On Determinism of Game Engines used for Simulation-based Autonomous Vehicle Verification

Abanoub Ghobrial, Greg Chance, Kevin McAreavey, Severin Lemaignan, Tony Pipe, Kerstin Eder

***Abstract*—Game engines are increasingly used as simulation platforms by the autonomous vehicle (AV) community to develop vehicle control systems and test environments. A key requirement for simulation-based development and verification is determinism, since a deterministic process will always produce the same output given the same initial conditions and event history. Thus, in a deterministic simulation environment, tests are rendered repeatable and yield simulation results that are trustworthy and straightforward to debug. However, game engines are seldom deterministic.**

This paper first reviews and identifies the potential causes of non-deterministic behaviours in game engines. ~~This is then followed by a~~ A case study using CARLA, an open-source autonomous driving simulation environment powered by Unreal Engine, ~~which highlights~~ is then presented to highlight its inherent shortcomings in providing sufficient precision in experimental results. Different configurations and utilisations of the software and hardware are explored to determine an operational domain where the simulation precision is ~~satisfactory~~ sufficiently low i.e.~~variance between repeats~~ variance between repeated executions becomes negligible for development and testing work.

Finally, a method of a general nature is proposed, that can be used to find the domains of permissible variance in game engines simulations for any given system configuration.

ENSURE PAST/PRESENT IS USED CONSISTENTLY AND APPROPRIATELY THROUGHOUT.

***Index Terms*—Autonomous Driving, Autonomous Vehicles, Determinism, Physics Engines, Verification and Validation (V&V), Simulation, Testing**

I. INTRODUCTION

Simulation-based verification of autonomous driving functionality is a promising counterpart to costly on-road testing, that benefits from complete control over (virtual) actors and their environment. Simulated tests aim to provide evidence to developers and regulators of the functional safety of the vehicle or its compliance with commonly agreed upon road conduct [64], national rules [56] and road traffic laws [58] which form a body of safe and legal driving rules, termed assertions, that must not be violated.

Design confidence is gained when the autonomous vehicle (AV) can be shown to comply with these rules, e.g. through assertion checking during simulation. There have been a number of fatalities with AVs, some of which could be attributed to insufficient verification and validation (V&V), e.g. [44]. Simulation environments offer a means to explore the vast parameter space in a safe and efficient manner [28] without the need for millions of miles of costly on-road testing [24]. In particular, simulations can be biased to increase the frequency at which otherwise rare events occur [27]; this includes testing how the AV reacts to unexpected behaviour of the environment [20].

Increasingly, the autonomous vehicle community is adopting game engines as simulation platforms to support the development and testing of vehicle control software. CARLA [7], for instance, is an open-source simulator for autonomous driving that is implemented in the Unreal Engine [62], a real-time 3D creation environment for the gaming and film industry as well as other creative sectors [15].

State-of-the-art game engines provide a convenient option for simulation-based testing. They offer sufficient realism [27] in the physical domain combined with realistic rendering of scenes, potentially suitable for perception stack testing and visual inspection of accidents or near misses. Furthermore, they are easy to setup and run with respect to on-road testing and are simple to control and observe, both with respect to the environment the AV operates in as well as the temporal development of actors [59]. Finally, support for hardware-in-the-loop development or a real-time test-bed for cyber-security testing [23] may also be required. Compared to the vehicle dynamics simulators and traffic-level simulators used by manufacturers [47], game engines offer a simulation solution that meets many of the requirements for the development and functional safety testing of AVs in simulation. However, while game engines are designed primarily for performance to achieve a good user experience, the requirements for AV verification go beyond that and include determinism.

II. PRELIMINARIES

A. Definitions

~~A list of definitions are given here which~~ Several definitions are introduced in this section. These are used in the subsequent

Manuscript received ...; revised ...; accepted.... Date of publication ...; date of current version ....

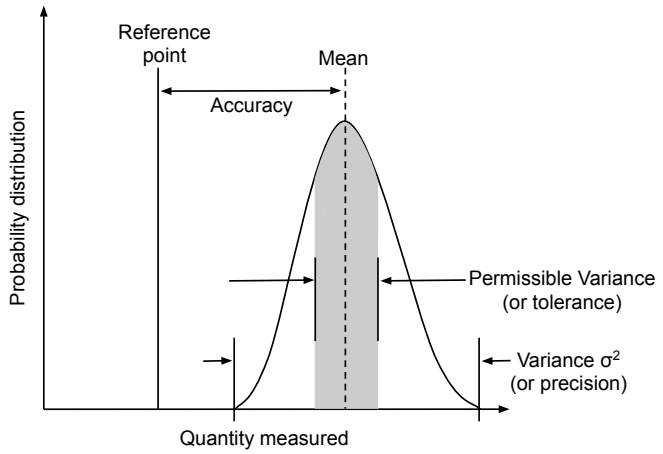
Abanoub Ghobrial and Greg Chance contributed equally to this paper and are both corresponding authors.

This research has in part been funded by the ROBOPILOT and CAPRI projects. Both projects are part-funded by the Centre for Connected and Autonomous Vehicles (CCAV), delivered in partnership with Innovate UK under grant numbers 103703 (CAPRI) and 103288 (ROBOPILOT), respectively.

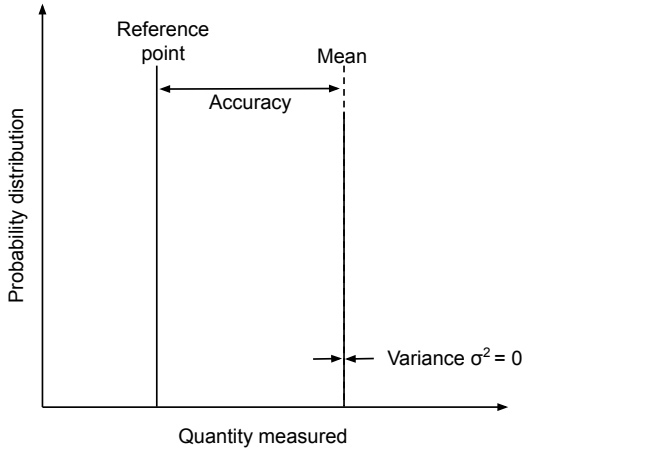
Abanoub Ghobrial (e-mail: abanoub.ghobrial@bristol.ac.uk), Greg Chance (e-mail: greg.chance@bristol.ac.uk), Kevin McAreavey (e-mail: kevin.mcareavey@bristol.ac.uk), and Kerstin Eder (e-mail: kerstin.eder@bristol.ac.uk) are with the Trustworthy Systems Lab, Department of Computer Science, University of Bristol, Merchant Ventures Building, Woodland Road, Bristol, BS8 1UQ, United Kingdom.

Severin Lemaignan (e-mail: severin.lemaignan@brl.ac.uk) and Tony Pipe (e-mail: tony.pipe@brl.ac.uk), are with the Bristol Robotics Lab, T Block, University of the West of England,Frenchay, Coldharbour Ln, Bristol, BS34 8QZ, United Kingdom.

Digital Object Identifier ....



(a) Non-deterministic



(b) Deterministic

Fig. 1: Demonstration of variance, precision, tolerance and determinism

discussion. Refer to Fig. 1 throughout this section.

1) Determinism: Schumann describes determinism as the property of causality given a temporal development of events such that any state is completely determined by prior states [48]. However, in the context of simulation this should be expanded to include not just prior states but also the history of actions taken by all actors. Therefore, a deterministic simulation will always produce the same result given the same history of prior states and actions.

A simulation can be thought of as the ~~generation or production of~~ process of generating or producing experimental data. In the case of a driving simulator, kinematics will describe future states of actors given the current conditions and actions taken, thereby generating new data. If a simulation is deterministic, Fig. 1 (b), then there will be no variation in the generated output data, i.e. all future states are perfectly reproducible from prior states and actions. However, if a simulation is non-deterministic, Fig. 1 (a), then there will be a variation in the generated output data.

2) Variance, Precision & Tolerance: We adopt ~~some terms~~ terminology from the mechanical engineering and statistics ~~domain that best~~ domains to describe when there is variation in the generated output data [3]. *Variance* is used here to define the spread, or distribution, of the generated output data with respect to the mean value. *Precision* is synonymous with *variance* although inversely related mathematically. Therefore, variance can indicate the degree to which a simulation can repeatedly generate the same result when executed ~~with~~ under the same conditions and actions~~taken~~. *Tolerance* is defined as the permissible limit of the variance, or in short the *permissible variance*.

As an analogy, the simulator can be thought of as a manufacturing process ~~, producing dataand to find the value of the precision then the~~ that produces data. To determine the precision of this process, the output must be measured and analysed for differences when the process is repeated. Those differences describe the spread or variance in the process output~~and a~~. A hard limit on the variance can then be defined, Fig. 1~~a~~ (a), beyond which the output fails to meet the required tolerance, e.g.~~rejected for~~ the output is rejected by quality control. Real manufacturing fails to achieve absolute precision~~and hence the~~. Hence, there is a need for tolerances ~~on design specifications to~~ to be specified to account for the variance in real-world manufacturing processes.

If a simulator is deterministic then it will produce results with absolute precision or zero variance, Fig. 1~~b~~ (b), and hence will be within ~~acceptable tolerance~~ tolerance by design. If the simulator is non-deterministic then there will be a measurable, non-zero variance in the output ~~which can be measured~~data.

3) Accuracy: Precision and tolerance should not to be confused with *accuracy*, which describes how closely the mean of the generated output of a process aligns to a known standard or reference value. We therefore define accuracy as the difference between the true value, or reference value, and what has been achieved in the ~~simulation or generation process~~ data generation process or simulation. For a driving simulation the reference value may be the real world that the simulation seeks to emulate, where any divergence from this standard is termed the *reality gap*. ~~In most cases such accuracy will not be possible due to~~ In practice full accuracy will often not be achievable due to modelling and associated computational demands of ~~such an exact replicaand in most casesis unnecessary, where~~ creating and executing exact replica. In most cases, in fact, it is unnecessary and some authors state that 'just the right amount of realism' is required to achieve valid simulation results [27].

4) Simulation Trace: A simulation trace is the output log from the simulator consisting of a time series of all actor positions (x, y, z) in a 3D environment recorded at regular time intervals. This definition could be extended to include other variables. A set of simulation traces derived from the same input and starting state then forms the experimental data on which variance is calculated for a given simulation run.

5) Simulation Variance & Deviation: If the simulator is non-deterministic then how can ~~this~~ the simulation variance be measured? This can be achieved by monitoring ~~any simulated~~ the values of any of the recorded output variables that should be consistent from run to run. ~~Actor path variance, derived from the simulation trace, is chosen over other variables as this distance based metric forms the basis for many downstream verification tests.~~ For example, actor position variance is a distance-based metric that can be derived from simulation traces. The actor position over time, i.e. the actor path, is often used in assertion checking, e.g. to determine whether vehicles keep within lanes or whether minimum distances to other vehicles and road users are being maintained. ~~Therefore the~~ Thus, in the case study presented in this paper, the term *simulation variance*, measured in SI unit m^2, refers to a measure of actor position variance in the simulation with respect to time, assuming fixed actions. Unclear: Why wrt. time, that would then be the actor path, or not? ~~*Deviation*~~ Case study results are presented using deviation (SI unit m)~~is~~, the square root of variance~~(SI unit m^2), which~~, rather than variance, as this is a more intuitive ~~value~~ measure to comprehend when ~~considering interpretation of verification tests.~~ interpreting test results.

6) Scene, Scenario & Situation: We ~~will~~ adopt the terminology defined for automated driving in [59], where *scene* refers to all static objects including the road network, street furniture, environment conditions and a snapshot of any dynamic elements. Dynamic elements are the elements in a scene whose actions or behaviour may change over time; these are considered actors and may include the AV, or *ego vehicle*, other road vehicles, cyclists, pedestrians and traffic signals. The *scenario* is then defined as a temporal development between several scenes which may be specified by specific parameters. A *situation* is defined as the subjective conditions and determinants for behaviour at a particular point in time.

B. When is Determinism needed?

Determinism is a key ~~prerequisite~~ requirement for simulation during AV development and testing. A deterministic simulation environment guarantees that tests are repeatable~~. A deterministic simulator can be considered to have~~, i.e. repeated runs of a test, given the same initial conditions and event history, produce the same output data. Thus, a deterministic simulator has zero *variance*. ~~If the variance is~~

A simulator with non-zero ~~it~~ variance can no longer be considered deterministic~~but this~~. Non-deterministic simulators may be sufficient for certain applications as long as ~~it is~~ their variance is permissible, i.e. *within tolerance*. Therefore, *tolerance* is the acceptable degree of variability between repeated ~~simulation traces~~simulations. When the simulation output is within tolerance, coverage results are stable and, when a test fails, debugging can rely on the test producing the same trace and outcome when repeated. This ensures that software bugs can be found and fixed efficiently, and that simulation results are trustworthy.

If the simulation is non-deterministic~~and has a non-permissible~~, e.g. it has a non-zero variance in, for example, actor positions, ~~this may lead to assessment errors~~then this may, in the best case, lead to intermittent assertion failures, making it difficult to re-produce, understand and remove bugs and rendering verification results unstable. In the worst case, however, bugs that could have been identified in simulation remain undetected, ~~resulting in~~ leading to false confidence in the safety of the AV's control software.

When used for gaming, game engines do not need to be deterministic nor do they ~~even~~ have any requirements on the limits of permissible variance; there are no safety implications from non-determinism in this domain, nor is finding and fixing all the bugs ~~related to non-determinism~~ a high priority for games developers. It could even be argued that simulation variance is a feature that enhances gaming and improves the user experience. However, the situation is very different for AV development and testing. Thus, our main research ~~question is~~questions are: *How can one assess ~~the extent to which~~ whether a simulation environment is deterministic~~or has a permissible level of variance?~~?* and *How can one determine and control the simulation variance?*

In this paper we ~~investigated how the non-determinism of Carla,~~ investigate non-determinism and how it affects simulation results on the example of CARLA, an open-source autonomous driving simulation environment based on the Unreal game engine~~, when used for simulation-based AV verification affects the simulation results~~. In our case study, scenarios between pedestrian and vehicle actors ~~have been analysed to identify conditions that result in non-deterministic simulation output by analysis of~~ are investigated to determine the actor position variance ~~.~~ in the simulation output for repeated simulation runs. ~~By analysing actor position variance we~~ We find that the ~~Carla~~ CARLA simulator is non-deterministic under certain conditions. Say what the observed variance range is, and what we consider permissible for this study. ~~Carla~~ CARLA exhibits a permissible ~~level of~~ variance when system utilisation is restricted to 75% or less and ~~ensuring termination of scenarios~~ the simulation is terminated once a vehicle collision has been detected. Why "vehicle" collision and not just "collision"?

The insights gained from this case study motivated the development of a general step-by-step method for AV developers and verification engineers to determine the simulation variance for a given simulation environment. Knowing the simulation variance will help assess the ~~impact that using~~ suitability of a game engine for AV simulation~~may have on verification tasks.~~. In particular, this can give a better understanding of the effects of non-determinism and to what extent simulation precision may impact on verification results.

This paper is structured as follows. Section III briefly introduces how game engines work before investigating in Section IV the potential sources of non-determinism in game engines. Our case study of simulation variance for a number of scenarios involving pedestrian and vehicle settings using CARLA is presented in Section V. Section VI presents the step-by-step method to assess the suitability of a simulation system for AV verification in general. We conclude in Sec-

tion VII and give an outlook on future work.

III. Background

There are numerous game engines with their associated development environments which could be considered suitable for AV development, e.g. Unreal Engine [62], Unity [61], CryEngine [10]. Specific autonomous driving research tools have been created to abstract and simplify the development environment, some of which are based on existing game engines, e.g. CARLA [7], AirSim [1], Apollo [2], and some have been developed for ~~cloud based~~ cloud-based simulation, e.g. Nvidia Drive Constellation [42].

Investigating the determinism of game engines has not attracted much research interest since performance is more critical for game developers than accurate and repeatable execution. Ensuring software operates deterministically is a non-trivial task~~and catching~~. Catching intermittent failures, or flaky tests [53], in a test suite that cannot be replayed makes the debugging process equally difficult [51]. This section gives an overview of the internal structure of a game engine and what sources or settings in the engine may affect *simulation variance*.

Central to a game engine are the main game logic, the artificial intelligence (AI) component, the audio engine, and the physics and rendering engines. For AV simulation, we focus on the latter two. The game loop is responsible for the interaction between the physics and rendering engines. Fig. 2 depicts a simplified representation of the process flow in a game engine loop, where initialisation, game logic and decommissioning have been removed [60]. A game loop is broken up into three distinct phases: processing the inputs, updating the game world (Physics Engine), and generating outputs (Rendering) [19].

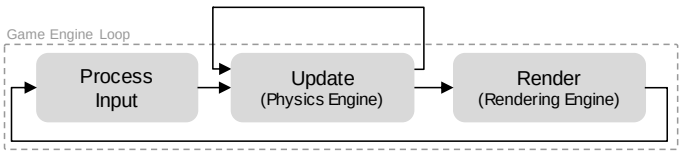


Fig. 2: Game engine loop block diagram [43].

The game loop cycle starts with initialising the scene and actors. Input events from the User or AI are then processed followed by a physics cycle which may repeat more than once per rendered frame if the physics time step, dt, is less than the render update rate. This is illustrated by the loop in the physics update in Fig. 2. The render update will process frames as fast as the computational processing will allow up to the maximum monitor refresh rate [31]. When the frame is rendered the game loop cycle returns to processing inputs. An intuitive and more detailed description of the interplay between the physics and render cycles is given in [4].

The physics engine operates according to a time step, dt. The shorter this time step is, the smoother the interpretation of the physical dynamics will be. To use a fixed physics time step, the user's display refresh rate needs to be known in advance. This requires an update loop to take less than one render tick (one frame of real world time). Given the range of different hardware capabilities, a variable delta time is often implemented for game playing, taking the previous frame time as the next dt. However, variable dt can lead to different outcomes in repeated tests and in some cases unrealistic physical representations [16]. Semi-fixed or limited frame rates ensure dt does not exceed some user-defined limit to meet a minimum standard of physical representation but allow computational headroom for slower hardware. Some engines provide sub-stepping which processes multiple physics calculations per frame at a greater CPU cost, e.g. Unreal Engine [54]. If the engine tries to render between physics updates, *residual lag* can occur, which may result in frames arriving with a delay to the simulated physics. Thus, extrapolation between frames may need to be performed to smooth transition between scenes. Note that both residual lag and extrapolation could affect perception stack testing. In exceptional cases, where computational resources are scarce, the fixed time step can be greater than the time between render ticks and the simulation will exhibit lag between input commands and rendered states, resulting in unsynchronised and unrealistic behaviour as can be experienced when games are executed on platforms not intended for gaming.

Considering the objectives for gaming and comparing them to these for AV development and testing, there are fundamental differences. Providing game players with a responsive real-time experience is often achieved at the cost of simulation accuracy and precision. The gamer neither ~~wishes~~ needs a faithful representation of reality, i.e. the gamer will accept low *accuracy*, nor do they require for repeated actions to result in the same outcome to within a particular tolerance~~, i. e. can be low *precision*~~. In contrast, the precision required for AV development and testing is ~~much higher~~very high, especially for perception stack testing, but also~~for~~, more generally, for obtaining reliable verification results and for visual inspection of accidents or near misses. If necessary, it is acceptable to achieve the required ~~level of tolerance on the precision~~ tolerance at the cost of real-time performance. For example, for perception stack testing the sensors need to get input that is repeatable so that if any software bugs are found they can be re-played and the issue resolved. This may only be realisable by slowing down the rendering to enable more extensive physics calculations.

IV. Potential Sources of Non-Determinism

The following review discusses the potential sources of non-determinism that were found in the literature or found as part of our investigation into game engines. We have examined hardware- as well as software-borne sources of non-determinism that occur at different layers of abstraction. A good analysis of potential sources is given by Strandberg et al. [53], although the AV simulation domain introduces it's own unique challenges that were not considered in that paper.

A. Floating-Point Arithmetic

Floating-point representation of real numbers is limited by the fixed bit width available in the hardware, resulting in the

finite precision with which numbers can be represented in a computation. Thus, the results of arithmetic operations must fit into the given bit width, which is achieved through rounding to the nearest representable value. This gives rise to rounding errors [36, 18].

PROBLEMATIC: Even operations such as the average of an array can result in issues with overflows ~~, underflows and~~ and underflows (THESE CAN HAPPEN WITH INTEGERS TOO) and, non-associative addition [25] (ONLY WHEN THE EXEC ORDER IS CHANGED) which may result in non-deterministic behaviour. ~~Calculation resultsmay differ~~

As a consequence, floating-point arithmetic is not associative. Thus, for arithmetic expressions beyond two operands, such as $x+y+z$, executing $x+(y+z)$ rather than $(x+y)+z$, can produce different results [25]. This is because rounding is being performed at different stages in the computation, depending on the respective execution order. Execution order changes can occur due to the ~~nature of non-associative floating point arithmetic if there is a change in compiler , a change in the execution order or even~~ use of a different compiler or as a consequence of parallelisation, e.g. within the runtime environment or at the hardware level. Furthermore, floating-point calculation results may differ if the execution is performed on a GPU rather than a CPU which may have different register widths [66]. ~~Some authors even suggest avoiding the use of floating point entirely for assertion testing due to imprecision and non-deterministic behaviour [30].~~

In the context of AV simulation, such rounding errors could result in accuracy issues of, for example, actor positions within the environment, leading to falsely satisfied or failed assertions. Thus, one could ~~jump to the conclusion~~ conclude that floating-point ~~errors cause non-deterministic behaviour~~arithmetic causes non-determinism, resulting in loss of repeatability. In fact, some authors suggest avoiding the use of floating-point representation entirely for assertion testing due to imprecision and non-deterministic behaviour [30].

~~However, these precision issues~~ In contrast, we argue that the precision issues related to floating-point operations are better described as *incorrectness* ~~, rather than them being a source of~~ that is in fact repeatable; they do not cause non-determinism ~~.~~ per se. Should this be IMPRECISION rather than incorrectness? It is normally reasonable to assume that the processors on which game engines run are ~~designed to be~~ deterministic with respect to floating-point arithmetic. Thus, ~~when a rounding error occurs, then~~ given the same operands and sequence of operations, the same processor should always produce the same ~~*incorrect* result given the same operands.~~ bit pattern. So, even if the result of a floating-point operation is incorrect due to floating-point rounding errors, it should always be equally *incorrect* ~~as long as the implementation conforms~~ for implementations that conform to the IEEE floating-point standard [21].

~~Beyond hardware limitations,~~ However, as illustrated above on the example of non-associative addition, a change of execution order can lead to sequences of floating point operations producing different results. Thus, any uncontrolled options to modify execution order, e.g. at the compiler level, within the runtime environment or at the hardware level, can cause simulation results to differ between runs, resulting in non-zero variance and loss of determinism.

Beyond that, aggressive optimisations by the compiler can also introduce incorrectness in floating-point arithmetic [39]. However, the same executable should still return the same output for identical input. ~~Therefore, it can be concluded that~~

In conclusion, floating-point arithmetic does not ~~contribute to~~ cause non-zero *simulation variance* for repeated ~~tests~~ simulation runs when using the same executable on the same hardware with exactly the same configuration and execution order.

B. Scheduling, Concurrency and Parallelisation

Runtime scheduling is a resource management method for sharing computational resources between tasks of different or equal priority where tasks are executed dependent on the scheduler policy of the operating system. A scheduler policy may be optimised in many ways such as for task throughput, deadline delivery or minimum latency [29]. ~~Changing~~ In principle, changing the scheduling policy and thread priorities may increase simulation variance.

It is, however, unlikely that changes to thread priorities or scheduling policy would occur during repeated controlled tests for the same hardware and operating system configuration. Given that hardware is considered deterministic and if the thread execution order does not change between tests, then for the same hardware and operating system configuration the same output should be given.

However, if some aspects of the game loop are multi-threaded [65], then, even with a clear thread scheduling order, any background process may interrupt the otherwise deterministic sequence of events. This may, for example, alter the number of physics calculations that can be performed within the game loop and hence result in simulation variance. Using multiple threads has been found to affect initialisation ordering for training machine learning models which can lead to unpredictable ordering of training data and non-deterministic behaviour [49, 5]. Interference may also occur when a scheduler simply randomly selects from a set of threads with equal priority, resulting in variation of the thread execution order.

Similar to thread scheduling, scheduling at the hardware level on a multi-core system determines on which processor core to execute processes. This may be decided based on factors such as throughput, latency or CPU utilisation. If due to the CPU utilisation policy the same single-threaded script executes multiple times across different physical cores of the same CPU type, then execution should still produce the same output. This is because the processor cores are identical and any impurities across the bulk silicon and minor perturbations in the semiconductor processing across the chip that may exist should have been accommodated for in design and manufacturing tolerances. However, scheduling multiple processes across several processing cores, where the number of cores is smaller than the number of processes, can result in

variation of the execution order and cause simulation variance unless explicitly constrained or statically allocated prior to execution.

Indeed, the developers of the ~~de-bugging program~~ debugging program ~~RRrr~~ [46] took significant steps to ensure deterministic behaviour of their program by executing or context-switching all processes to a single core, which avoids data races as single threads cannot concurrently access shared memory. This allowed control over the scheduling and execution order of threads promoting deterministic behaviour by design [51].

Likewise, simulation variance may be observed for game engines that use GPU parallelisation to improve performance by offloading time-critical calculations to several dedicated computing resources simultaneously. While this would be faster than a serial execution, the order of execution arising from program-level concurrency is often not guaranteed.

Overall, scheduling, concurrency and parallelisation may be reasons for *simulation variance*.

C. ~~NUMA~~ Non-Uniform Memory Access (NUMA)

For a repeated test that operates over a number of cores based on a CPU scheduling policy, memory access time may vary depending on the physical memory location relative to the processor. Typically a core can access its own memory with lower latency than that of another core resulting in lower inter-processor data transfer cost [38]. Changes in latency between repeated tests may, in the worst case, cause the game engine to operate non-deterministically if tasks are processed out of sequence using equal priority scheduling, or, perhaps, simply with an increased data transfer cost, i.e. slower. By binding a process to a specific core for the duration of its execution, the variations in data transfer time can be minimised.

D. Error Correcting Code (ECC) Memory

ECC Memory is used ubiquitously in commercial simulation facilities and servers to detect and correct single bit errors in DRAM memory [12]. Single bit errors may occur due to malfunctioning hardware, ionising radiation (background cosmic or environmental sources) or from electromagnetic radiation [13]. If single bit errors go uncorrected then subsequent computational processing will produce incorrect results, potentially giving rise to non-determinism due to the probabilistic nature of such errors occurring. Estimating the rate of error is difficult and dependent on hardware, environment and computer cycles [33].

Any simulation hardware not using ECC memory that runs for 1000's of hours, typical in AV verification, is likely to incur significant CPU hours and is therefore subject to increased exposure to these errors. To counter this, commercial HPC and simulation facilities typically employ ECC memory as standard.

E. Game Engine Setup

The type and version of the engine code executed should be considered, paying attention to the control of pseudo-random numbers, fixed physics calculation steps, (dt), fixed actor navigation mesh, deterministic ego vehicle controllers and engine texture loading rates especially for perception stack testing. For example, in Unreal Editor the *unit* [52] command can be used to monitor performance metrics such as *Frame* which reports the total time spent generating one frame, *Game* for game loop execution time and *Draw* for render thread time. With respect to perception stack testing, weather and lighting conditions in the game engine should be controlled as well as any other dynamic elements to the simulation environment, e.g. reflections from surface water, ensuring textures are not randomly generated.

F. Actor Navigation

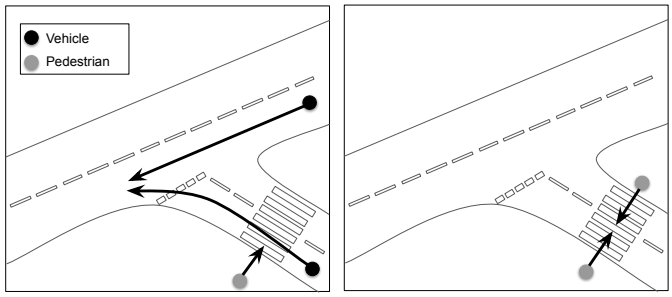
There is existing evidence to suggest actor navigation, typically pedestrians, could be the cause of non-deterministic simulation behaviour [8]. Unreal Engine is the underlying framework for the CARLA simulator which uses the A* algorithm for actor navigation [63]. The A* algorithm [26] will give deterministic outcomes as long as the environment is deterministic [34, 55].

A navigation mesh is a fixed area of the environment where actors are free to navigate. Navigation meshes are used by the path planning algorithm to find the shortest distance between navigable points in the environment and could be considered as a potential source of non-determinism if the navigation mesh is not a fixed entity. However, the environment management in CARLA implements fixed binary files for navigation meshes that are linked to each road scene or driving environment~~and~~. These cannot be unknowingly modified and therefore can be considered fixed.

Unless other sources that can alter the fixed environment exist, the A* algorithm should give deterministic results. Thus, actor navigation should not be considered a source of non-determinism.

G. Summary

We have investigated the potential sources of non-determinism affecting game engines and explored the impact they may have on simulation variance. Memory checking not withstanding, errors associated with the lack of ECC are likely to be minimal unless there is significant background radiation or 1000's of hours of computation are expected. To ensure precise simulation outcomes the physics setting, dt, must be fixed, along with any actor navigation meshes, ~~random number seeds~~seeds for random number generation, game engine setup and simulation specific parameters. ~~NUMA~~ Non-uniform memory access (NUMA) should only affect interprocessor data transfer cost and, without control measures, will only make the computation cycle longer. Relative access times between different caches are likely to be small although may have a more pronounced impact on high throughput systems, e.g. HPC. Not sure this is correct, see the NUMA section above. Why "only"? If one computation out of many takes longer, then their exec order can be affected, or not? Basic thread scheduling should not affect the simulation's determinism unless changing scheduling policy, operating system or migrating between machines with different setups. However,



(a) (b)

Fig. 3: Schematic of test scenarios for (a) Tests 1-4, (b) Tests 5-6. Descriptions are given in Table I.

should new and unexpected threads start during the simulation, then the interruption to execution order or additional resource demand may affect timing of subsequent steps, thus reducing the number of physics updates within a game loop. Likewise, uncontrolled allocation of hardware resources such as CPUs or GPUs can potentially give rise to non-determinism.

V. Case Study of Simulation Variance

We present an empirical investigation into using game engines for simulation-based verification of autonomous vehicles with a focus on characterising sources of non-determinism in order to understand the impact they have on simulation variance. Gao et al. [17] took a similar approach investigating Java applications, where a set of sources of non-determinism (termed factors) were shown to impact on repeatability of testing. Ultimately, our objective is to control ~~non-determinisim~~ non-determinism to minimise simulation variance.

We first describe the context, scene and scenario of interest before discussing and defining a tolerance for what is considered an acceptable simulation variance in this context. and then..? Please cover the entire section - consider splitting setup from evaluation.

A. Context, Scene and Scenario

This case study draws on a setup used to verify an urban mobility and transport solution, where the primary verification objective is to test the behaviour of an ego vehicle in an urban environment at a T-junction in response to pedestrians and other vehicles. Thus, the scene for our investigation is the T-junction ~~(T-intersection) and the scenarios are~~ with the scenarios as shown in Figure 3.

This scene was used to create a number of scenarios involving pedestrians and vehicles in order to identify any changes in the actor ~~trajectories~~ paths over repeated tests executed under a variety of systematically designed conditions and hence study any simulation variance. The vehicles and pedestrians were given trajectories, via pre-defined waypoints, that would result in either colliding with or avoiding other actors.

B. Tolerable Simulation Variance

To achieve stable verification results over repeated test runs, the simulated actor states must be precise to a specific tolerance. Deterministic behaviour would result in zero variance of the simulated actor states but if this cannot be achieved then what is permissible? ~~This~~ The tolerance must be appropriate to allow accurate assertion checking and coverage collection in the simulation environment, but not so small as to fail with minor computational perturbations. WHAT WOULD FAIL? Thus, a tolerance must be defined to reflect the precision at which repeatability of simulation execution is required.

For this case study a tolerance on actor position of 1m would be insufficient when considering the spacial resolution required to distinguish between a collision and a near-miss event. A very small value, e.g. $1 \times 10^{-12}m$, may be overly-sensitive to minor computational perturbations and generate false positives. Therefore, for this case study and~~more broadly~~ , more broadly, for our general verification requirements, a tolerance of 1cm has been selected. Thus any variance of less than 1cm is permissible. To put this another way, we can accept a precision with a tolerance of $\leq \pm 1cm$.

~~Case~~ In the following, case study results are shown in terms of the maximum deviation, $\max \sigma$, from the mean actor path over the entire simulation history where any value higher than the specified tolerance is considered non-permissible.

C. Actor Collisions

Previous investigations into the Unreal Engine indicated that collisions between dynamic actors Why DYNAMIC? and solid objects, termed *blocking physics bodies* in Unreal Engine documentation [9], can lead to high simulation variance~~, [22].~~ [22]. NOT ACCESSIBLE Collisions and the subsequent physics calculations that are processed, termed *event hit* callback in Unreal Engine, were ~~seen~~ identified as potentially key aspects to the investigation into simulation variance.

The tests used for the case study are listed in Table I~~and were chosen to~~ . They cover a range of interactions between actor types. Tests 1 & 2 involve ~~2~~ two vehicles meeting at a junction where they either do not collide (~~test~~ Test 1) and ~~when they do tiggering~~ where they do collide (Test 2), thereby triggering an *event hit* callback in the game engine~~(test 2)~~. In both cases the trajectories of the vehicles are hard-coded to follow a set of waypoints spaced at 0.1m intervals using a PID controller. ~~In test~~ In Test 3 a mixture of different actor types ~~are~~ is introduced where two vehicles drive without collision and a pedestrian walks across the road at a crossing point. ~~In test~~ In Test 4 this pedestrian collides with one of the vehicles at the crossing, triggering ~~a~~ an *event hit* callback, see Fig 3a. Similar to vehicles, pedestrians navigate via a set of regularly spaced waypoints at 0.1m intervals using the A* algorithm which is the default controller for the CARLA pedestrian actors. Tests 5 & 6 involve only pedestrians that either do not collide (~~test~~ Test 5) and that do collide (~~test~~ Test 6), see Fig. 3b.

D. ~~Experiment Description~~Evaluation Metric

For each test the position of each actor ~~was~~ is logged at $0.1s$ intervals providing a trace of that actor's ~~trajectory~~ path

with respect to simulation time. ~~Repeated traces~~ The logs from repeated tests are compared at the same time index t for each actor a to provide a value for the variance of actor a at time index t, $\sigma_a^2(t)$, at each point in time. This gives a variance function over time for each actor.

The logs from repeated tests are sourced to establish a value for the variance associated with each actor, a, at each time point t, giving a variance function over time for each actor, $\sigma_a^2(t)$.~~Herein the~~

Instead of using ..., PLEASE COMPLETE herein the results are given in terms of the deviation, $\sigma_a(t)$, which indicates the dispersion of the actor position data relative to the mean and is helpfully in the same units as actor position, i.e. metres (m), for ease of interpretation.

The maximum variance over the entire set of n ~~simulations is therefore~~ repeated simulations, i.e. the overall observed worst case, is defined as the largest variance of any actor at any time in ~~the simulation,~~ any of the simulation runs, as given in Equation 1.

$$\max_{a,t} \sigma_a^2(t), \tag{1}$$

~~and therefore the~~ The maximum deviation is ~~simply~~ the square root of the maximum variance The absolute value of the square root? and herein referred to as $\max \sigma$ for brevity. So, "a" is any actor or a specific actor? Likewise, "t" is any time point or a specific time point? NOTE that any actor any time means a in A and t in T, where A is the set of actors and T is the set of all time points.

The maximum deviation, $\max \sigma$, ~~was analysed against~~ can be analysed for the different scenarios and settings that were identified as potential sources of non-determinism, and compared against the limit of *permissible variance* to indicate if the simulation ~~was reliable~~ is sufficiently accurate for verification purposes.

E. Internal Settings

Within Unreal Engine there are numerous internal settings relating to the movement and interaction of physical bodies in the simulation. Settings can be adjusted to alter how actors interact and path plan via the navigation mesh of the environment, e.g. *Contact Offset* and *Navmesh Voxel Size*, or can be changed to improve the fidelity of physics calculations between game update steps, e.g. *Physics Sub-Stepping* and *Max Physics Delta Time*. Other options such as *Enable Enhanced Determinism* were investigated along with running the engine from the command line with options for more deterministic behaviour `-deterministic`, ~~floating point~~ floating-point control `/fp:strict` and headless mode `-nullrhi` along with running the test as a packaged release by building and cooking [45]. An initial study into the Unreal Engine using a pedestrian and a moving block was used to investigate simulation variance against these settings. The results were compared to a baseline of the default engine settings. However, none of these options improved simulation variance significantly and all internal setting were set restored to the default values. Details on this previous investigation can be found on the Trustworthy Systems GitHub [22]. NOT ACCESSIBLE

F. Initial Results and Discussion

This is for unrestricted, so this does not come as a surprise later.

G. External Settings

After discovering that internal engine settings could not achieve a simulation variance that met the required *tolerance*, WHEN HAS THIS BEEN REPORTED? Should there be an "Interim Results" section? settings external to the game engine were explored. Game engines, and simulation hardware more generally, will utilise available system resources such as central and graphical processing units (CPU and GPU respectively), to perform physics calculations and run the simulation environment. For high performance simulations the demand on these resources may be a significant fraction of the total available and an initial hypothesis was that as this ratio tended toward one, there would be an increase in *simulation variance*. This hypothesis was supported by our initial work performed on the unreal engine [22] and was explored more fully in this work using the CARLA platform. From a reviewer's point of view this is a mysterious initial study that has been referred to often but nothing has been presented. Either take out or include, at least wrt. the overall observations, such as the support for this hypothesis.

To replicate in a controlled manner the high computational loads that may be anticipated for high performance simulations, software that artificially utilises resources on both the

Test	Actors	Collision	Collision Type	n	$\max \sigma$ (m) (unrestricted)	$\max \sigma$ (m) (restricted)
1	Two vehicles	No	N/A	1000	0.03	7.0×10^{-3}
2	Two vehicles	Yes	Vehicle and Vehicle	1000	0.31	9.8×10^{-3}
3	Two vehicles and a pedestrian	No	N/A	1000	0.07	5.2×10^{-4}
4	Two vehicles and a pedestrian	Yes	Vehicle and Pedestrian	1000	0.59	1.5×10^{-12}
5	Two pedestrians	No	N/A	1000	5.6×10^{-13}	5.6×10^{-13}
6	Two pedestrians	Yes	Pedestrian and Pedestrian	1000	5.6×10^{-13}	5.6×10^{-13}

TABLE I: A description of the test scenarios showing the test number, the actors included, ~~if~~ whether a collision occurred and if so then between which actors. ~~Where n is~~, the number of repeats is set to 1000 and $\max \sigma$ is the maximum simulation deviation. The term *unrestricted* refers to an unrestricted account of the results including results of any resource utilisation. To understand the impact of collisions and high resource utilisation, the *restricted* column shows a subset of the results where post-collision data and experiments above 75% resource utilisation have been removed.

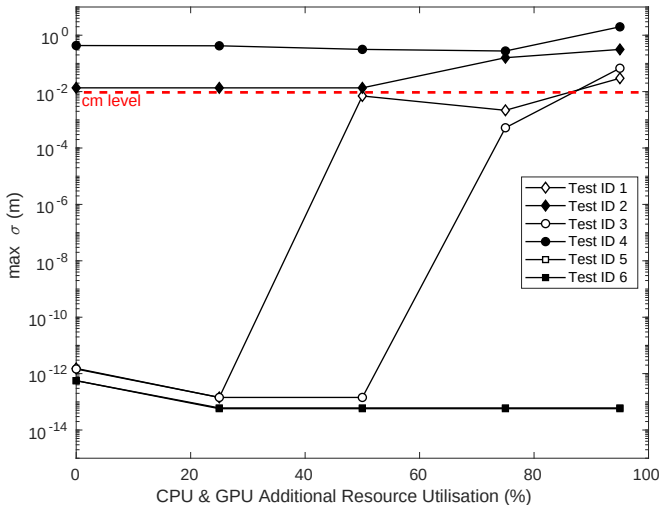


Fig. 4: Summary of results showing maximum deviation for each scenario against different resource utilisation levels. Tests 5 and 6 overlap having almost identical results.

CPU and GPU were executed alongside the simulation. Resource utilisation was artificially increased for both CPU and GPU devices to include a range of values from ~~0~~-950 to 95% (see Section VI) using reported values of the system monitors `htop` and `nvidia-smi` respectively. Resource utilisation figures reported here should be considered approximate values. Somewhere the experimental setup needs to clarify that one set was executed "unrestricted" and another "restricted" - say what this means and why this was done.

Practitioners should also be aware that many libraries for calculating variance itself may require attention to get precise results. For example the `numpy` method of variance is sensitive to the input precision and will return an incorrect answer if the wrong parameters are set [41]. In `matlab`, the calculation of variance may switch from single thread to a multi-threaded execution not obviously apparent to the user when the input data size becomes large enough, opening up the potential for ~~concurrency~~ concurrency-induced imprecision [14].

H. Experimental System Configuration and ~~Screening~~Pre-Screening

~~These tests~~ The experiments were carried out on an Alienware Area 51 R5 with an i9 9960X processor with 64GB non-ECC DDR4 RAM at 2933MHz with an NVIDIA GeForce RTX 2080 GPU with 8GB GDDR6 RAM at 1515 MHz. ~~Operating~~ The operating systems was Linux Ubuntu 18.04.4 LTS. ~~To ensure reliability of results, each test was repeated 1000 times.~~ Tests were carried out in CARLA (v0.9.6 and Unreal Engine v4.22) using synchronous mode with a fixed dt of 0.05s. Each test was repeated 1000 times. Justify why 1000 - see my later comment in the Method section. A detailed guide for reproducing ~~these~~ the experiments along with the scripts used are provided on github[1]. To eliminate some of the potential sources of non-determinism outlined in Section IV a series of screening tests and analyses were performed on our system. These were:

- System memory: `memtest86` [32] full suite of tests ran, all passed.
- Graphical memory: `cuda_memtest` [11], [50].
- Non-uniform memory access (NUMA): `numactl` [40] was used to fix the simulator and test script to single cores, but only a minor (2%) improvement in simulation variance was observed~~and therefore~~. Therefore this was not used for subsequent testing. But if we aim to min the sources of non-det, then that should be considered/used?

I. Final Results and Discussion

Consider spitting to introduce the first part of the table, which motivates the 2nd lot of experiments under load restrictions. I'd find this more logical.

A summary of all the results ~~are shown~~ is shown in Table I~~in~~. In the column $\max\sigma$ (unrestricted) ~~, where~~ the value reported is the maximum deviation across all resource utilisation levels, i.e. the worst case for a given scenario. From these results it is clear that scenarios with only pedestrian actors (~~tests 5-6~~Tests 5 & 6) display results within tolerance over all resource utilisation levels with or without a collision where $\max\sigma$ is 5.6×10^{-13} or 0.56pm. However, all other scenarios involving vehicles or a mixture of actor types ~~breach the tolerance that has been set~~do not meet the required tolerance, with some deviation in actor path as large as 59cm. Clearly~~a deviation in results of~~ , such a large ~~amount is unacceptable if simulation is to be used as a reliable~~ deviation cannot be acceptable for simulation to be considered a credible verification tool.

Resource utilisation ~~level~~ was found to have a significant impact on *simulation variance*. Figure 4 shows $\max\sigma$ against the artificially increased resource utilisation level, where the x-axis indicates the approximate percentage of resource utilisation (for CPU & GPU). In this figure, ~~anything~~ any $\max\sigma$ above the 1cm level (indicated by a dashed line) is considered non-permissible according to our ~~specific tolerancelevel informed from our verification requirements~~specified tolerance. Note that the non-permissible results in Figure 4 (all those above the dashed line) are the worst case account of the situation, as per Equation 1, as the maximum variance is taken over the entire simulation period.

A general pattern in the results indicates that ~~;~~some scenarios consistently fail to produce results within ~~permissible levels of variance at any resource utilisation level (~~tolerance, irrespective of resource utilisation (cf. Fig. 4 Test ~~ID~~ 2 & 4 are above the dashed ~~cm level),~~ 1cm line), while some are consistently within tolerance (~~Test ID~~ cf. Fig. 4 Test 5 & 6 both are with pedestrians only), and ~~that~~ some cases only fail to meet the tolerance requirement ~~when at higher~~ at higher resource utilisation levels, ~~breaching the tolerance~~ i.e. above 75% resource utilisation (~~Test ID~~ cf. Fig. 4 Test 1 & 3). ??? about 85% based on the figure - or not???

~~However, the non-permissible results in Figure 4 (all those above the dashed line) are the worst case account of the~~

[1]https://github.com/TSL-UOB/CAV-Determinism/tree/master/CARLA_Tests_setup_guide

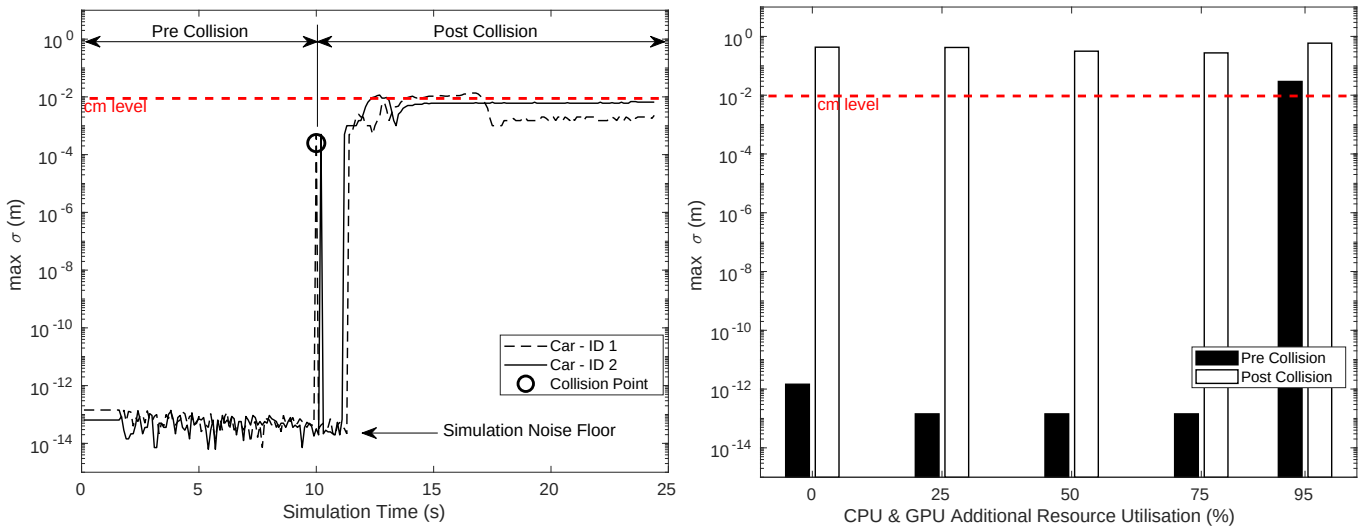


Fig. 5: Vehicle to vehicle collision (~~test~~ Test 2) showing (a) maximum deviation against simulation time for 25% resource utilisation and (b) maximum deviation pre- and post-collision against resource utilisation. The simulation noise floor is shown in (a) which is the empirical lower limit of deviation for the hardware reported in this study.

~~situation, as per Equation 1, as the maximum variance is taken over the entire simulation period.~~

Examining specifically the results from ~~tests~~ Tests 2 & 4 as a function of simulation time reveals further information about the simulation variance before and after an actor collision. Fig. 5a shows this examination for vehicle to vehicle collisions (~~test~~ Test 2), where $\max\sigma$ switches from permissible prior to the vehicle collision to non-permissible ~~post-collision~~post collision. ~~This~~ The pattern of permissible results prior to collision and non-permissible ~~post-collision~~ post collision is maintained up to a resource utilisation level of approximately 75% Why 75%?, see Fig. 5b. This time series examination was repeated for vehicle to pedestrian collisions (~~test~~ Test 4) and the results are shown in Fig. 6a. Similarly to vehicle-to-vehicle collisions, the variation of $\max\sigma$ for vehicle to pedestrian collisions indicates permissible pre-collision behaviour with up to 75% resource utilisation, see Fig. 6b. This is a key finding~~of this work~~; it suggests that verification engineers should consider terminating tests at the point of a collision, as any post-collision results will be non-permissible.

The second key finding of this work is illustrated in Fig. 6a. In this scenario (~~test~~ Test 4), there is a collision between a vehicle (Car ID 2, solid line) and a pedestrian (Ped ID 3, dot dash line) which occurs at a simulation time of approximately 6s and a second vehicle actor (Car ID 1, dashed line), which is *not involved in the collision*. There are three observations~~here~~; firstly that the vehicle directly involved in the collision (Car ID 2) displays high simulation variance immediately after the collision. Secondly, that the maximum deviation of the pedestrian involved in the collision (Ped ID 3) is at a tolerable level throughout the test[2]. Thirdly, we observed a delayed effect on Car ID 1 showing high simulation variance with a 5s delay *even though this vehicle was not involved in the collision*. This final point should be of particular concern to verification engineers~~and research practitioners~~, developers and researchers in the field as it implies that *any collision between actors can affect the simulation variance of the entire actor population* and could potentially result in erroneous ~~assertion checking.~~ simulation results.

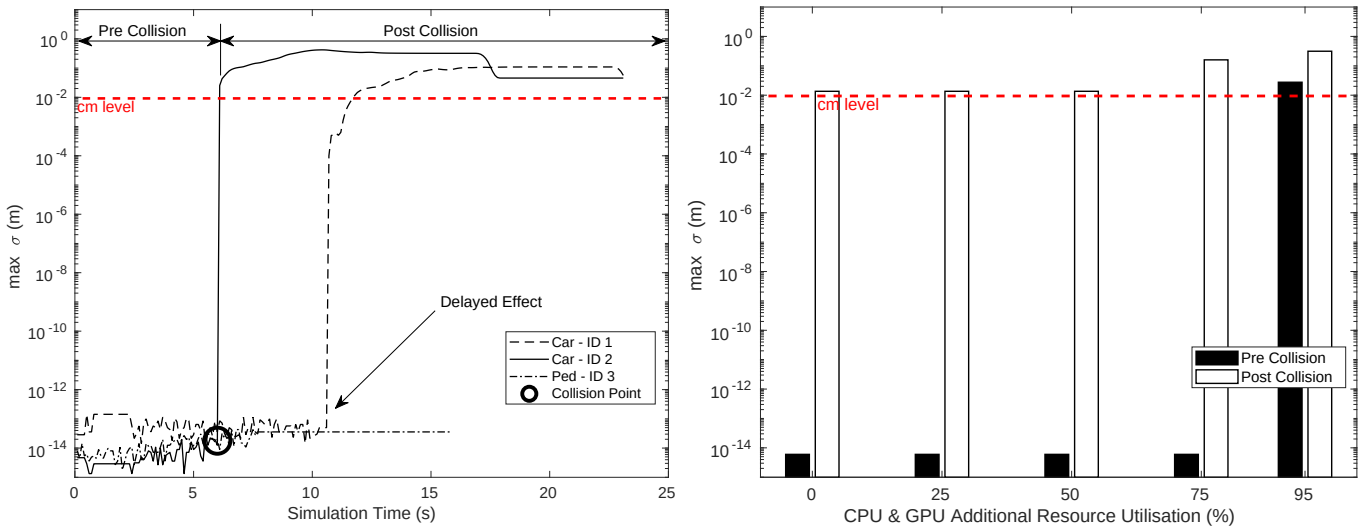
~~The~~ To conclude, the main findings of this ~~work~~ case study suggest a working practice that would minimise the factors that give rise to the non-deterministic effects observed in this investigation. By limiting simulation results to pre-collision data and ensuring resource utilisation levels do not exceed 75%, the permissible variance of 1cm is achievable as shown in the *restricted* column in Table~~.~~ I. By applying this set of restrictions upon the simulation the maximum observed deviation across all experiments was 0.98cm which is within the target tolerance we set out to achieve. Practitioners may wish to set a stricter resource utilisation level, such as less than 50% to further reduce the potential dispersion of results if this is required for their chosen application.

J. Process Scheduling Priority

An investigation into the ~~response~~ impact of process scheduling on the simulation variance was undertaken say why. The experiment was repeated ($n = 1000$) using ~~test~~ Test 1 but altering the process scheduling priority using the program `NICE`.[3] Setting a higher priority for the simulator process with respect to the resource utilisation processes, it was possible to determine if scheduling could account for the increased simulation variance when the system is under high resource utilisation. To give a process a high priority a negative NICE value is set with the lowest being -20. To decrease the

[2]However, please note that in CARLA the pedestrian object is destroyed post-collision hence the flat line from $t = 6$s onwards.

[3]http://manpages.ubuntu.com/manpages/bionic/man1/nice.1.html



(a) (b)

Fig. 6: Vehicle to pedestrian collision (~~test~~ Test 4) showing (a) maximum deviation against simulation time for 25% resource utilisation and (b) maximum deviation pre- and post-collision for different resource utilisation levels.

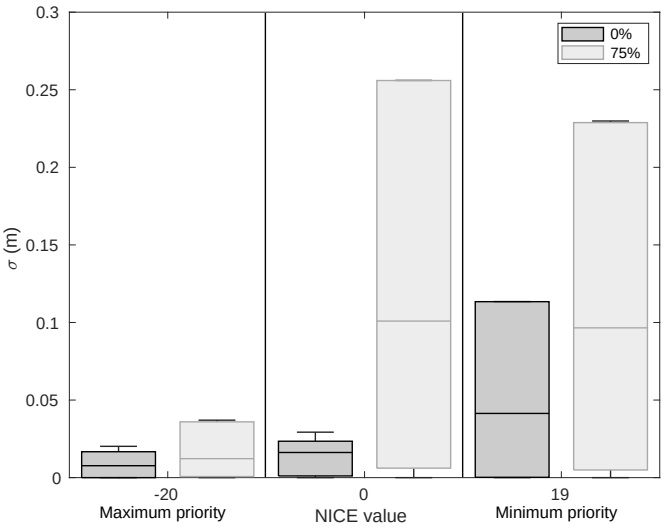


Fig. 7: ~~Summary~~ Variance range of ~~investigating the effect of~~ three NICE priority ~~setting~~ settings for ~~0% and 75% additonal~~ additional CPU & GPU resource utilisation of 0% and 75%.

priority a positive value is set, up to +19. The default NICE value is 0.

The results are presented in Fig. 7 What is on the y axis? where the box denotes the inter-quartile range of???, non-outlier limits by the whiskers and a horizontal bar for the median. The figure shows that decreasing the priority of the simulator process (right hand side of the plot) has little effect on simulation variance when compared to a default NICE value of 0 (central ~~plot bars~~ bars in the plot). Increasing priority (left hand side of ~~plot) reduced~~ the plot) significantly reduced the variance for the 75% resource utilisation ~~but this did~~ experiment, but this does not account for all the difference in the observed results. This ~~is demonstrated~~ can be seen in the maximum priority setting where the bars in the plot are not equal, indicating an additional contribution to variance not accounted for by the NICE scheduling. ~~This unaccounted~~ The remaining difference in the variance between the two resource utilisation levels may be due to the lack of absolute control that `NICE` has over ~~the~~ process scheduling.[4]

K. Investigation Summary

Factor into overall conclusion of the case study

This empirical investigation has highlighted the shortcomings of using a games engine for simulation based verification and advice for best working practice. However, these results are specific to the hardware and software used in the study and may not be transferable to other systems directly. ~~Therefore we now present~~ Therefore we have derived a general methodology that practitioners can follow to find the *operational domains of permissible variance* for ~~any hardware configuration.~~ a game-engine-based simulation environment. This methodology is presented in the next section.

VI. ~~Methodology~~Method to determine the Variance of a Simulationfor what???

In this section a method for assessing the actor path variance in AV simulation environments that are based on game engines is presented in the form of a work flow, see Fig. 8. This method can be used to determine the operational domains of permissible variance of a simulation environment. MAKES NO SENSE: In this section a method for determining the variance of simulated actor path trajectories (KIE: a trajectory is a path), termed simulation variance, ~~termed *simulation variance*,~~ and resolving the operational domains of permissible variance of the simulation presented here as a work flow, see Fig. 8. In additiona ~~number of~~ , recommendations and best ~~practices are suggested that other~~

[4]https://askubuntu.com/questions/656771/process-niceness-vs-priority

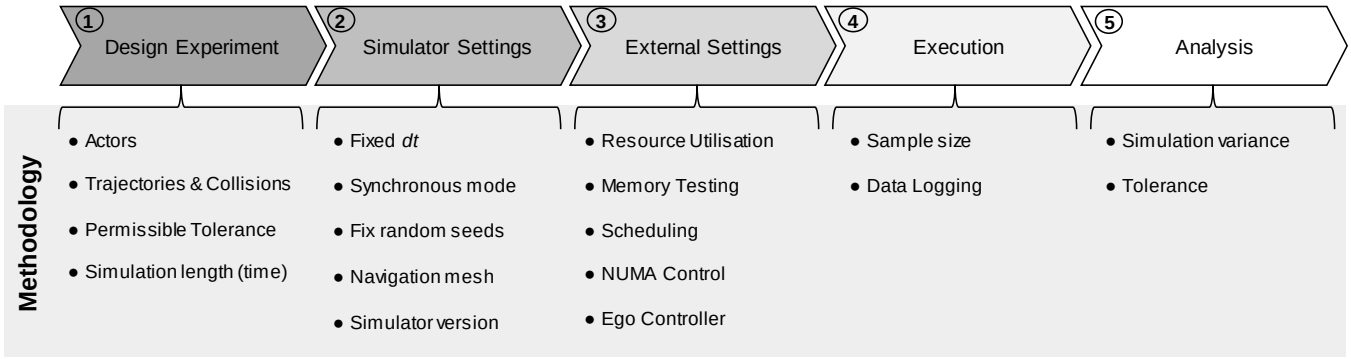


Fig. 8: ~~Flow diagram~~ Stages of the ~~methodology~~ method proposed to determine the ~~simulation~~ variance of a simulation. PLEASE UPDATE TO MATCH TEXT

~~practitioners can follow to minimise simulation variance .~~ practice guidelines for minimising simulation variance are suggested. HIGHLIGHT THESE IN THE TEXT.

The method consists of a sequence of five stages; experimental design, simulator configuration, external configuration, execution and analysis. In the following, each stage is described in detail with reference to the items listed for each stage in Fig. 8. NOTE: renamed stages.

A. *Experimental* Design~~Experiment~~

SURELY, A SERIES OF SYSTEMATICALLY DESIGNED EXPERIMENTS, EACH REPEATED X TIMES, SHOULD BE SET UP, WHERE EACH EXPERIMENT VARIES ONE OF THE SUSPECTED SOURCES OF NON-DETERMINISM? THEN, THE ANALYSIS OF THE DATA FROM EACH EXPERIMENT SHOULD REVEAL THE ACTUAL SOURCES OF NON-DETERMINISM FOR THE GIVEN SETTING? IN OTHER WORDS, WE ARE AIMING TO FIX AS MUCH AS WE CAN CONTROL (EASILY), AND THEN WE INTENTIONALLY VARY (IDEALLY) SINGLE VARIABLES, E.G. THE RESOURCE UTILISATION, TO FIND OUT WHAT MAKES A DIFFERENCE. WE NEED A CLEAR DEFINITION OF THE OBJECTIVES OF THE EXPERIMENTAL DESIGN, I.E. WHAT ONE SHOULD VARY AND WHAT ONE WANTS TO "NAIL DOWN".

1) Actors: Different actor types may be handled differently by the game engine~~and~~ . For instance, our investigation indicated that CARLA pedestrians (termed *walkers*) do not suffer from simulation variance under any conditions that were tested. However, the introduction of CARLA vehicles increased the simulation variance of the pedestrian-only test by 12 orders of magnitude. THIS NEEDS REPHRASING AS IT APPEARS TO CONTRADICT THE STATEMENT ABOVE. As such, all actors that could be included in any future simulation should be tested, including any non-standard CARLA or bespoke actors. ... SUCH AS THE EGO VEHICLE...? [SEE SECTION ON "EXTERNAL SETTINGS" LAST POINT - EGO CONTROLLER.]

2) Trajectories and Collisions: Actor paths, or the sequence of actions required to generate paths, should be hard-coded to ensure repeatability. These paths should include collisions between actors and potentially collisions between actors and static scenery if this is likely to occur in the simulation or as part of the verification process. Actor paths without collision are also important to include as these will serve as a baseline to the other tests. WHAT BASELINE?

3) ~~Setting a Tolerance~~Permissible Variance: The verification engineer should set the *permissible variance*, also termed *tolerance*~~based on their specific requirements which refers to the acceptable level of actor path variance. . This tolerance level is analogous to the analogue signal level associated with each of the binary states in digital circuits. At some low analogue voltage the signal is interpreted as binary 0, this may be, for example, any voltage below 0.5V, and the same signal is interpreted as binary 1 at some higher voltage, e.g. above 3.5V. This convention enables the abstraction from the noisy physical signal values that are analogue voltages into the digital domain with binary representation.~~ The tolerance ~~that the user must set~~ depends on the objectives of the simulation and the granularity at which the simulation environment operates. For example, this tolerance must be sufficiently small to enable accurate assertion checking ~~. In our empirical investigation we set~~ and coverage collection. In Section V-B a tolerance of 1cm ~~which is~~ was considered sufficient for urban scenario assertion checking. In practice, it may be necessary to determine this tolerance experimentally.

4) Simulation ~~Length~~Time: SHOULD THIS NOT BE PART OF THE EXECUTION SECTION? The simulation time should be sufficient to record ~~an interaction~~ interactions between actors but not so long that the testing take an inconvenient time to complete. In our empirical investigation a simulation time of $10 - 20$s was sufficient to monitor the distinct change in events such as the pre- and post-collision including the delayed effect seen by other actors, shown in Fig. 6a. POINT TO THE SECTION THAT SPECIFIED THAT.

B. Simulator Settings

The settings internal to the game engine or other simulation environment should be set to ensure a fixed physics time step, dt. If using CARLA, a small fixed value, say $0.05s$, can be set by using ~~the following:~~ `setting.fixed_delta_seconds = 0.05`~~, whereas in Unity~~. In Unity, the default fixed time step is set to $0.02s$ [35].

In CARLA, synchronous mode must be used to allow communication to external controllers which ensures no sensor data are passed out of order to the simulator which is particularly important if a complex ego controller is used [6].

The use of random numbers must be controlled through ~~the use of fixed seedswhich~~ fixed seeds, resulting in pseudo-randomness. Random numbers might be used in the simulation environment to control variations of background effects, e.g. weather patterns, or the navigation of random pedestrian actors, external vehicle controllers or other clients connected to the simulation environment. Actors that navigate through the environment should use a fixed navigation mesh.

The version number of the CARLA and Unreal environment has also be shown to affect results, see [22]~~, so ensuring~~. Therefore, ensuring a consistent version number throughout testing is also important.

C. External Settings

1) Resource Utilisation: The resources available to the simulator have been shown to have a significant effect on the ~~simulation variance of simulated vehicles and exploring this as a variable the analysis is required.~~ variance of the path of simulated vehicles. Thus, it is important to understand at what level of resource utilisation the system running the simulation becomes susceptible to *simulation variance*.

CPU utilisation software, such as the linux `stress` tool, which is a workload generator program, can be used to spawn workers on any number of cores or virtual threads on a system. This can be used to artificially ~~increased~~ increase the load on the system~~to explore at what point your system may become susceptible to *simulation variance*~~. For GPU utilisation, `gpu-burn` ~~was~~ can be employed using the `fur test`~~where different~~. Different resolutions and multiple instances can be used to tune graphical utilisation levels [57]. Reported values of resource utilisation can be obtained using the system monitors `htop` and `nvidia-smi` for CPU and GPU, respectively. These values ~~can also be written into the data logging for completeness.~~

should be added to the data logs. Alternatively, in place of artificial resource utilisation, multiple instances of the simulation could be executed simultaneously. However, the granularity of control with this approach may be reduced.

2) Memory Testing: Prior to experimental execution the system hardware should be tested for memory conformity and to ensure no single bit errors are occurring, see Section V-H. For mainboard memory `memtest86` can be used on most platforms to run a series of pre-defined memory test patterns. This memory testing software can also be used for ECC enabled hardware. Similarly, to test memory on Nvidia based graphical adaptors `cuda_memtest` can be used to ensure no memory errors exist.

3) Scheduling: We hypothesise that thread scheduling may be a major contributor to the non-deterministic results found in the empirical study. However, gaining fine control over the scheduling policy and thread execution order ~~may be non-trivial. Such policies are defined by the operating system and may execute threads with equal priority. In~~ is known to not be trivial [51].

The operating system schedules threads according to a specified scheduling policy, potentially based on equal thread priority. Thus, in such a case, all tasks non-essential to the simulator should be terminated to prevent ~~interruption to the simulator~~interference with the simulation.

~~Setting the simulator process to~~ Assigning a higher priority to the simulator process may help to alleviate conflicting task scheduling which can be achieved by using, for example `TaskSettings.Priority` in Windows [67] or `NICE` in Linux [37].

4) NUMA Control: Control over a Non-Uniform Memory Access policy can be achieved using `numactl` for multiprocessors with shared memory. This control allows the simulator program to be fixed on a single core~~reducing~~, reducing and unifying memory access time. ~~Initial screening tests done~~ Note that initial screening tests performed with NUMA control ~~gave moderate~~ resulted in only minor improvements in simulation variance~~of the order of a few percent~~, see Section V-H. This control may assist if simulation variance is borderline to the tolerance but not seen as essential. UNCLEAR: WHAT IS NOT ESSENTIAL?

5) Ego Vehicle Controller WHY IS THIS NOT COVERED UNDER ACTORS? [SEE MY NOTE IN THE ACTORS SECTION]: CLEARLY SEPARATE THE INVESTIGATION FROM THE METHOD. THE FACT THAT THE CASE STUDY DID NOT HAVE AN EGO VEHICLE SHOULD BE MENTIONED AS A LIMITATION OF THE CASE STUDY, RATHER THAN ELABORATED ON HERE, BUT WHY DOES THIS MAKE A DIFFERENCE? The impact of the ego vehicle should be considered as an actor but ensuring that any randomness used in the controller is controlled with fixed seeds and that any adaptive learning algorithms should be controlled in the same manner. LONG AND CUMBERSOME SENTENCE WITH STRANGE GRAMMAR: NO EGO VEHICLE WAS USED IN THE EMPIRICAL INVESTIGATION BUT THIS SHOULD BE CONSIDERED AS AN ACTOR BUT ENSURING THAT ANY RANDOMNESS USED IN THE CONTROLLER IS CONTROLLED WITH FIXED SEEDS AND THAT ANY ADAPTIVE LEARNING ALGORITHMS SHOULD BE CONTROLLED IN THE SAME MANNER.

D. Execution

1) Sample Size: Initial testing [22] indicated an actor path deviation of 1×10^{-13}cm for 997 out of 1000 tests~~but with 3~~, with three tests reporting a deviation of over ~10cm. While executing 100 repeats may seem sufficient, this sample size may fail to observe ~~these~~ events that occur with low probability, ~~inferring a~~ giving false confidence in the results. AGAIN, SEPARATE THE STUDY FROM THE METHOD. THIS OBSERVATION SHOULD REALLY HAVE BEEN MADE IN THE SECTION ON "EXPERIMENTAL SYSTEM CONFIG AND PRE-SCREENING, TO JUSTIFY THE FIGURE OF 1000 REPEATS. It is therefore recommended that the sample size is ~~found~~ determined empirically dependent on the number of results that exceed the tolerance. HOW WOULD ONE KNOW?

2) Data Logging: ~~Data logging should be used to record the actorpositions~~ Unique identifiers should be assigned to each experiment, each repeat and each individual actor. The

time-stamped actor positions should then be recorded at fixed time intervals throughout the simulation in order to determine the variance in actor path. Additional information ~~can~~ should also be logged such as the CPU and GPU utilisation levels and engine specific metrics such as game loop latency. ~~For the subsequent analysis, the actor should have an identifier and the repeat number and simulation time should also be captured.~~

E. Analysis

~~The~~ For each experiment, the maximum value of actor path deviation over all time samples and actors, $\max\sigma$, should be analysed to identify which of the candidate sources of non-determinism require restriction or control ~~for permissible operation. Finding these boundaries will identify the *domains of permissible variance* for the user specific hardware.~~to reach the domain of permissible variance within the simulation environment.

UNCLEAR HOW THIS RELATES TO ANALYSIS. IS THIS A SUMMARY OF THE METHOD? IF SO, IT SHOULD BE MOVED TO THE FRONT OR TO THE CONCLUSION. This method can be extended to a broader actor states and actions including actor orientation, speed, and any other status indicators that may be of interest and also including appropriate actions that may be useful for verification purposes. This broader As such, Equation 1 would need to be adjusted to include these new variables. SOMETHING IS MISSING IN THIS SENTENCE - PLEASE CORRECT BEFORE MOVING TO A DIFFERENT PLACE, OR REMOVING.

VII. Conclusions & Future Work

~~The autonomous vehicle community are adopting game engine simulators for such purposes as control system~~ Game engines offer simulation environments that are used for the development and verification of autonomous driving functions. ~~Having a deterministic~~ Determinism of a simulator is required ~~for precise results,~~ to achieve repeatability, which is essential to find and fix software bugs ~~and ensure~~ efficiently, and to ensure simulation results are trustworthy. If a simulator is non-deterministic then practitioners should at least be aware of, and know how to find, the operational domains ~~for reliable results. During an~~ where *simulation variance* is tolerable.

An investigation into the CARLA simulator ~~, a *simulation variance* was observed~~ revealed a significant simulation variance for repeated tests with the same initial conditions ~~indicating non-deterministic execution~~and event history, indicating non-determinism of the simulation. ~~During the study we hypothesized and uncovered several parameters~~ We then researched, identified and discussed potential sources for non-determinism in this context. A systematic case study of the CARLA simulator uncovered the actual factors that contribute towards greater simulation variance~~and hence~~, giving rise to non-deterministic execution~~, such as~~. In particular, actor collisions and ~~system resource utilisation~~ system-level resource utilisation were identified as key contributors. ~~The results of the investigation are hardware and software version specific, so we proposed a general methodology that~~ SAY WHAT THE RECOMMENDATIONS WERE FOR THE REQUIRED TOLERANCE OF THE CASE STUDY, I.E. SUMMARISE OUR RESULTS/RECOMMENDATIONS HERE AND POINT TO GITHUB FOR DATA/DETAIL. THIS SEEMS TO HAVE BEEN COMMENTED OUT BUT WAS HERE BEFORE.

A general method to assess the actor path variance of a game-engine-based simulation environment is then proposed. The method can be used to find the *domains of permissible variance* ~~for any~~ of a simulation environment for a given system configuration. ~~This methodology will allow the AV verification community using simulation tools to ensure their results are reliable and trustworthy.~~

This can give AV developers and verification engineers increased confidence in the simulation results and reduce debug time. ~~A potential avenue of exploration is developing a version of the simulator which has been designed specifically for verification. In this version the simulator would be deterministic because the execution is suitably controlled along with sources of~~ As future work the method can be extended to criteria other than the actor path, e.g. actor orientation and any status indicators that may be of interest, also including actions, sequences and timings that may be useful for verification purposes.

An ambitious avenue for future work is the development of a deterministic simulator for AV development and verification. This requires controlling all potential sources of non-determinism, including randomness and scheduling. ~~A similar task was achieved with the~~, very similar to the development of the record-and-reply debugger ~~`RRprogram [46]`~~`rr` [46], originally developed to catch low-frequency non-deterministically failing tests at Mozilla [51].

~~With the advent of more AI systems becoming adaptive, meaning that the system may change it's response to the same stimulus over time, the notion of repeatability will need reassessing in the context of verification. The main research question here is what will constitute 'the same' which will be important for passing or failing verification tests but also for determining coverage.~~

References

[1] *AirSim drones simulator*. https://microsoft.github.io/AirSim/. Accessed: 2020-011-13.

[2] *Apollo autonomous driving soultion*. http://apollo.auto/. Accessed: 2020-011-13.

[3] T. Atkins and M. Escudier. *A Dictionary of Mechanical Engineering*. Oxford University Press, 2013.

[4] J. Austin. *Fix Your Unity Timestep*. [Accessed: 04-03-2020]. URL: https://johnaustin.io/articles/2019/fix-your-unity-timestep.

[5] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley. "The ML test score: A rubric for ML production readiness and technical debt reduction". In: *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017* 2018-January (2017), pp. 1123–1132.

[6] Carla. *Carla: Configuring the simulation*. [Accessed: 04-03-2020]. URL: https://carla.readthedocs.io/en/0.8.4/configuring_the_simulation.

[7] *CARLA: Open source simulator for autonomous driving research*. http://carla.org/. Accessed: 2020-011-13.

[8] F. Codevilla. *CARLA 0.8.2 Driving Benchmark*. http://carla.org/2018/04/23/release-0.8.2/. [Accessed: 26-03-2019].

[9] *Collision Overview - Unreal Engine 4*. https://docs.unrealengine.com/en-US/Engine/Physics/Collision/Overview/index.html. Accessed: 2020-011-13.

[10] *Cry Engine*. https://www.cryengine.com/. Accessed: 2020-011-13.

[11] *Cuda memtest - Tests GPU memory for hardware errors and soft errors using CUDA*. https://github.com/ComputationalRadiationPhysics/cuda_memtest. Accessed: 2020-011-13.

[12] T. J. Dell. "A white paper on the benefits of chipkill-correct ECC for PC server main memory". In: *IBM Microelectronics Division* (1997), pp. 1–23.

[13] P. E. Dodd and L. W. Massengill. "Basic mechanisms and modeling of single-event upset in digital microelectronics". In: *IEEE Transactions on nuclear Science* 50.3 (2003), pp. 583–602.

[14] *Does MATLAB use all cores by default when running a program? - MATLAB Answers*. https://uk.mathworks.com/matlabcentral/answers/317128-does-matlab-use-all-cores-by-default-when-running-a-program. Accessed: 2020-011-13.

[15] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. "CARLA: An Open Urban Driving Simulator". In: *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*. 2017, pp. 1–16.

[16] G. Fiedler. "Fix Your Timestep". In: (2004). Accessed: 2019-12-18.

[17] Z. Gao, Y. Liang, M. B. Cohen, A. M. Memon, and Z. Wang. "Making System User Interactive Tests Repeatable: When and What Should We Control?" In: *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. Vol. 1. 2015, pp. 55–65.

[18] D. Goldberg. "What every computer scientist should know about floating-point arithmetic". In: *ACM Computing Surveys (CSUR)* 23.1 (1991), pp. 5–48.

[19] J. Gregory. *Game Engine Architecture*. 2nd ed. CRC Press, 2017. Chap. 7.

[20] C. Hutchison et al. "Robustness Testing of Autonomy Software". In: *International Conference on Software Engineering Software Engineering in Practice Track* (2018).

[21] "IEEE Standard for Floating-Point Arithmetic". In: *IEEE Std 754-2019 (Revision of IEEE 754-2008)* (July 2019), pp. 1–84.

[22] *Investigating Unreal Engine for deterministic behaviour*. [Accessed: 06-03-2020]. URL: https://github.com/TSL-UOB/CAV-Determinism/tree/master/UnrealEngineTests.

[23] A. Y. Javaid, W. Sun, and M. Alam. "UAVSim A simulation testbed for unmanned aerial vehicle network cyber security analysis". In: *2013 IEEE Globecom Workshops, GC Wkshps 2013* (2013), pp. 1432–1436.

[24] N. Kalra and S. M. Paddock. "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" In: *Transportation Research Part A Policy and Practice* 94 (2016), pp. 182–193.

[25] N. Kapre and A. DeHon. "Optimistic parallelization of floating-point accumulation". In: *Proceedings - Symposium on Computer Arithmetic* (2007), pp. 205–213.

[26] L. Kliemann and P. Sanders. *Algorithm Engineering Selected Results and Surveys*. 2nd ed. Springer, 2016.

[27] P. Koopman and M. Wagner. "Toward a Framework for Highly Automated Vehicle Safety Validation". In: *SAE Technical Paper Series* 1 (2018), pp. 1–13.

[28] K. Korosec. *Waymo's self driving cars hit 10 million miles*. https://techcrunch.com/2018/10/10/waymos-self-driving-cars-hit-10-million-miles. Accessed: 2019-11-26. 2019.

[29] C. L. Liu and J. W. Layland. "Scheduling algorithms for multiprogramming in a hard-real-time environment". In: *Journal of the ACM (JACM)* 20.1 (1973), pp. 46–61.

[30] Q. Luo, F. Hariri, L. Eloussi, and D. Marinov. "An Empirical Analysis of Flaky Tests". In: *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. FSE 2014. Hong Kong, China: Association for Computing Machinery, 2014, pp. 643–653.

[31] T. d. Margerie. *Precise frame rates in Unity*. [Accessed: 24-06-2020]. URL: https://blogs.unity3d.com/2019/06/03/precise-framerates-in-unity/.

[32] *MemTest86 - The standard for memory diagnostics*. https://www.memtest86.com/. Accessed: 2020-011-13.

[33] N. Mielke et al. "Bit error rate in NAND flash memories". In: *2008 IEEE International Reliability Physics Symposium*. IEEE. 2008, pp. 9–19.

[34] S. Miglio. *AI in Unreal Engine learning through virtual simulations*. https://www.unrealengine.com/en-US/tech-blog/ai-in-unreal-engine-learning-through-virtual-simulations. [Accessed: 26-03-2019].

[35] *MonoBehaviour - Unity Game Engine*. https://docs.unity3d.com/ScriptReference/MonoBehaviour.FixedUpdate.html. Accessed: 2020-011-13.

[36] J.-M. Muller et al. *Handbook of Floating-Point Arithmetic*. 2nd ed. Springer International Publishing, 2018.

[37] *NICE - run a program with modified scheduling priority*. https://linux.die.net/man/1/nice. Accessed: 2020-011-13.

[38] J. Nieplocha, R. J. Harrison, and R. J. Littlefield. "Global arrays: A nonuniform memory access programming model for high-performance computers". In: *The Journal of Supercomputing* 10.2 (1996), pp. 169–189.

[39] A. Nötzli and F. Brown. "LifeJacket: Verifying Precise Floating-Point Optimizations in LLVM". In: *Proceedings of the 5th ACM SIGPLAN International Workshop on State Of the Art in Program Analysis*. SOAP 2016. Santa Barbara, CA, USA: Association for Computing Machinery, 2016, pp. 24–29.

[40] *Numactl - Control NUMA policy for processes or shared memory*. https://linux.die.net/man/8/numactl. Accessed: 2020-011-13.

[41] *NumPy Variance*. https://numpy.org/doc/stable/reference/generated/numpy.var.html. Accessed: 2020-09-03. NumPy, 2020.

[42] Nvidia. *Nvidia drive constellation*. [Accessed: 24-06-2020]. URL: https://www.nvidia.com/en-gb/self-driving-cars/drive-constellation/.

[43] R. Nystrom. *Game Programming Patterns*. 1st ed. Genever Benning, 2011.

[44] "Preliminary Report Highway HWY18MH010". In: *National Transportation Safety Board* (2018).

[45] *Releasing Your Project - Unreal Engine 4*. https://docs.unrealengine.com/en-US/Engine/Deployment/Releasing/. Accessed: 2020-011-13.

[46] *RR lightweight tool for recording*. https://github.com/mozilla/rr. Accessed: 2020-011-13.

[47] Z. Saigol and A. Peters. "Verifying automated driving systems in simulation framework and challenges". In: *25th ITS World Congress, Copenhagen* (2018).

[48] J. Schumann, P. Gupta, and Y. Lui. "Application of Neural Networks in High Assurance Systems: A Survey". In: *Applications of Neural Networks in High Assurance Systems. Springer, Berlin, Heidelberg,* (2010), pp. 1–19.

[49] D. Sculley et al. "Hidden technical debt in machine learning systems". In: *Advances in Neural Information Processing Systems* 2015-January (2015), pp. 2503–2511.

[50] G. Shi, J. Enos, M. Showerman, and V. Kindratenko. "On testing GPU memory for hard and soft errors". In: *Proc. Symposium on Application Accelerators in High-Performance Computing*. Vol. 107. 2009.

[51] C. Staff. "To Catch a Failure: The Record-and-Replay Approach to Debugging". In: *Commun. ACM* 63.8 (July 2020), pp. 34–40.

[52] *Stat Commands*. https://docs.unrealengine.com/en-US/Engine/Performance/StatCommands/. Accessed: 2020-011-13.

[53] P. E. Strandberg, T. J. Ostrand, E. J. Weyuker, W. Afzal, and D. Sundmark. "Intermittently Failing Tests in the Embedded Systems Domain". In: *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ISSTA 2020. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 337–348.

[54] *Substepping UE4*. https://docs.unrealengine.com/en-US/Engine/Physics/Substepping/. Accessed: 2020-011-13.

[55] U. S. Team. *AI and Behavior Trees*. https://docs.unrealengine.com/en-us/Gameplay/AI. [Accessed: 26-03-2019].

[56] *The Highway Code*. https://www.gov.uk/guidance/the-highway-code. Accessed: 2019-11-25. Department of Transport, UK, 2015.

[57] V. Timonen. *GPU burn*. [Accessed: 17-12-2019]. URL: https://github.com/wilicc/gpu-burn.

[58] *UK Road Traffic Act 1988*. http://www.legislation.gov.uk/ukpga/1988/52/contents. Accessed: 2019-10-03.

[59] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer. "Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving". In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* 2015-Oct (2015), pp. 982–988.

[60] Unity. *Order of Execution for Event Functions*. [Accessed: 24-06-2020]. URL: https://docs.unity3d.com/Manual/ExecutionOrder.html.

[61] *Unity Game Engine*. https://unity.com/. Accessed: 2020-011-13.

[62] *Unreal Engine 4*. https://www.unrealengine.com/. Accessed: 2020-011-13.

[63] *Unreal Engine 4 AI Programming Essentials*. https://www.oreilly.com/library/view/unreal-engine-4/9781784393120/ch04s03.html. Accessed: 2020-011-13.

[64] *Vienna Convention on Road Traffic*. https://treaties.un.org. Accessed: 2019-10-03.

[65] *What is multithreading?* https://docs.unity3d.com/Manual/JobSystemMultithreading.html. Accessed: 2020-011-13.

[66] N. Whitehead and A. Fit-Florea. *Precision & Performance: Floating Point and IEEE 754 Compliance for NVIDIA GPUs*. https://developer.nvidia.com. 2011.

[67] *Windows Task Settings Priority property*. https://docs.microsoft.com/en-us/windows/win32/taskschd/tasksettings-priority. Accessed: 2020-011-13.