Fixed-Block Signalling

The current rail network uses *fixed block* signalling, illustrated in figure 1. Once released by the start signal, the train may enter the block. If the line ahead of the home signal is not clear, the distant signal will order the train to halt. The braking distance (and safety margin) from the distant signal is calculated such that the

moving train may halt at the home signal. After adding the signal sighting distance, these lengths determine the block size for each controlled section. Blocks are calculated according to worst-case braking distances and best visible locations for signal placement, which means that some blocks are much larger than needed for safe braking; line occupancy is therefore lower.

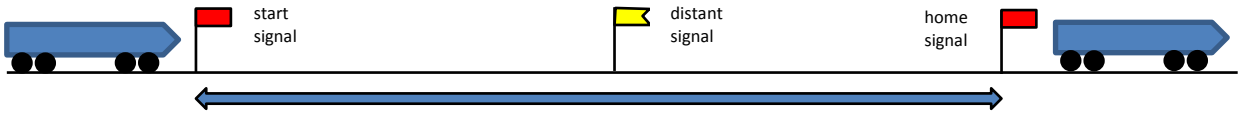

Figure 1: Absolute block (fixed block) signalling

2.2 ERTMS-3 Centralised Control System

Figure 2 illustrates the ERTMS-3 notion of *dynamic block* signalling, with all trains in continuous radio communication with a regional control centre. The movement authority is attached to the train, rather than to a section of track. The block-size must exceed the braking distance (green) and safety margin (yellow), which in turn are determined according to the relative speeds of the two trains and the underlying track and weather conditions. The movement authority of a train ends when it enters the dynamic block allocated to the preceding train. Since dynamic blocks are continuously optimised for train speeds, they support a much higher line capacity than fixed-block approaches. However, under ERTMS-3, this all depends heavily on position and speed information transmitted continuously by each train to the regional control centre, and on revised headways transmitted back by the control centre to the trains.

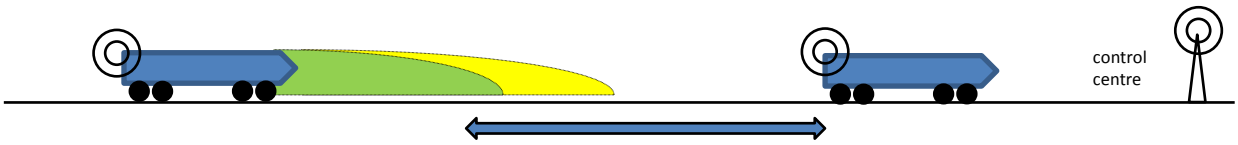

Figure 2: Variable block (dynamic block) signalling

2.3 Ferromone Trails: Self-Organisation

Our proposed self-organising strategy provides variable-block capability for an infrastructure cost similar to ERTMS-2 and with less dependence on high-volume GSM-R radio traffic. Figure 3 depicts the laying and sensing of "*ferromone trails*". Trains equipped with a Balise Transmission Module perform data uplink (front of train) and data downlink (rear of train), reading and writing to regularly-spaced balises. In single-line operation, the preceding train deposits its ID, time and speed of passing the local balise checkpoint. The following train reads the most-recently deposited ID, time, speed and balise position and continuously recalculates its safe braking curve from this and its own data. In figure 3, the train on the left is reading time and speed information that was deposited by the preceding train on the right when it passed that balise. From the timestamp and speed, it estimates the current position and speed of the preceding train (with some error-margin, depending on balise spacing), and from this calculates its own braking curve.

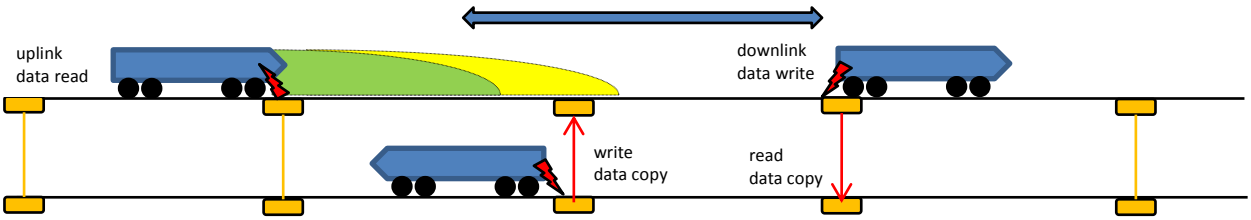

Figure 3: "Ferromone trail" single and dual line operation

In dual line operation, balises are linked in pairs across the up- and down-lines such that data may be read and written about traffic in both directions (viz. each balise has induction-loops for accessing the up- and down-line data). Trains passing in one direction also copy the data for the opposite line, updating balises on the remote line with more recent information about the train that passed earlier on that line. In figure 3, the train on the left is reading older time and speed information that was deposited by the preceding train; however the train heading in the opposite direction has just updated the next balise ahead of the first train with more recent time, speed and position information about the train on the right, which it originally read from a balise just after it passed that train on the other line. This is the most up-to-date information about that train, and it will be copied to all remote balises, until another train is passed on the other line. This dual-line configuration has the benefit that, as the lines become more fully utilised, speed and position estimates get more accurate, supporting better headway planning (with smaller error-margins).

2.4 Ferromone Trails: Emergency Operation

In a similar manner to ant pheromone trails, in which different chemicals are overlaid, our control system consists of two layers, including one that utilises GSM-R as a packet-switching network to support local movement authority and emergency operations. The most important concern is fail-safe operation in an emergency. Whereas the balise-based control system can reliably regulate speeds and headways for normal operations up to normal service braking, in an emergency, there will always be cases that exceed the capability of this system. Figure 4 illustrates a scenario in which the preceding train has become derailed or has had to apply maximum braking force. The following train may not detect the rapid deceleration of the first train, until it reads a balise (shown in red) inside its own braking safety margin.

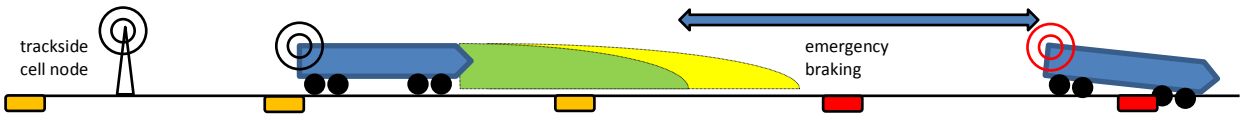

Figure 4: "Ferromone trail" fail-safe emergency operation

This is mitigated by the parallel control system, in which trains constantly emit a radio "squawk" to indicate nominal status and position at regular intervals (e.g. every second). In an emergency, such as applying full service braking, or maximum emergency braking, or becoming derailed, this will change to an emergency "squawk" giving the position of the train. The following train receives signals from the preceding train via the GSM-R cellular network (a handshake protocol allows signals to propagate only as far as needed). If this changes to an emergency "squawk", *or if the signal is lost due to an accident*, the following train will immediately recalculate its safe braking curve, based on the worst-case estimate of the previous train's speed and position. In figure 4, the following train may receive accurate position information (by emergency squawk); *or it may lose the signal* and have to recalculate a safe braking curve, based on the last-known position of the preceding train, assuming it applied emergency braking after that. The latter protocol guarantees fail-safe operation in the case of equipment failure.

2.5 Aims and Objectives of the Current Project

While the ultimate aim of this research is to explore how self-organising trains may in future be able to re-plan dynamically all aspects of their behaviour (speed, headway, emergency response and route-selection, where alternative routes exist), the current proof-of-concept project has focused firstly on developing and refining the model-concepts needed to simulate the *ferromone trail* style of control; and secondly on seeking to demonstrate that this kind of control system may be operated safely.

The project was executed in four phases:

- Model Construction: building the landscape of railway routes, junctions, stations and signals; building the elements of the control infrastructure and controllers; and populating the routes with trains.
- Single/Dual Line Evaluation: simulating the movement of trains under traditional fixed block control; under single-line *ferromone trails* control; and under dual-line *ferromone trails* control.
- Emergency Operation: simulating the emergency response of trains to a derailed train, firstly using the balise-based control system; and secondly, using the GSM-R based emergency response system.
- System Degradation: simulating the system under poor weather and line conditions; and simulating the system with a percentage of damaged or destroyed control infrastructure elements.

The project was formally announced on 14 February, 2017. The researcher, Dr Sina Shamshiri, was employed from 20 February, 2017. The project is due to conclude on 15 June, 2017.

2.6 Achieved Scope and Future Extensions

In the four month period, we have built a simulator that allows the construction of railway routes for single- and dual-line traffic, with intermediary stations and junctions controlled by signals. Routes are travelled by trains, under different systems of fixed-block and dynamic-block control, in order to measure the effects on line capacity. We have focused on the problem of optimising the speed and headway of a series of trains proceeding along a route (where the route and its pinch-points may be varied arbitrarily). We have deliberately excluded the problem of trains joining and leaving at point-switches (turnouts) and assume that trains are always safely routed inside stations. Junctions are modelled only to the extent that they constitute obstacles to the passage of trains, controlled by signals. We have, for now, excluded the interesting problem of dynamic route re-planning, which will require branching and joining rail routes. We hope to return to these problems at a later date.

The focus of this prototype is to demonstrate that self-organising control can be operated safely. In previous discussions, Network Rail had judged the self-organising approach to be radical and high-risk. By providing this successful demonstration, we hope to convince Network Rail and other partners to support us in following up this work with a much larger project, which will simulate the UK rail network at scale, using the FLAME massive multi-agent simulator developed at Sheffield [5, 6]. This has been applied successfully in the past to model massive-scale multi-agent behaviour in a variety of domains, including skin-cell regrowth [7], chemical interactions [8], ant-colony behaviour [4, 9], European macroeconomic policy [10] and pedestrian crowd behaviour [11]. FLAME is quite different from the regular kinds of Monte-Carlo simulators (such as VISION, SPA, CUI, TRAIL and RailSys [12, 13]) used by Network Rail, in that it models the detailed behaviour of individual agents exactly – it will reveal chaotic behaviour, rather than smooth it out.

3. Simulator Implementation

This section describes the implementation of the prototype simulator, including:

- the development of model-concepts for the simulator,
- the backing data from physics and railway research used to inform the simulator,
- the software implementation technology chosen, and
- the railway landscapes constructed for simulation purposes.

From the outset, it was determined that the simulator must keep a clear separation of concerns between its models of the real world (material reality, which we shall call the *materium*[1]) and error-prone perceptions of this world constructed by agents that sense aspects of the real world (constructed reality, which we shall call the *simulacrum*). We define these two terms in order to avoid using terms like "real-world" ambiguously to denote the materium, since the materium is itself only an approximation to reality. The separation of concerns between these two layers of the simulator allows it to check whether trains operating on internal rules within the simulacrum have violated physical rules for safe movement in the materium.

3.1 The Materium

The *materium* is anything that forms part of the simulator's model of the real world. It is a discrete sampling of reality. It obeys real-world physics at some quantized level (typically updated every 100ms; but this is a parameter in our simulator). The materium must be accurate enough to provide a reality-check for the simulacrum (for example: a train may believe it is at a certain point on a line, but its estimate may be 1-3% incorrect with respect to the materium). The main concepts of the materium are listed in table 1.

Table 1: Concepts in the Materium

Concept	Description
Landscape	Base terrain containing railway route, stations, junctions
Timer	Global scheduler controlling quantization of the simulation
Oracle	Global controller comparing simulacrum/materium, reporting violations
Track	Track sections (blocks) with length, incline, curvature, line-condition
Station	Controlled stopping-points with multiple platforms, dwell-time and signals
Signals	Home/distant signal pair, linked to a lineside radio control block and balise
Trains	Generic trains with parametric characteristics (simulating HST, ICX, LST traffic)
Balise	Track-bed transponder beacon capable of recording data from passing trains
Radio Mast	Packet-switching relay in the cellular GSM-R radio network

Key general features of the materium are:

- The Landscape may be constructed from a simple text-file in the YAML format, which allows rapid building of different kinds of simulation, with different railway features and types of trains;
- The track sections may correspond to fixed blocks when used in traditional operation, or are joined as continuous track in dynamic block operation, by adding or omitting signal controls on entry;
- The trains are parameterised by length (number of cars) and maximum speed to simulate High Speed Trains (HST), Inter-City Express (ICX), or Local Stopping Trains (LST) traffic;
- Stations are assumed to hold a number of trains, which are safely routed to different platforms once entry has been gained, and trains are released in departure-order through signals;
- Signals consist of a home signal, and one distant signal (or many repeater-signals), controlled by a lineside electronic unit (LEU) and radio control block (RCB) linked to the GSM-R network;
- Junctions are finessed in this prototype as portions of the main route controlled by signals (viz. representing when the route ahead is blocked, due to a switched junction aspect);
- Trackside GSM-R radio masts are positioned at regular intervals, assumed sufficient to ensure full cellular connectivity and so are not varied during simulations;
- Balises are positioned at regular intervals on the track-bed; and their frequency may be adjusted to give denser or sparser ferromone trails.

[1] The term *materium* was apparently first coined in the role-playing game "Warhammer 40,000" to distinguish real space-time, in which the conventional laws of physics apply, from the hyper-space dimension.

Of specific interest are the infrastructure features supporting self-organising control. The balises come in up to four different types[2]. The first three have a single induction coil, the last has a double induction coil, supporting reading and writing of data on both the up and down lines.

- Passive balise – stores information about its location, track gradient, and advisory line speed;
- Active balise – stores signal aspect information, energised by a lineside electronic unit (LEU);
- Ferromone balise (single) – stores the timestamp, position, speed and ID of the last train to pass over;
- Ferromone balise (dual) – is linked to its partner on the opposite track, with an extra induction coil.

The GSM-R radio masts are assumed to support a packet-switching protocol; however any protocol that supports train-to-train communication in short bursts will suit. Radio transmissions are in the form of short "squawks" that are routed back to following trains, which accept and acknowledge packets via a handshake protocol. GSM-R performs the following functions:

- Authority to proceed is given by the RCB linked to the signal ahead of a (halted) train;
- A continuous live signal is emitted by trains running on the line, with ID and position;
- A safe signal is emitted by trains that have stopped safely at a station platform;
- An emergency signal is emitted by trains in any kind of emergency;
- If the signal is lost for more than a defined period, this is treated as an emergency.

The Oracle continuously checks the correspondence between the materium and the simulacrum and reports when any train has violated rules of safe operation, such as exceeding a safe line speed or colliding with another train.

3.2 The Simulacrum

The *simulacrum* is anything that forms part of a train's internal view of reality, which is constructed using whatever sensing and communications equipment the train carries. The most important fact about the simulacrum is that it is only an approximation of the materium. This is because communications are intermittent (a train receives updates at discrete intervals, rather than continuously), empirical sensors are possibly unreliable (odometer and Doppler readings have an error factor), measured distances may vary according to the weather and line condition and equipment may possibly fail. The main concepts of the simulacrum are listed in table 2.

Table 2: Concepts in the Simulacrum

Concept	Description
Clock	The train's internal notion of time, synchronized daily with the global Timer
Driver	Human controller responding to fixed-block signals and lineside information
Sentinel	Automated controller responding to balise data and GSM-R radio signals
Disruptor	Agent applying the effects of sensor errors, or poor line and weather conditions
Balise Reader	Balise transmission module, capable of performing data uplink (read) function
Balise Writer	Balise transmission module, capable of performing data downlink (write) function
GSM-R Radio	Train-based GSM-R transmitter and receiver, with packet-switching function
Movement Sensor	Abstraction over axle transducers, accelerometers, radar units to detect motion
Occupancy Sensor	Abstraction over track circuits and axle-counters used to detect track entry/exit

[2] To simplify the simulator implementation, signals and their aspect-reporting balises are combined, such that placing a signal automatically assumes a balise further up the track within "sighting distance" of the signal. Eventually, only "read-only" balises, and "read-write" balises, are modelled explicitly.

Key general features of the simulacrum are:

- The Driver controller behaves like a human train driver, responding to signals within sighting distance and reading trackside line speed indicators (balises are used to represent trackside indicators);
- The Sentinel controller automatically optimises the train's headway and speed, using balise data for line speeds, signal aspects and preceding train information; and monitors GSM-R for emergencies;
- The Movement Sensor calculates a train's estimated position, based on the last balise giving exact position, and an internal estimate of distance travelled since (with an error-margin);
- The Occupancy Sensor reports when a controlled track block is entered or exited (we assume fixed configuration trains, so axle-counting errors are less significant);
- Train integrity (no accidental detachment of cars) is finessed through the trailing placement of the Balise Writer, which if detached, will slow to a halt, recording car deceleration (see also below);
- Trains travelling at up to the maximum advisory line speeds and applying no more than *full service braking* may function safely under the control of the balise-based system;
- Trains needing to apply *emergency braking*, or becoming derailed, or detecting a loss of train integrity (detached cars) must emit an emergency signal.

For the sake of determining the likely accuracy of any self-organising control system, it was important to be able to model realistic kinds of error that might be introduced in the simulacrum. Sources were researched to determine the likely accuracy of the technologies used. Kinds of error that we considered are:

- The possibility that the train's internal Clock becomes de-synchronised with the global Timer; to mitigate this, we assume that train clocks are re-synchronised every day to global time;
- The possibility that the train's Movement Sensor becomes worn or inaccurate (due to wheel-wear and less reliable axle transducers – we assumed inaccuracies up to 3% (where ERTMS asks for just 1%) ;
- The possibility that a track Occupancy Sensor may miscount the number of axles on a train (due to low-hanging mass); we assumed measured inaccuracies of up to two axles per train;
- The possibility that track inclines or declines (gradients up to 4%) will affect planned acceleration or deceleration, such that the train will experience slower or increased rates-of-change.

For the sake of determining the likely degraded behaviour of the system under adverse conditions, we also considered that the following may possibly happen:

- Signal-passed-at-danger (SPAD) events are currently not modelled explicitly (this would result in overshooting to the next block, and would trigger proximity violations), but are assumed to be mitigated by repeater-balise reading ahead of the home signal;
- The Balise Reader/Writer may fail if the balise is damaged or destroyed; Siemens states that the protocol ensures that data is either successfully transferred, or no data is changed [20-22];
- The GSM-R cellular network may suffer a temporary drop-out – so we added two seconds hysteresis to the monitoring of live signals, to prevent premature emergency responses;
- The line may be wet, or icy, or covered with leaf-slime ("leaves on the line") – so we modelled the changed physics, such that loss of traction may lead to proximity violations.

The simulator finesses the implementation of track inclines, declines and loss of adhesion by adding acceleration and braking factors to the affected sections of track. One feature omitted in the prototype, which would be needed in any extension to the whole UK rail network, is the concept of a Route Map, the train's internal model of the alternative routes it may possibly take, in the case of obstruction on the normal route.

3.3 Support from Railway Research

The business case for this project was informed by the 2007 DfT Sustainable Railway White Paper [1] and other rail strategy documents [14, 15]. General information about the ERTMS project was available from the ERTMS website [2] and from the published 1997 specification [16], with the different levels of ERTMS/ETCS operation summarised here [17]. The initial experimental deployment of ERTMS-2 in the UK from 2011 onwards was reported by UNIFE [18]. A useful critique of the likely long-term cost savings of ERTMS-2 and ERTMS-3 in the UK was prepared by the Transport Research Laboratory in 2010 [3].

We approached Siemens UK to establish the feasibility of read-write balises, which is fundamental to our *ferromone trails* concept. The original specification for the Eurobalise [19, 20] required both a data uplink (reading) and a data downlink (writing) function. A later version of the specification [21] and the Siemens production documents for the *Trainguard Eurobalise* S21 and S22 [22] removed the data downlink function. We contacted Mark Glover and Ewan Spencer at Siemens Rail Automation Holdings Ltd., UK to comment on this, in case this function had been removed due to a later-discovered technical infeasibility [23]. Fortunately, there was no technical issue, but the downlink function had simply no longer been needed in the revised ERTMS 1-3 specifications. According to Siemens [22], the balise transmission module can read a balise data telegram (with at least 3 successful scans) on trains travelling at up to 500 km/h (in excess of the anticipated 300 km/h speeds of HST-class trains). The size of the data telegram is 1023 bits (830 bits of useful data after parity checking), but larger telegrams could be transmitted if trains travel at slower speeds [19]. Our simulator assumes the existing telegram-size restriction, and introduces separate balises to perform the *ferromone trails* function. The distance from a balise to a Lineside Electronic Unit (LEU) may be a maximum of 5000 m [22], so cross-linking pairs of balises over much shorter distances across the track is also feasible.

The current and future plans for the GSM-R radio network were informed by Network Rail guides [24, 25]. At present, all of the UK is covered by 2,500 GSM-R radio masts, with over 15,000 km of railway lines covered (the whole UK, apart from some remote Scottish areas); and 4,056 trains are connected. The system was originally designed for analogue driver-to-signaller radio communications, although it is recognised that this has capacity limitations; and ERTMS requires a move away from analogue to digital [24]. The system is "being upgraded to support data mode (circuit switched) so that it can be used for the ETCS data communications between trackside and trains. As the capacity of a GSM-R cell is very constrained when using circuit switched, it is expected to be upgraded to packet switched (GPRS) which improves the effective capacity" [23].

Our simulator assumes packet-switched data transmissions, where the track-side masts function like routers in a packet-switched network; although any protocol that supports train-to-train communication in short bursts would suffice. We assume that Lineside Electronic Units (LEUs) operate motorised points with their interlocking and signals, and these are connected to a local Radio Block Controller (RBC), which issues movement authorities to trains via GSM-R. Whereas a train may read signal-aspect information from balises as it approaches a signal, if the train eventually halts at a home signal, it must wait for a movement authority to be transmitted via GSM-R (if all visual signalling has been removed), or else if the visual signal aspect changes (if visual signals are retained). Our emergency system also uses GSM-R in a fail-safe mode (see above).

We investigated typical train configurations and lengths from widely-available data shown in table 3. Our simulator currently uses a standardised car-length of 20 m and supports trains of different lengths (most simulations were conducted with 16-car HST and 5-car ICX trains). We were also keen to ensure that the performance of trains in our simulator was informed by physics and railway operating standards. In this, we were helped by the work of Piers Connor [26, 27], who has provided an analysis of future rail capacity for HST-class trains (like the Eurostar, TGV or ICE trains). These documents give realistic acceleration and braking figures for HST-class trains. We supplemented this with figures from Victor Winter's analysis of suburban

services in the Bay Area Rapid Transit (BART) system [28]. From this, we established the ball-park data for acceleration and braking given in table 4.

Table 3: Train parameters informing the simulation

Parameter	Description
300 km/h	Maximum operating speed of HST-class trains (ICE, TGV, Eurostar).
225 km/h	Maximum operating speed of ICX-class trains (Pendolino class 390, Hitachi class 800)
160 km/h	Maximum operating speed of LST-class trains (Turbostar class 170)
20 cars / 16 cars	Typical lengths of HST-class trains (nominally 400 m / 320 m), each car ~20 m long
9 cars / 5 cars	Typical lengths of ICX-class trains (nominally 234 m / 130 m), each car ~26 m long
4 cars / 2 cars	Typical lengths of LST-class trains (nominally 96 m / 48 m), each car ~24 m long

Table 4: Acceleration and braking parameters used in the simulation

Parameter	Description
0.4 m/s^2	Average acceleration for an HST-class train (amortized over whole speed-range)
- 0.2 m/s^2	Minimum service braking ("low" braking force), minimum in normal conditions
- 0.5 m/s^2	Normal service braking ("medium" braking force), maximum in wet conditions
- 0.7 m/s^2	Full service braking ("high" braking force), normal maximum in dry conditions
- 1.3 m/s^2	Maximum emergency braking ("maximum" braking force), may cause damage
- 0.1 m/s^2	Compromised braking ("slipping" on icy, or leaf-slime coated rails)

The acceleration of an HST-class train is actually non-linear and initially closer to 0.5 m/s2, but later falls off to 0.3 m/s^2 as the train's speed increases, so 0.4 m/s^2 is used as an amortized figure over the whole period (c.f. [26]). Trains may apply braking force within a normal range ("service braking") or in an emergency may apply full braking force ("emergency braking"). Emergency braking is only used as a last resort, since this causes damage to the train and track: wheels that lock and slide may develop wheel-flats and the rails may also suffer profile damage. We assume a "normal service braking" force of - 0.5 m/s^2, which takes into account poor rail conditions in wet weather [27, 28], where in dry conditions the "full service braking force" of - 0.7 m/s^2 is assumed as the normal maximum [28].

Under poor rail conditions in wet weather, "normal service braking" is the maximum force that may be safely applied without risk of wheel-slide. Under icy rail conditions or during the autumnal leaf-fall where, in certain areas, there is a risk of leaf-slime forming on the rail head, wheel-to-rail adhesion is seriously compromised [29]. The coefficient of static friction may be reduced from 0.5 to 0.05, and braking distances may increase ten-fold. Our simulation models this, such that trains applying braking force greater than the safe maximum force must declare an emergency.

The maximum gradient on adhesion-based railways in the UK is 4%. This was based on data collected from railway blog sites, where drivers reported "worst hills" on the UK rail network:

- Lickey incline, near Bromsgrove: 1:37, or 2.7%
- Dainton bank in Devon: 1:36, or 2.8%
- Approach to Farringdon: 1:27, or 3.7%

Other materials consulted included documents on the safe movement of trains [30, 31], the requirements for sighting and placing signals [32, 33] and a state-of-the-art analysis of train position estimation using a fusion of odometer, accelerometer and Doppler radar [34]. We also commend the work of Wu, et al., who investigated counter-measures to potential cyber-security attacks on balises (including jamming, telegram faking, balise re-positioning, and telegram replaying) [35].

3.4 Software Architecture

The software for the simulator was developed in the Java programming language. The main advantage of Java, as an object-oriented programming language, is that it supports direct 1:1 modelling in software of the various simulation concepts described above in sections 3.1 and 3.2. We chose Java because of its ability to support rapid prototyping. In any future large-scale simulation, we would seek to encode the same concepts in the input languages to FLAME [5] or FLAME GPU [6], which support scaling up to massive levels.

The main statistics of the simulator software package are:

- The Java archive (JAR) file for the main simulator package is 27Mb in size;
- There are around 9000 lines of code; of which 90% is Java and 10% is JavaScript/HTML;
- Typical simulation run-times are 1 – 30 minutes, depending on their complexity;
- A simulation job is imported as a text file (in YAML format) describing the landscape;
- Simulation statistics are exported as compressed, comma-separated value (CSV) files;
- The simulator software has been released as Open Source [36] on GitHub[3] with an accompanying website [37].

Figure 5 gives a high-level overview of the simulator architecture. This shows dependencies between the different packages responsible for the map, the track, the trains, the signalling and the journey. Each of these packages contains many datatypes responsible for different concepts of the simulation. Full advantage was taken of Java language features to capture generality when specifying components with common behavioural interfaces. The design includes a number of unit and integration testing concepts that execute parts of the simulation to help assure correct behaviour.

The main simulator is a standalone Java program that reads railway landscapes as text files and then executes the given simulation. The simulator program is linked, via web sockets, to a web-interface developed in HTML and JavaScript that displays current status information for each running train and supports limited user-interaction with trains. The web interface includes a visualisation that renders the positions of trains along a given route populated by stations and signals. The simulation may be run at arbitrary fine-grained to coarse-grained tick-intervals, and typically executes many times faster than real-time (the speed depends on the tick-interval and the number of objects simulated).

The simulator is currently set up to run in two modes:

One mode is interactive, linked to the web visualisation interface; this is a useful mode of operation for observing the behaviour of queued trains along the whole route and for detecting any bottleneck effects (such as “Mexican wave”-style behaviour). Figure 6 is a screen-shot of the web interface, showing above, the visualisation of the progress of trains along the route (with a station and signals also indicated) and below, the controls indicating the current status of individual trains.

The other operating mode is a batch-mode, where the simulator is run offline, for the sake of collecting journey statistics. The statistics are exported as compressed CSV-files. We used the statistical environment R to analyse the results and generate views of this data. Because of the large numbers of simulations to be conducted, each taking from 1-30 min to execute, we scheduled these to run in parallel on the Sheffield Advanced Research Computer (ShARC), a high-performance grid computing engine dedicated to research computing [38].

[3] See GitHub repository: https://github.com/sinaa/train-simulator and GitHub website: https://sinaa.github.io/train-simulator/

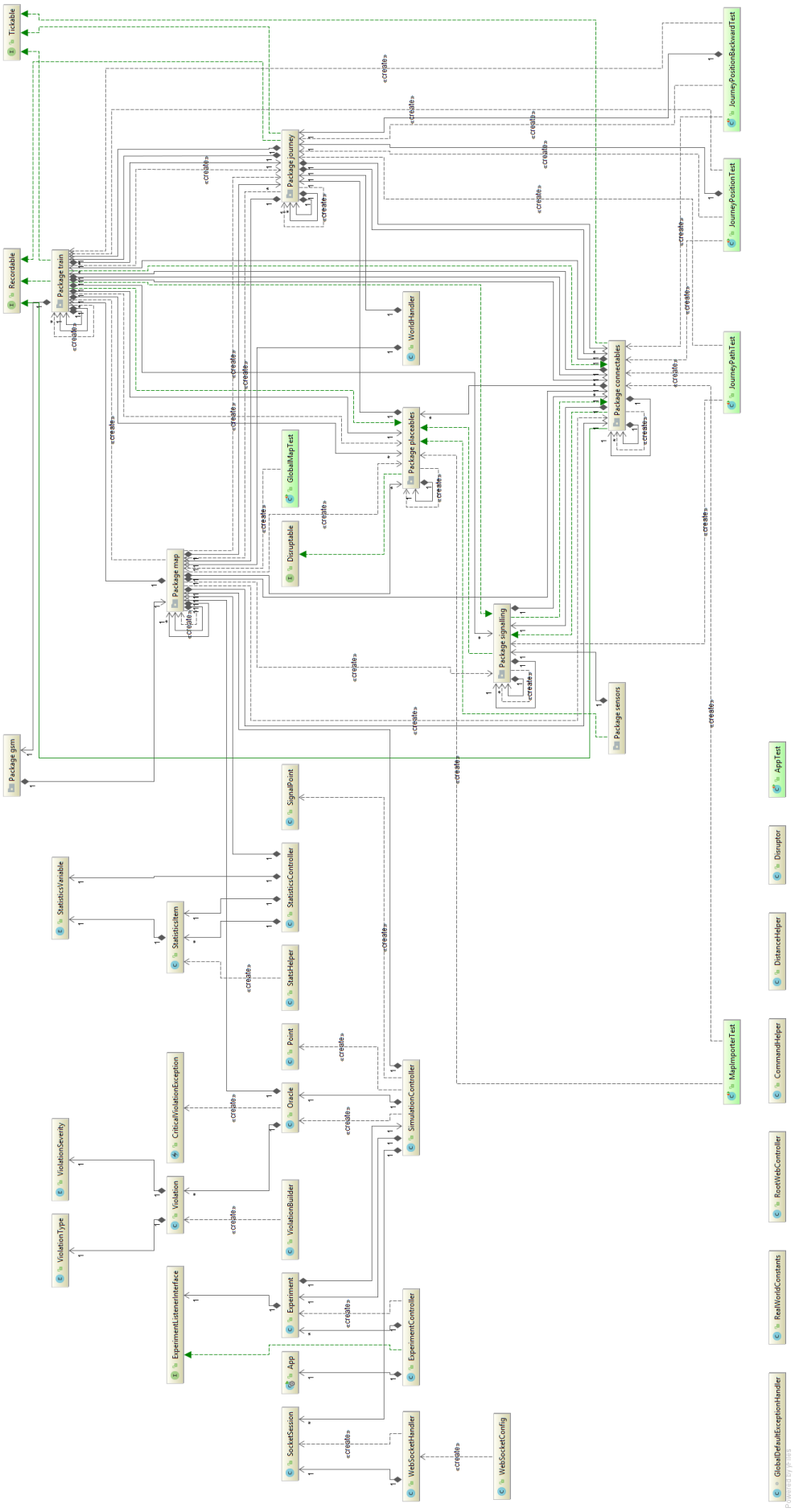

Figure 5: High-level Overview of the Simulator Architecture

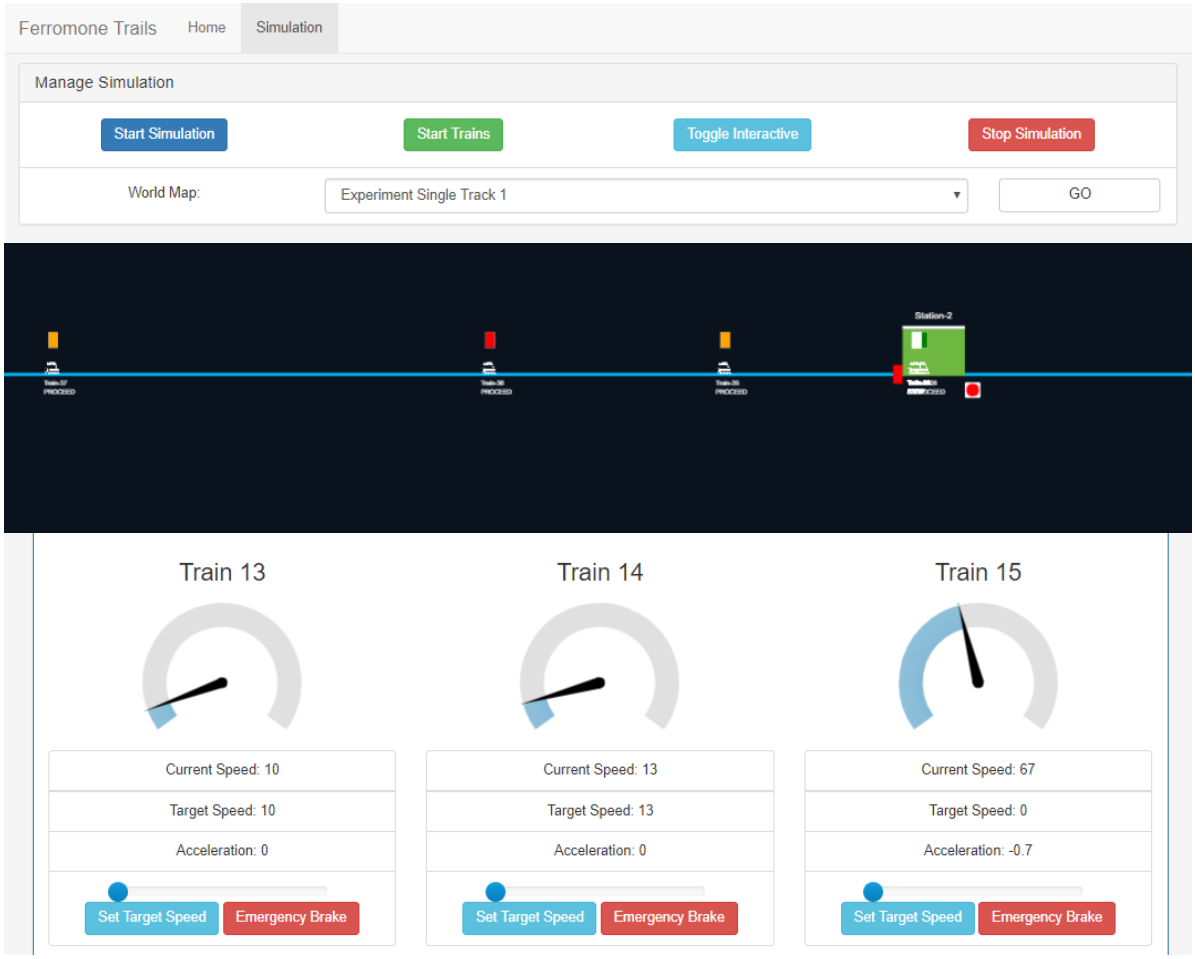

Figure 6: Screenshot of the Web Visualisation Interface

3.5 Railway Landscapes for Simulation

The initial input to the simulator is a text file (in the YAML format) giving a configuration of the landscape, composed of different track sections (of various lengths, with different line speeds) and with different numbers of intermediate stations and optional signal-controlled pinch-points, and a different schedule of trains. Table 5 illustrates the typical set-up for a High-Speed Train simulation:

Table 5: High-Speed Landscape with End-to-End Run

Configuration	Description
Route	Fast end-to-end line between termini, of total length 202 km
Stations	Two stations at the beginning and end of the route
Signals	Signalled 1 km blocks on entry/exit at each station; no other pinch-points
Track Sections	15 block sections of minimum length 1 km to maximum length 10 km
Balise Spacing	50 m spacing of ferromone balises
Line Conditions	10 km bend; 20 km incline; 20 km decline
Train Sets	HST-class trains in 16-car sets; ICX-class train sets in 5-car sets
Traffic Volume	Run simulations for 1, 10, 20, 30, 50, 100 train sets
Departure Cadence	Trains released at 1, 2, 3, 5, 10, 20 minute cadences
Simulation #1	Fixed-block control of all trains, as a null hypothesis
Simulation #2	Dynamic-block simulation, using single-line ferromone trails
Simulation #3	Dynamic-block simulation, using dual-line ferromone trails

This high-speed run only contains two station termini and 132 km of uninterrupted track (with maximum realistic speeds of up to 300 km/h). To introduce some variation into the run, a 10 km long bend in the line is introduced shortly before the half-way point; and then another high-speed section leads to a 20 km long incline causing the train to lose acceleration, followed by a 20 km long decline, causing the train to gain acceleration, before reaching the final terminus station. There is only one signal-controlled section of track, which is 1 km long, situated immediately after the departing station. This is used to ensure that a 400 m long train has cleared the station environs, before allowing the next train to depart.

Table 6: Inter-City Landscape with Intermediate Station Halt

Configuration	Description
Route	Fast end-to-end line with intermediate station, of total length 132.4 km
Stations	Three stations at the beginning, middle (400 m platform) and end of the route
Signals	Signalled 1km blocks on exit from each station; no other pinch-points
Track Sections	15 block sections of minimum length 1 km to maximum length 10 km
Balise Spacing	50 m spacing of ferromone balises
Line Conditions	10 km bend; 20 km incline; 20 km decline
Train Sets	HST-class trains in 16-car sets; ICX-class train sets in 5-car sets
Traffic Volume	Run simulations for 1, 10, 20, 30 train sets
Departure Cadence	Trains released at 1, 2, 3, 5, 10, 20 minute cadences
Simulation #1	Fixed-block control of all trains, as a null hypothesis
Simulation #2	Dynamic-block simulation, using single-line ferromone trails
Simulation #3	Dynamic-block simulation, using dual-line ferromone trails

Table 6 illustrates the typical set-up for an Inter-City Express simulation. This inter-city run includes an intermediate station, at which all trains must halt, in addition to the two station termini. The landscape is otherwise similar to the high-speed line (we allow trains to run up to 300 km/h, rather than 225 km/h). The intermediate station is placed shortly before the 10 km bend. The departing and intermediate stations both have 1 km signal-controlled sections of track, to ensure that the station environs have been cleared, before admitting the next train onto the main line. The intermediate station adds 400 m to the length of the route. Station dwell-times are set at 2 min for the intermediate station, after which the train is released onto the main line, subject to signalling.

Table 7: Extreme Emergency-Triggering Landscape

Configuration	Description
Route	Fast end-to-end line with intermediate station, of total length 132.4 km
Stations	Three stations at the beginning, middle (400 m platform) and end of the route
Signals	Signalled 1km blocks on exit from each station; no other pinch-points
Track Sections	15 block sections of minimum length 1 km to maximum length 10 km
Balise Spacing	50 m spacing of ferromone balises
Line Conditions	Sudden 90% line speed reduction; 20 km incline; 20 km decline
Train Sets	Super-HST-class (460 km/h) train sets in 9-car and 5-car configurations
Traffic Volume	Run simulations for 1, 10, 20, 30 train sets
Departure Cadence	Trains released at 1, 2, 3, 5, 10, 20 minute cadences
Simulation #1	Fixed-block control of all trains, as a null hypothesis
Simulation #2	Dynamic-block simulation, using single-line ferromone trails
Simulation #3	Dynamic-block simulation, using dual-line ferromone trails

We also simulated more extreme landscapes, with widely-varying advisory line speeds, to examine the ability of trains to brake suddenly, avoiding collision with the preceding train. One of these landscapes shown in

table 7 allowed line-speeds up to 460 km/h and immediately followed this fast section by a restricted section, where the line-speed was only 46 km/h (one tenth of the previous line-speed). We expected this to trigger emergency braking in some line-running scenarios.

4. Simulation Results

This section describes the results obtained from simulating different railway landscapes. Altogether, we conducted 744 experiments under normal conditions and a further 126 experiments under degraded line conditions (after validating the simulator set-up, as described below), which included all combinations of:

- end-to-end high speed run, versus inter-city run with intermediate station halt;
- fixed-block signalling, versus variable dynamic block control using ferromone trails;
- single-line ferromone trail laying, versus dual-line ferromone trail laying;
- HST-class train sets with 16 cars, versus ICX-class trains with 5 cars;
- Departure cadences set at 1, 2, 3, 5, 10, 20 minutes;
- Balise spacing set at 50, 100, 200, 500 metres.

The most important statistics collected from the simulation were:

- the station departure cadence (rate of releasing trains);
- the total number of trains occupying the route (line occupancy);
- the maximum, minimum and average journey duration (journey times);
- the standard deviation of journey speeds (smoothness of travel).

The departure cadence is directly related to the notion of rail capacity, such that capacity increases as the interval between departures drops, so it is important to know the smallest safe interval. In fixed-block operation, the maximum line occupancy is dictated by the total number of block sections, so any increase in line-occupancy under dynamic-block operation is also significant. As trains self-organise, they regulate their own speeds independently, so journey times may vary, and the limits of this are important to know. As speed and headway planning becomes more accurate, the smoother the journeys will be, because trains will not make as many braking and accelerating adjustments; and so will be more energy-efficient.

4.1 Initial Validation

The simulator was validated initially by observation, running an example route and visualising the behaviour of a sequence of trains on the route (see figure 6). We ran at tick-speeds of 10ms (100 updates per sim-second), but later changed this to 100ms (10 updates per sim-second) without noticeable loss of fidelity. Note that simulated time (in sim-seconds) runs at many times real-time; complete simulation runs execute in variable real-time (typically from 1-30 minutes) depending on their complexity.

- We tested the approach to a station by placing passive balises with decreasing advisory line speeds to control deceleration, and observed trains following the correct braking pattern.
- We tested the self-correcting nature of balise-based control, by building routes with sudden changes in advisory line speeds, and observed trains applying stronger corrective braking force.
- We tested the realism of the physics engine by manually overriding the speeds of trains (through the user interface controls, see figure 6), to force them to approach stations too rapidly, such that they applied maximum braking force and still overshot their platform.
- We also deliberately triggered safety violations, by manually overriding the speeds of trains to force them to approach preceding trains too rapidly, to ensure that the simulator reported proximity events, when the trains became too close.

4.2 Testing Extreme Limits

The first evaluation experiments were carried out with an extreme landscape (see table 7). This was a 132.4 km route with an intermediate station after 41 km. The line allowed line-speeds up to 460 km/h, but featured a severe speed restriction to 46 km/h after 21 km, an incline (lasting 20 km) and decline (also lasting 20 km) after 62.4 km, followed by a fast 30 km approach to the terminal station. Trains were controlled by signals for 1 km upon exit from the stations. The dwell-time for the intermediate station was 2 minutes.

Trains were deliberately allowed to run at unrealistically high speeds (up to 460 km/h) on open line-sections, so that they would be forced to brake suddenly when the line speed was reduced to 1/10th of the open line speed. This was set up to observe the braking performance of trains as they formed queues on the line. The incline and decline were so arranged as to cause a 20% increase, or decrease in the rate-of-change of actual train velocities. This was set up to observe the braking performance of trains under adverse line conditions on the down-hill slope, where only 80% of nominal braking-force could be achieved.

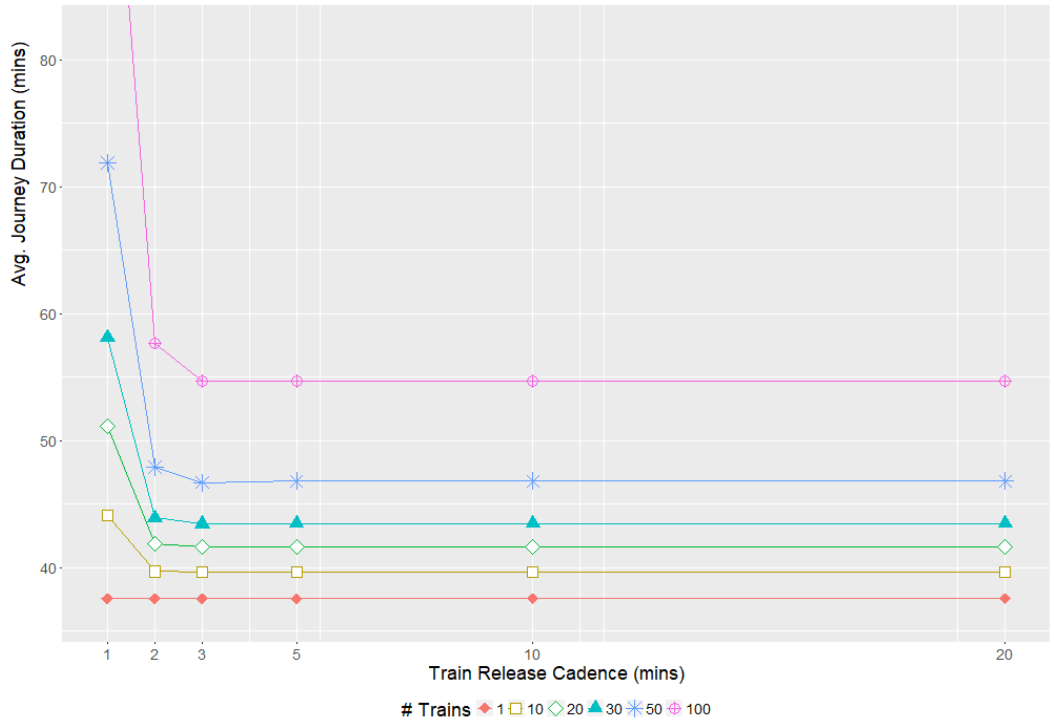

Figure 7: Average Journey Duration against Departure Cadence

The first results for super-HST train-sets with 5 cars are shown in figure 7. This showed, remarkably, that the self-organising control offered by the ferromone-trail sensing system was by itself sufficient to regulate the passage of traffic, even where trains applied strong braking force at the sudden discontinuity in line-speed after 21 km. Figure 7 shows the plot of average journey time against the departure cadence, in minutes, for volumes of traffic ranging from 1–100 trains.

For one train, the optimum journey time over this route was 37.5 minutes (subject to physical acceleration and deceleration limits – see section 3.3). This corresponds to an amortised speed of 212 km/h, which though unrealistically fast (trains may reach 460 km/h), gives best opportunity for disasters. For lines occupied by 10, 20, 30 trains, average journey times increased by 2 minutes per 10 trains scheduled, so long as the departure cadence was no shorter than every 3 minutes. These small increases were due to trains adjusting their headways to account for trains in front; and this had a cumulative effect. At shorter intervals, trains queued to clear the 1 km signal-controlled section after each station; but at 3-minute, or longer intervals, the performance was stable and predictable from the number of trains scheduled.

The given route started to become saturated when 20-30 trains were scheduled. At 2-minute departure cadences, the main line was occupied by 10-11 trains (not counting trains halted in stations); and at 1-minute departure cadences, the main line was occupied by 11-12 trains. Line occupancy could be increased by scheduling 50-100 trains, in which case 2-minute departures achieved 12-16 trains on the line, and 1-minute departures achieved 14-18 trains on the line, for greatly-increased journey times, due to queueing. The maximum theoretical capacity of the same route under fixed-block control would be 15 trains.

We expected the safety characteristics of this extreme simulation to be challenging, because of the sudden braking required after 21 km. However, we found that for shorter 5-car trains, the leading trains applied *normal service braking*, and following trains calculated a braking curve between this and *full service braking*, adjusting their speeds to match the estimated positions of the trains in front. When many trains were scheduled, they tended to bunch up as a cohort. When we ran this extreme simulation with longer 16-car trains, we observed proximity violations during *full service braking*; this was due to the longer 320 m train-length not giving sufficient train-separation. However, when the emergency control layer was added, trains applied increasingly stronger braking as above, until the first train exceeded *full service braking* and emitted an emergency signal, which caused all following trains to brake immediately. When trains were closely-spaced, this typically triggered a cascade of emergency braking in all following trains.

4.3 High-Speed End-to-End Performance

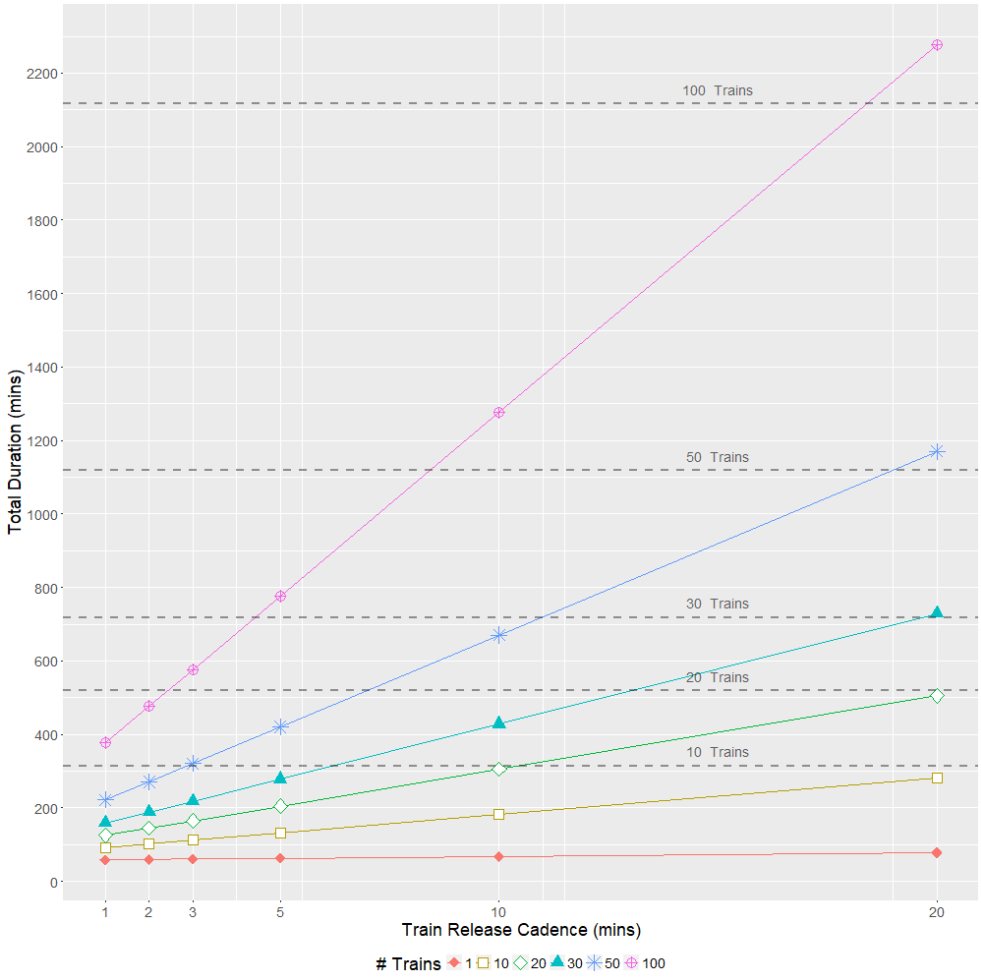

Figure 8: Comparing High-Speed Capacity Achieved with Dynamic- and Fixed-Block Control

The first set of more realistic experiments were conducted with ICX-class 5-car configurations on a high-speed 202 km run between two termini lasting about 64 minutes (see table 5). Figure 8 shows the headline gains in

capacity for dynamic block signalling using ferromone trails, against fixed-block signalling. The X-axis shows the variable departure cadence under dynamic-block control; and the Y-axis shows the total duration of the simulation run (total time taken to move all scheduled trains from one terminus to the other). Fixed-block sizes were set at 10 km (after advice from Connor [26, 27] that over 7 km is needed to bring an HST-class train to a halt). The shortest total journey durations for fixed-block control are marked as horizontal thresholds: these are constant for given numbers of trains, since trains may only be released as fast as the blocks will admit them. This shows that with ferromone-trail based dynamic block control, a *five-fold capacity gain* can be made for 3-minute departure cadences (we can move 50 trains, versus 10 trains, in the same period); a *three-fold capacity gain* can be made for 5-minute departure cadences; a *two-fold capacity gain* for 10-minute departure cadences; and dynamic-block capacity only degrades to fixed-block capacity at 20-minute departure cadences.

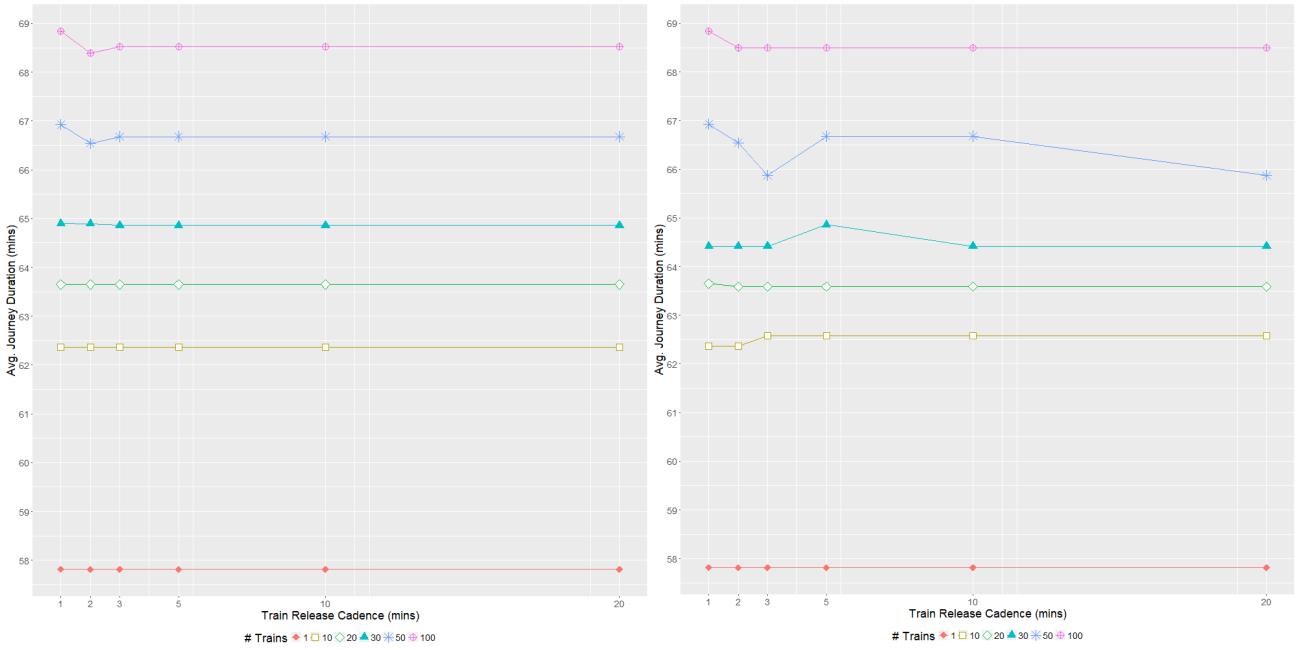

Figure 9: Fast Journey Times for (a) Single-Line and (b) Dual-Line *Ferromone Trail* Control

Figure 9 shows the average journey times on the Y-axis for different traffic volumes, at different departure cadences on the X-axis. Whereas a single unimpeded train can complete the route in 58 minutes, 10 trains adjusting their headways will take 62.5 minutes, and for every 10 more trains, this increases the journey time by about 1.5 minutes. Figure 9 (a) is stable across all departure cadences (c.f. the extremes in figure 7), due to the realistic and smoothly-changing advisory line-speeds. Journey durations in figure 9 (b) are broadly similar under dual-line ferromone trail laying, with the best time-gain of 1 minute for 50 trains running at 3-minute departure cadences. Note that dual-line control only starts to improve headway planning after the lines are full in both directions: in low-traffic simulations, the trains only cross paths after half the run. Journey times for dual-line control were largely unimproved over single-line control; but we thought that dual-line control might smooth out speed-variations over the journey.

To test this hypothesis, we measured variations in speed. Figure 10 shows standard deviations (SD) of speeds in m/s, sampled every sim-second on the Y-axis, for different traffic volumes, at different departure cadences on the X-axis. The ground truth is shown for a single train in figure 10 (a), where the SD over the whole journey is 30.75 m/s (the train accelerates up to 300 km/h and brakes to enter the terminus). For 10 trains this increases to 32.25 m/s and thereafter by increments of around 0.3 m/s per 10 trains. Figure 10 (a) is stable over all departure cadences for single-line ferromone trail control. Figure 10 (b) shows the same measurement under dual-line ferromone trail control. The results are interesting, because they are not stable.

While most SDs are just slightly lower, for 20 trains the SD drops sharply at a departure cadence of 2 minutes; whereas for 10 trains the SD increases sharply at a cadence of 3 minutes. This is not what we expected; and requires some interpretation. Our best guess is that the headway-planning benefit of receiving more accurate data about the preceding train must depend on the precise intervals and places at which trains cross each other's paths; and the emergent behaviour is slightly chaotic (in the mathematical sense).

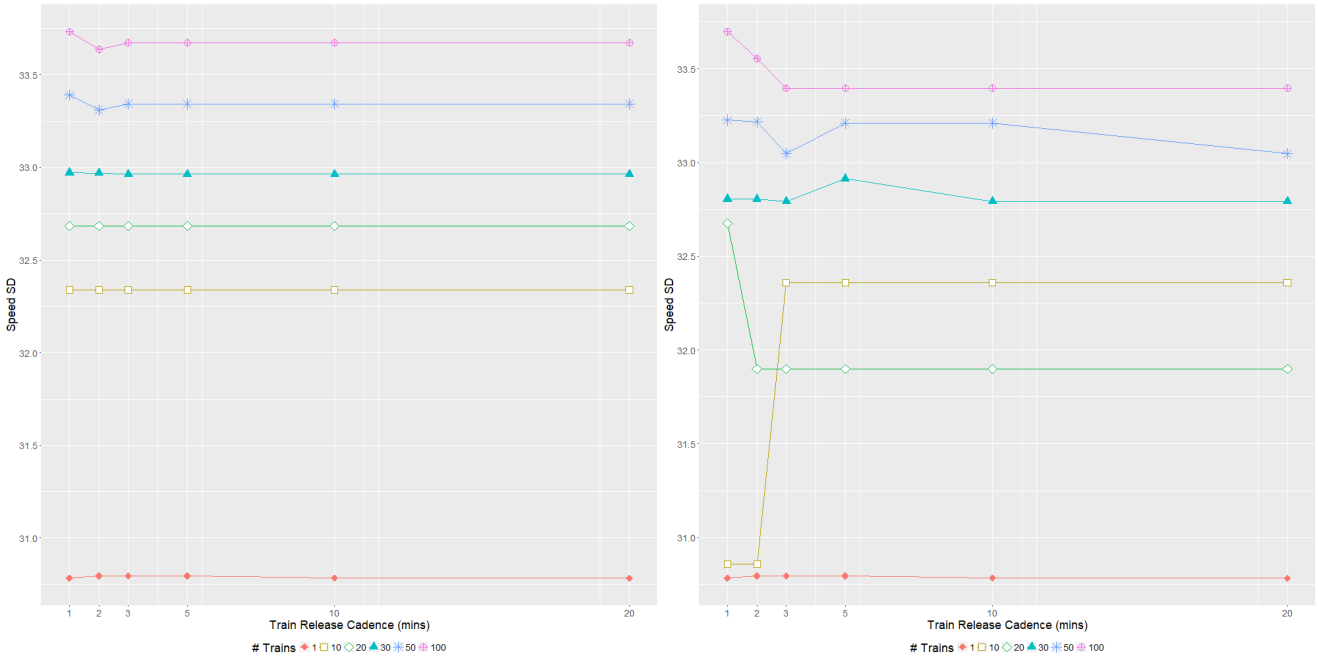

Figure 10: Speed Variation for (a) Single-Line and (b) Dual-Line *Ferromone Trail* Control

Figures 8 to 10 show performance data for ICX-class 5-car train configurations, operating at up to 300 km/h. The second set of experiments was conducted on the longer HST-class 16-car train configurations, whose journey performance was found to be similar to the above (the trains accelerate and brake at the same rates), so we do not duplicate these figures. However, there were other small differences.

We also recorded any safety issues. For the ICX-class 5-car train configurations, we experienced no proximity violations and triggered no emergencies under any of the above-described operating conditions, even when operating at higher than expected maximum speeds. For the longer HST-class 16-car trains, there were no safety problems when balises were spaced closely at 50 – 100 m intervals (c.f. the extreme scenario, in which emergencies did occur – see section 4.2). However, when balises were spaced more generously at 200 – 500 m intervals *and* the train departure cadence was set at 3-minute intervals or shorter, we did experience 20 proximity violations (that would trigger the emergency shut-down of the route). This suggests that HST lines should have more in-fill balises and leave larger intervals between departures.

4.4 Inter-City Station Serving Performance

The third set of experiments was conducted with ICX-class 5-car configurations on a slightly lower-speed 132.4 km run between cities, with an intermediate station placed after 41 km (see table 6). Trains were allowed to accelerate up to 300 km/h in the open stretches, and did indeed reach this speed on the open stretches between stations (see figure 13). The dwell-time at the intermediate station was set at 2 minutes. Figure 11 (similar to figure 8) shows the headline capacity gains for dynamic-block over fixed-block control. The Y-axis displays the total time taken to move a given volume of traffic from one terminus to the other, via the intermediate station. If we use the same cadence-intervals as above, we gain a 3-fold increase in capacity at a 3-minute departure cadence, a 2.5-fold increase in capacity at a 5-minute departure cadence, and a 1.5-fold increase at a 10-minute departure cadence.

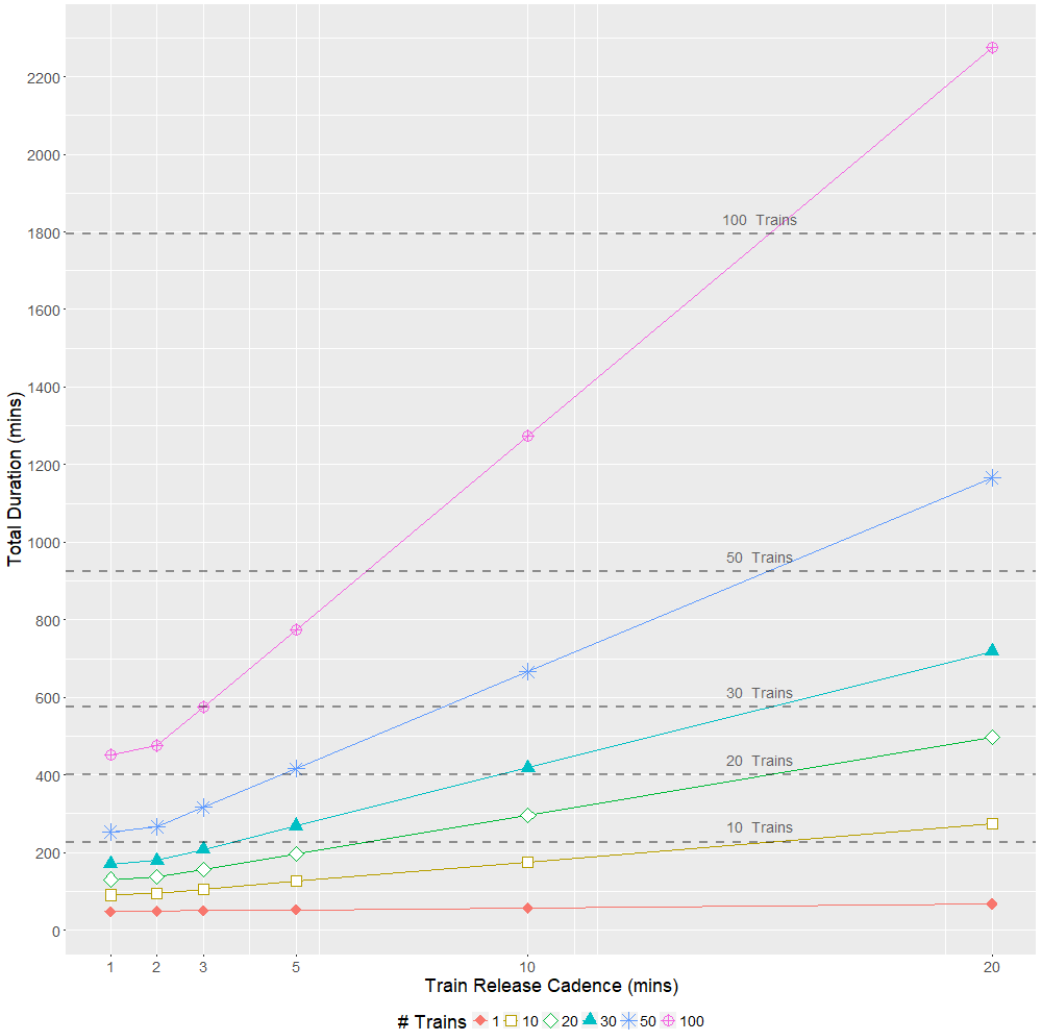

Figure 11: Comparing Inter-City Capacity for Dynamic and Fixed-Block Control

Figure 12 compares journey durations for single-line and dual-line ferromone control (note that the Y-axes are not displayed to the same scale in figure 12 (a) and 12 (b), due to the different dynamic ranges). The Y-axis gives average journey times for the stopping journey, against different departure cadences and traffic volumes on the X-axis. The ground truth is given by one train, which takes 47 minutes to complete the whole journey, following the advisory line-speed indications (so travelling at an amortized speed of 174 km/h). Figure 12 (a) illustrates single-line control. Here, for 10 trains, the average journey time stabilises at 54 minutes (from 2-minute departure cadences). For 20 trains, this increases to 55 minutes, and for 30 trains, to 56 minutes. All journey times seem stable for a departure cadence of 2 minutes, or longer. At a 1-minute cadence, increased traffic results in much longer journeys. This is due to queueing to get past the controlled blocks on entry to, and exit from, stations.

Figure 12 (b) illustrates dual-line control, in which trains receive updated information about the preceding train. Journey times are only slightly improved (by around a minute) over the stable range of departure cadences. 10 trains take on average 53 minutes, 20 trains take 54 minutes and 30 trains take 54.5 minutes (from 3-minute departure cadences). Dual-line ferromone control seems to bring most advantage for very short departure cadences of 1-2 minutes. Whereas 50 trains take on average 77 minutes under single-line control, under dual-line control this reduces to 60.5 minutes at a 1-minute cadence, and to 60 minutes at a 2-minute cadence. By contrast, it would take on average 84 minutes for 50 trains to make the same journey under fixed-block control.

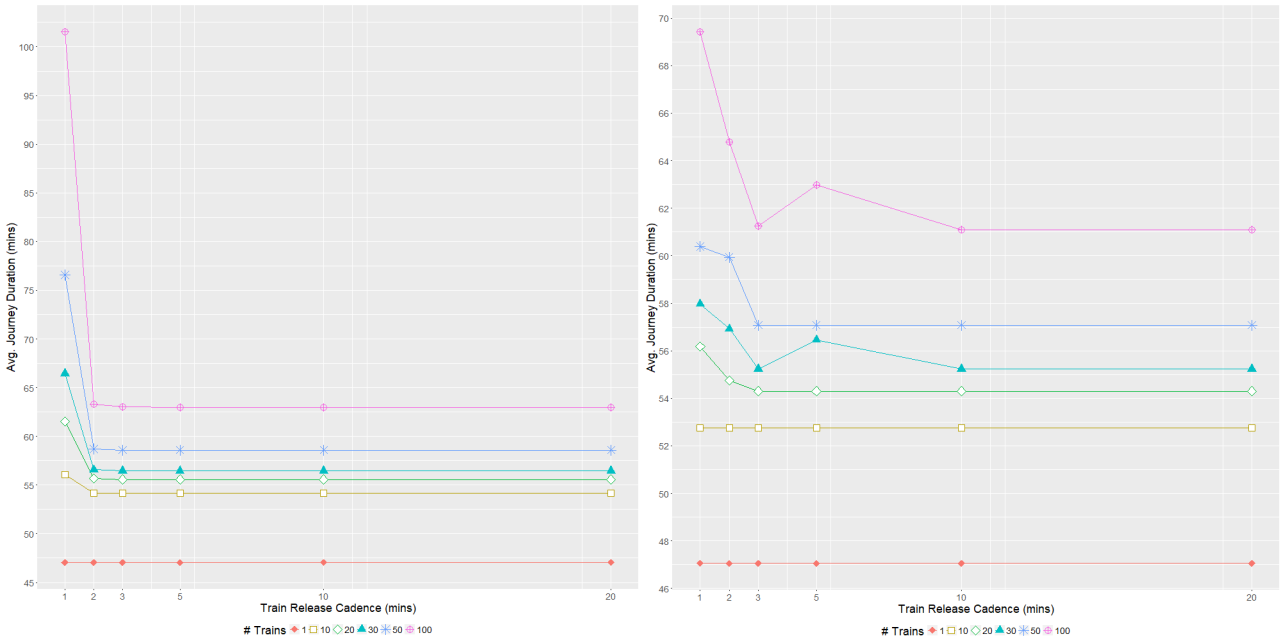

Figure 12: Stopping Journey Duration for (a) Single-Line and (b) Dual-Line Ferromone Trails

While pushing 50 or 100 trains onto the line at the cost of longer journey times may seem like a bad idea, a much higher line-occupancy statistic may be obtained in this way. The highest line-occupancy we achieved (trains active on the main line; not in any station) was for HST-class trains in the 16-car configuration, where we achieved 36 active trains when pushing out 100 trains at a 1-minute cadence. The second-highest line occupancy was for ICX-trains in the 5-car configuration, where we achieved 18 active trains (for similar loading and cadence). Trains took an average of 102 minutes to complete this journey. Under fixed-block control, the maximum line capacity is 15 trains.

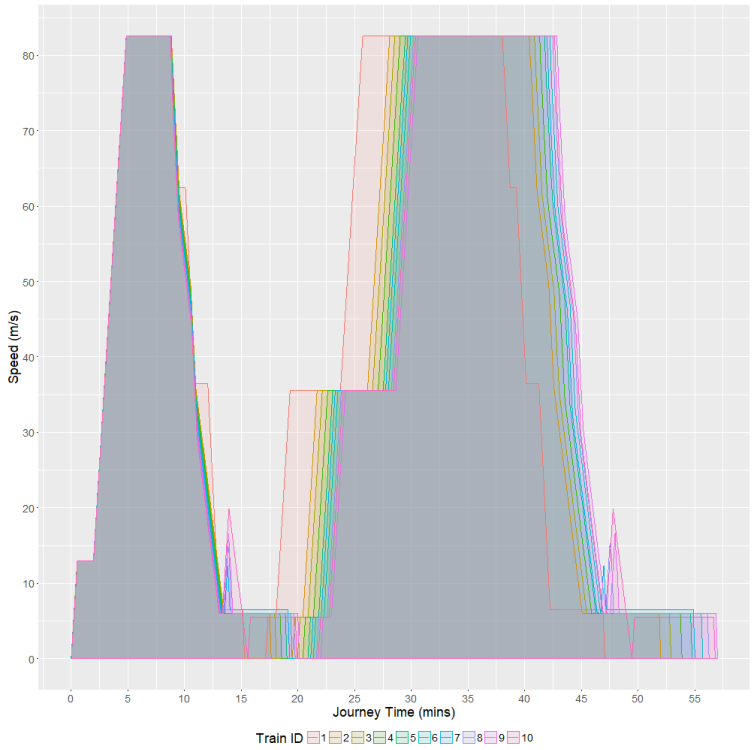

Figure 13: Speed Profiles for 10 Inter-City Trains, with 3-min Departure Cadence

Figure 13 shows the speed profiles of 10 trains, released with a 3-minute departure cadence, overlaid to show slight differences between earlier and later trains. They reach their maximum speed of 83.33 m/s (300 km/h) on the open line stretches between the stations, and follow similar braking patterns on approach to the intermediate station. However, there is some negotiation on entry to this station, such that they depart with more staggered delays. The later-arriving trains in the queue experience a slight "Mexican Wave" effect on approach to stations, where they initially brake harder, but then accelerate again as the line clears. Figure 13 also shows how acceleration is limited to 35 m/s at the bend placed after the intermediate station. Following this, the uphill and downhill gradients affect slightly the acceleration and braking of the trains.

4.5 Safety under Balise Degradation

One of our goals was to measure how safe the balise-based control system would be under conditions where the trackside equipment was degraded. We combined this with an investigation into ideal balise-spacing. Siemens *Eurobalises* are normally spaced at 0.5–1 km intervals, typically in pairs spaced 3 m apart, with one passive balise and one active balise linked to a Lineside Electronic Unit (LEU). Pairs of numbered balises tacitly allow trains also to detect their direction of travel, from the order in which the numbers are received.

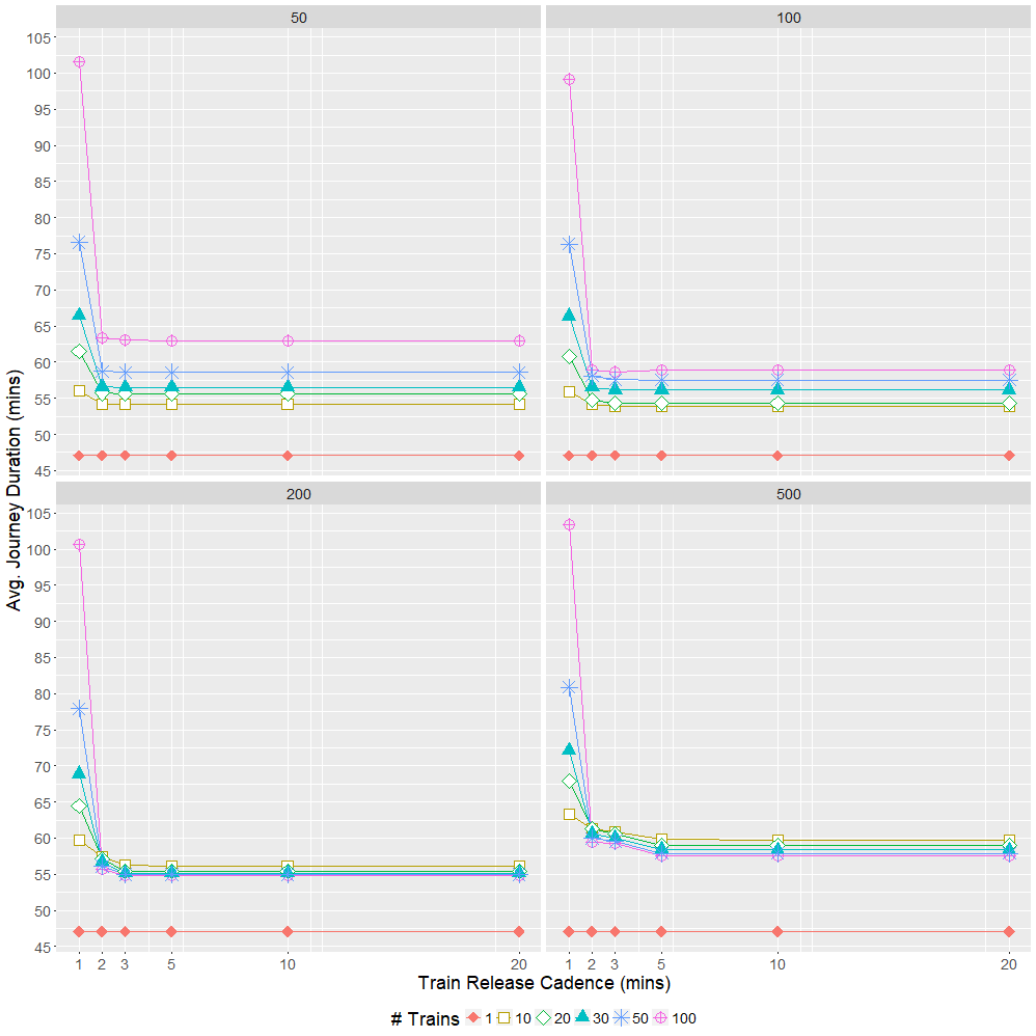

Figure 14: Effect of Balise-Spacing on the Duration of Inter-City Journeys

We explored installing *ferromone trail* read/write balises at different intervals ranging from 50–500 m. The shorter intervals are based on the notion that trains should benefit from more regular updates about the timestamp, speed and position of preceding trains. The longer intervals represent either a lower-cost installation option, or evenly-spaced balise degradation. Figure 14 shows the effect of variable balise spacing

(50, 100, 200 and 500 m) on the average journey times of trains on the Inter-City route (see table 6). Here, we must discount the journey time of the single train that always takes 47 minutes in each scenario. Otherwise, for closely-spaced balises (at 50 – 100 metre intervals), it is clear that trains negotiate their headways much more carefully, as the volume of traffic increases. At 50-metre balise spacing, the spread of stable journey times ranges from 54 minutes (10 trains) to 63 minutes (100 trains). At 100-metre balise spacing, this range is becomes more compressed, from 54 minutes (10 trains) to 58 minutes (100 trains), as less information is present to plan headways. At 200-metre balise spacing, the range is the most compressed at around 55-56 minutes for all traffic; journey times are dominated more by the advisory line speeds, which all trains are following.

This seems to be a "sweet spot" since, at longer 500-metre balise spacing, the range spreads out again from 56-60 minutes; and journey times take two minutes longer overall. Note however, in these last two cases, that the plotted traces cross over at the 2-minute cadence mark, indicating a lack of continuity: the 100-train traffic takes slightly less time than the 10-train traffic! This could be noise; but the exact reversed order of traces in the 500-metre case suggests some other effect, such as averaging over larger groups of trains.

The "sweet spot" was also suspicious, since it suggested that trains might run autonomously under static line-speed indications alone. We tried this counter-example and found indeed that it worked well for sets of ICX-class 5-car trains that all followed the line speed profiles exactly, and were released at no shorter cadence than every 3 minutes. However, we also observed 25 proximity violations for 16-car trains, 5 during dual-line control in the 3-station configuration, and 20 during single-line control, where balise spacing was dropped to 200, or 500-metre intervals and trains were being released at shorter 1-3 minute cadences.

The answer is obvious in retrospect, which is that no special controls are needed if all trains observe exactly homogeneous line speeds and are released with sufficient headway; however, the ability to re-negotiate headways is critical, if you have sudden variations in line speed, or if you have mixed-class traffic. For this reason, we concluded that 100-metre balise spacing was as sparse as you should dare to go in a *ferromone trails* control system, if you wished to maintain normal operations with 3-minute departure cadences.

To test the effects of realistic balise degradation, we experimented with a range of probabilistic disruption ratios (1%, 2%, 5%, 10%, 20% and 50%) for balises spaced at 50-metre intervals. Each balise independently had this probability of failure. Under the lower 1-2% ratios, the chance of two consecutive balise failures was small; under the 50% failure rate, consecutive numbers of failed balises followed a coin-tossing expectation, with the possibility of longer runs. Under normal train operations discussed above, we found that even this did not disrupt how the network functions, that is, trains continued with their best information about headway and advisory speeds. However, it reduced the safety of train operations in artificially-induced extreme circumstances (such as unexpected hard decelerations, short of an emergency stop) combined with short 1-2 minute release cadences, in which proximity violations could occur. We concluded that placing the ferromone balises 100 metres apart is acceptable with 1-2% degradation, and placing them 50 meters apart is advisory otherwise.

4.6 Safety under Emergency Operations

We ran the simulator both with, and without the additional layer of GSM-R radio bursts to signal train liveness and emergencies. All the events described above as *proximity violations* occurred when a train got within one metre of another train under balise-based control. The cause was invariably due to late headway planning, combined with having (only) "full service braking" as the maximum braking force that could be applied. When the emergency layer was enabled, trains could apply "emergency braking" force, and also emitted the emergency signal to notify the immediately-following train.

This strategy was always successful in preventing proximity violations. We created one railway landscape with a special "derailing balise", which automatically triggered an accident in the first train to pass over it. This train applied emergency braking and emitted the emergency signal. Following trains either re-planned their normal braking curves, or if they were already too close to the derailed train, they also applied emergency braking and passed on the emergency to the next train following. If trains were closely spaced in queues, this triggered a cascade of emergency braking. There were no collisions.

The emergency system works in a fail-safe way, so if the live signal is lost for any reason, this must be treated like an emergency. To cater for the possibility that the GSM-R live signal may be lost temporarily, for example because of trackside geography, we allowed a gap of up to two seconds (two chances to recapture the signal), before this was treated as a fail-safe emergency.

In our experiments above, we assumed that the live signal only contains the train ID; and the emergency signal also contains the train's position. This was in order to keep the size of transmitted data packets as small as possible, to avoid GSM-R network congestion. However, it might be feasible for the live signal to contain more information, including the train's latest estimated position and speed. In this case, the GSM-R layer could duplicate the information provided by the balise-layer, providing greater redundancy, in case either system failed temporarily.

In fail-safe operation, the following train must always make a worst-case assumption about the position of the preceding train (whose signal has been lost). The best information about the preceding train could be gained from balises (in our experiments), or from radio packets (if positions are given, up to the point of failure). The spacing of balises affects the frequency with which trains are updated. Assuming that the GSM-R signal updates a train every second about the preceding train, this frequency is equivalent to:

- balises spaced at 50 metres; and the train travelling at 180 km/h;
- balises spaced at 100 metres; and the train travelling at 360 km/h.

If balises are spaced at 100 metres, radio-based notification will update train positions more frequently. If balises are spaced at 50 metres *and* the trains are travelling at greater than 180 km/h, then balise-based notification will update positions more frequently (at intervals of less than one second). However, position information received by GSM-R will always be up-to-date and more accurate, compared to the estimated position calculated from balise data.

To test the effects of realistic GSM-R relay transmitter degradation, we experimented with a range of probabilistic disruption ratios (1%, 2%, 5%, 10%, 20% and 50%) for the GSM-R relay transmitters. Each transmitter independently had this probability of failure, when relaying a liveness or emergency signal. We found that the trains simply responded as per the described emergency protocols, that is, if they lost two consecutive liveness signal transmissions, they raised an emergency and halted. Even with 1% degradation, it was possible to trigger an emergency through two consecutive failures, in which case the line would be shut down safely (all trains in a queue would halt safely). We concluded that this was extremely safe, but possibly inconvenient. We did not have time to investigate other emergency protocols, such as increasing the number of dropped transmissions that were allowed, or using balise-data as a fall-back in case of dropped transmissions. We believe it may be possible to create a protocol that is still safe, but more liberal.

5 Conclusions

We have prototyped a multi-agent simulation of the UK national rail network, which allows trains to self-organise and dynamically re-plan their headway, speed, and reactions to adverse line conditions, by sensing and depositing simple electronic data known as "Ferromone Trails" about the passage of trains. This system

supports dynamic-block (variable-block) headway control and offers greatly increased capacity over traditional fixed block control. We have provided a GitHub repository with all the simulator software [36], and an accompanying website [37]. All of the experimental scenarios described above (870 in total, including the track degradation experiments) are archived on this site, so that our results may be replicated by other investigators.

5.1 Key Findings

We simulated High-Speed routes between two termini, and Inter-City routes with an intermediate station at which trains must stop. Capacity gains using *ferromone trails* dynamic block control are based on routes organised into block-sections that are 10 km long (which is sufficient for an HST-class train to brake to a halt within one block [26, 27]). Key findings were:

- for High-Speed routes, we gained a five-fold capacity increase at a departure cadence of 3 minutes (e.g. we can move 50 trains in the same time that it takes to move 10 trains under fixed block control);
- for Inter-City routes, we gained a three-fold capacity increase at a departure cadence of 3 minutes (e.g. we can move 30 trains in the same time that it takes to move 10 trains under fixed block control);
- average journey times increased by less than 1.3% per 10 extra trains scheduled, and are entirely predictable from the number of trains, for departure cadences of 3 minutes, or longer;
- lines became fully occupied when 20-30 trains are scheduled at 3-minute departure cadences, but traffic throughput is still stable for these and higher traffic volumes, and at longer cadences;
- routes may be run at over-capacity, with shorter departure cadences of 1-2 minutes, at a cost of steeply increased journey times (e.g. Inter-City routes may take 100 trains at a 2-minute departure cadence, for a 34% increase in journey time);
- line occupancy may be increased above the theoretical limit for fixed block control (e.g. line occupancy rose to 18 trains on Inter-City routes with a theoretical maximum of 15 trains);
- bi-directional laying and sensing of *ferromone trails* did not bring any noticeable improvement over single-line control, unless operating at over-capacity, at 1-2 minute departure cadences;
- *ferromone trail* balises are spaced ideally at 50-100 metre intervals, and headway-planning becomes less responsive and less accurate at greater 200-500 metre intervals;
- as more *ferromone trail* balises are removed, train behaviour reverts to simple autonomous control guided by passive balises giving advisory line speeds;
- *ferromone trail* balises are effective on lines with sudden speed discontinuities, simulating mixed-class (two-speed) traffic, and when operating at short departure cadences (1-3 minutes);
- the separate GSM-R emergency reporting system always brought trains to a safe stop, in a variety of induced emergency conditions;
- without the GSM-R emergency reporting system, it was possible to induce proximity violations on 16-car HST-class trains, by removing (incapacitating) balises, or shortening departure cadences.
- balises spaced at 100 metres may suffer a failure rate of 1-2% without unduly disrupting network operations; but at 5% or higher rates of failure, 50-metre spacing is advised.
- increasing braking safety margins (on top of actual braking distances) permits sparser balise spacing, but increases knock-on effects in train queues (decreases flexible responsiveness).

5.2 Limitations

The multi-agent simulation is currently capable of modelling a variety of railway landscapes, consisting of end-to-end routes with optional intermediate stations and other signalled pinch-points. It can be run both in fixed block, and dynamic block modes. It has the capability to model trains of different lengths, operating at different line speeds. However:

- it does not currently model junctions and crossings – these were ruled out of scope in a short project lasting less than four months; junction control can only be modelled using signals as pinch-points;
- wet weather, line gradients, ice and leaves-on-the-line are modelled abstractly as coefficients affecting acceleration and deceleration, rather than explicitly;
- the model of trains is slightly idealised with trains consisting of different numbers of 20-metre long idealised cars; trains accelerate at similar rates and may achieve similar maximum speeds;
- using homogeneous train-sets is a possible threat to validity of the simulation results; however we mitigated this by deliberately triggering extreme changes in train speeds.

At the moment, when trains encounter adverse line and weather conditions, they take longer to achieve their target speeds (we use a standard acceleration profile; and trains may calculate a deceleration profile between normal and full service braking). Slippery track results in trains initially misestimating the required acceleration or braking profile, but in the case of deceleration, they recalculate the required braking force continuously, based on measured actual speed. It would be possible in future to factor information about line gradients into the estimates made by the *Sentinel* controller, so that *simulacrum* expectations about required acceleration/deceleration profiles are closer to those experienced in the *materium*.

While we did not conduct mixed-traffic experiments, we argue that the extreme operating scenario posed a much more difficult problem than any mixed-traffic scenario. In the extreme scenario, trains suddenly encountered previous trains travelling at much slower than expected speeds. In mixed-traffic, a slower train will leave a more predictable trace of its passing, to which a fast-moving train will be able to adapt sooner. In our extreme scenario, we caused trains to apply full service braking, giving no notice to the following train until it crossed a balise logging the previous train's slower speed.

The simulation software, which is available through GitHub [36], is fully extensible, written in Java that closely follows the conceptual models as described. Alterations may be made to the simulation software to capture a greater level of detail, if this is desired, for example, in order to model heterogeneous traffic, or in order to support explicit weather conditions and convert these to line acceleration factors.

5.3 Costs and Benefits

The *ferromone trails* control system is estimated to be much cheaper than the projected ERTMS-3 centralised radio control system, since it uses existing equipment with small modifications.

We believe that it is worth the effort to re-instate the (currently abandoned) data downlink function, in order to provide balises with the ability to support *ferromone trails* control systems. According to Siemens' literature, balises are very cheap pieces of track-side equipment, consisting of a small induction loop and a circuit board with robust data-transfer logic, enclosed in a robust foam/plastic casing [22]. They are designed to be extremely durable. It should be cost-effective to manufacture these at large volumes and install them on the railway network.

We recommend the single-line method of sensing and depositing ferromone trails, which gave the biggest improvements over fixed block signalling. The dual-line method required balises connected across the up- and down-line in pairs, with additional wiring and added induction-loop complexity. Since this method did not bring significant improvements over the single-line method, it could be dropped to save costs.

We would need someone at Siemens to comment on the likely total cost per kilometre of installing the ferromone trail balises. Equipping the on-board Balise Transmission Module with both downlink and uplink functions is assumed to be an extension of the existing induction method, so more a matter of data transfer protocols than additional manufacturing costs.

For most of the simulations reported above, we achieved completely reliable train control using only the balise-based system, with 50-100 metre balise spacing. The GSM-R radio network could possibly be used as-is for the emergency system, without packet-switching, although the latter will be much more efficient. It merely requires a means of train-to-train addressing that transfers short data bursts. This could also use any future radio bands, such as freed-up "white spaces" between about 50 MHz and 700 MHz (arising from the move from analogue to digital television).

5.4 Next Steps

We would like to present this work to the Department for Transport, Network Rail, RSSB, and to Siemens and other principal manufacturers of balise and GSM-R equipment, to put together a larger research consortium to develop these ideas further. It is important that industry partners evaluate this work, so that they may provide contacts who wish to commit to this work.

The next steps will include a proposal for a larger project, possibly funded through EPSRC open calls, or possibly through the recently-announced Industry Strategy Challenge Fund, in which industry provides matched funding alongside EPSRC funding for the academic partner.

In the follow-up project, we would like to model the UK rail network at much larger scale, using the using the FLAME multi-agent simulator developed here at Sheffield [5, 6]. This would use the experience gained from this prototype project to create and design appropriate simulation elements that can execute at a massive scale, on GRID computers or on GPU accelerators. This project, if awarded, would allow modelling of the more complex elements ruled out of scope in this prototype, which would include:

- explicitly modelling trains joining or leaving the main line at point-switches
- explicitly modelling trains crossing over the main line at crossovers
- explicitly modelling mixed-traffic running on the same main line
- investigating alternative complementary data to be deposited on balises
- investigating alternative complementary data to transmitted by radio
- modelling further combinations of balise-based and radio protocols
- investigating better predictive mechanisms to improve journey quality
- investigating whether low-cost wireless technology for balises is feasible
- modelling novel fusions of balise- and GPS-based train location
- modelling novel fusions of on-train and track-side sensing and control

We remind the reader that all of our simulations are publicly available at the GitHub repository [36] and website [37].

References

[1] Delivering a Sustainable Railway, Rail White Paper Cm7176, Department for Transport, July, 2007.

[2] European Rail Traffic Management System, http://www.ertms.com, 2016.

[3] V Radmas, T Bradbury, S Deniss, D Chapman, R Bloomfield and D Fisher, ERTMS level 3 risks and benefits to UK railways, Client Project Report CPR798, Transport Research Laboratory, September, 2010.

[4] D Jackson, M Holcombe and F Ratnieks, Coupled computational simulation and empirical research into the foraging system of Pharaoh's ant (Monomorium pharaonsis), Biosystems 76 (1-3), August-October, 2004, 101-112.

[5] Software Engineering Group, STFC, FLAME website, http://www.flame.ac.uk, 2013.

[6] P Richmond, FLAME GPU website, http://www.flamegpu.com, 2015.

[7] D C Walker, G Hill, S M Wood, R H Smallwood and J Southgate, Agent-based computational modelling of wounded epithelial cell monolayers. IEEE Trans Nanobioscience 3, 2004, 153-163.

[8] M Pogson, R Smallwood, E Qwarnstrom and M Holcombe, Formal agent-based modelling of intracellular chemical interactions, Biosystems, 85 (1), Elsevier, 2006, 37-45.

[9] D E Jackson, M Holcombe and F L W Ratnieks, Trail geometry gives polarity to ant foraging networks, Nature, 432 (7019), December 16 (2004), 907-909.

[10] C Deissenberg, S van der Hoog and H Dawid, EURACE: A massively parallel agent-based model of the European economy, Applied Mathematics and Computation 204 (2), 2008, 541-552.

[11] T Karmakharm and P Richmond, Large scale pedestrian multi-simulation for a decision support tool, Proc. Theory and Practice of Computer Graphics (TPCG) 2012, 41-44.

[12] M Bussiek, T Winter and U Zimmermann, Discrete optimization in public rail transport, Mathematical Programming, 79, 1997, 415-444.

[13] J-P Benfeldt, U Mohr and L Müller, RailSys, a system to plan future railway needs, Proc. Int. Conf. Computers in Railways VII, The Built Environment, 50, Wessex Institute of Technology Press, 2000, 249-255.

[14] J Mills, J Spear and C Brown, Rail Industry Research Strategy, 071205 RIRS Final version, Department for Transport, 5 December 2007.

[15] Department for Transport, Rail Technical Strategy, ISBN 978-0-11-552890-3, The Stationery Office, July 2007.

[16] Transport Research APAS: Rail Transport VII – 35, European Rail Traffic Management System Requirement Specifications, Directorate-General Transport, European Commission, ISBN 92-827-9062-2, Brussels, 1997.

[17] A Ngai, What is ERTMS/ETCS?, Institution of Railway Signal Engineers (Hong-Kong Section) Newsletter, Issue No. 6, March, 2010, http://www.irse.org.hk/eNewsletter/issue06/issue06.htm.

[18] UNIFE (Union des Industries Ferroviaires Européennes), ERTMS Deployment in the UK: re-signalling as a key measure to enhance rail operations, ERTMS Factsheets No. 14, Brussels, 2014.

[19] Transport Research EURET: Rail Transport VII – 5, Eurobalise Sub-System, Directorate-General Transport, European Commission, ISBN 92-827-7993-9, Brussels, 1996.

[20] UNISIG Consortium (Alstom, Ansaldo, Bombardier, Invensys, Siemens, Thales), FFFIS for Eurobalise, Subset-036, Issue 2.4.1, September 27, 2007.

[21] UNISIG Consortium (Alstom, Ansaldo, Bombardier, Invensys, Siemens, Thales), FFFIS for Eurobalise, Subset-036, Issue 3.0.0, February 24, 2012.

[22] Siemens AG, Trainguard Eurobalise S21 and S22 for track-to-train communication, Siemens AG Mobility Division, Berlin, Germany, 2014.

[23] Mark Glover and Ewan Spencer (Siemens UK), Personal email communication, 30 March 2017.

[24] Network Rail, Guide to the GSM-R System, briefing pack, networkrail.co.uk, 2017.

[25] Network Rail, The what, why and how of the GSM-R System, networkrail.co.uk, 2017.

[26] P Connor, Rules for high-speed line capacity, Railway Technical Web Pages, Infopaper No. 3, 26 August, PRC Rail Consulting Ltd., 2011. http://www.railway-technical.com/

[27] P Connor, High-speed railway capacity: understanding the factors affecting the capacity limits for a high-speed railway, Proc. International High Speed Rail 1964-2014: Celebrating Ambition, Birmingham Centre for Railway Research and Education (BCRRE), University of Birmingham, 8-10 December 2014. http://www.railway-technical.com/about-2/books-papers--articles/high-speed-railway-capacity.pdf

[28] V L Winter, R S Berg and J T Ringland, Bay area rapid transit district advance automated train control system case study description , Chapter 6 in: High Integrity Software, ed. V L Winter, Kluwer Academic Publishers, 2001, 115-135. DOI 10.1007/978-1-4615-1391-9_6, ISBN 978-1-4613-5530-4.

[29] T Maciel, A leafy curse: the physics of leaves on the track, Physics Central: Physics Buzz Blog, 28 October, 2014. http://physicsbuzz.physicscentral.com/2014/10/a-leafy-curse-physics-of-leaves-on-track.html

[30] Rail Safety and Standards Board, Preparation and movement of trains, GE/RT8000/TW1 Rule Book, Issue 8, 2008.

[31] Office of Rail Regulation, Safe Movement of Trains, Railway Safety Publication 3, 2nd edition, ISBN 07176-2727-6, 2007.

[32] Rail Safety and Standards Board, Signalling positioning and visibility, Railway Group Standard GE/RT8037, Issue 1, December, 2003.

[33] Rail Safety and Standards Board, Signal sighting assessment requirements, Rail Industry Standard RIS-0737-CCS, Issue 1, June, 2016.

[34] M Malvezzi, G Vettori, B Allotta, L Pugi, A Ridolfi, F Cuppini and F Salotti, Train position and speed estimation by integration of odometers and IMUs, Proc. 9th World Congress on Railway Research, 22-26 May, Lille, 2011.

[35] Y Wu, J Weng, Z Tang, X Li and R H Deng, Vulnerabilities, attacks and countermeasures in balise-based train control systems, IEEE Trans. Intelligent Transportation systems, DOI: 10.1109/TITS.2016.2590579, 2016.

[36] S Shamshiri, Train Simulator: De-centralised Self-Driving Trains, GitHub source code repository, https://github.com/sinaa/train-simulator, June, 2017.

[37] S Shamshiri, Train Simulator: De-centralised Self-Driving Trains, public GitHub website, https://sinaa.github.io/train-simulator/, June, 2017.

[38] M Croucher and P Richmond, High Performance Computing at Sheffield, Research Computing Group, http://docs.hpc.shef.ac.uk/en/latest/, 2017.