

# A Containerized Microservice Architecture for a ROS 2 Autonomous Driving Software: An End-to-End Latency Evaluation

Tobias Betz, Long Wen, Fengjunjie Pan, Gemb Kaljavesi, Alexander Zuepke, Andrea Bastoni,  
Marco Caccamo, Alois Knoll, Johannes Betz

Technical University of Munich, Germany,

{tobi.betz, long.wen, f.pan, gemb.kaljavesi, alex.zuepke, andrea.bastoni, mcaccamo, k, johannes.betz}@tum.de

**Abstract**—The automotive industry is transitioning from traditional ECU-based systems to software-defined vehicles. A central role of this revolution is played by *containers*, lightweight virtualization technologies that enable the flexible consolidation of complex software applications on a common hardware platform. Despite their widespread adoption, the impact of containerization on fundamental real-time metrics such as end-to-end latency, communication jitter, as well as memory and CPU utilization has remained virtually unexplored. This paper presents a microservice architecture for a real-world autonomous driving application where containers isolate each service. Our comprehensive evaluation shows the benefits in terms of end-to-end latency of such a solution even over standard bare-Linux deployments. Specifically, in the case of the presented microservice architecture, the mean end-to-end latency can be improved by 5-8%. Also, the maximum latencies were significantly reduced using container deployment.

**Index Terms**—Software-Defined Vehicle, Autonomous Driving, Containerization, End-to-End Latency, Robot Operating System 2

## I. INTRODUCTION

The automotive market is shifting towards software-defined vehicles (SDV), enabling a more software-centric automotive ecosystem. For example, the open-source consortium SOAFEE [1], [2] specifically targets SDV and brings together OEMs, Tier 1s, and chip manufacturers to work on the challenges. The E/E architecture of SDVs is based on a central computing unit in which a powerful high-performance computer manages and coordinates diverse functionalities. These functions encompass processing of sensor data, operation of infotainment systems, advanced driver assistance systems, and communication with external systems. This enables the separation of software from hardware functionality to achieve greater modularity and scalability. Lightweight virtualization techniques such as containerization enable efficient resource utilization and isolation of software components. This design philosophy empowers applications and services to operate independently within their dedicated virtual environments. When employing virtualization technologies in SDVs, stringent real-time criteria such as latencies and deadlines must be met. This concern becomes particularly important in autonomous vehicles, where the timely processing of sensor data within a predefined time window is critical for enabling prompt decision-making and control. Typically, end-to-end latencies of 100 ms are considered acceptable, wherein the sensor data must be swiftly processed, and the resulting output variables must be made available from the vehicle’s trajectory controller [3]. Therefore, the end-to-end latency directly impacts the vehicle’s ability to

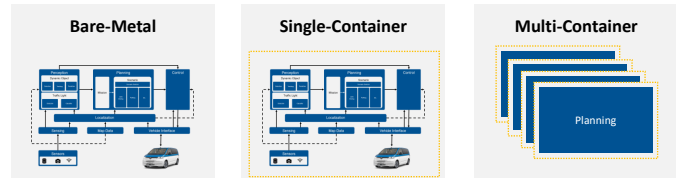


Fig. 1: Considered software deployments: bare-metal (no containers), one container, isolation via dedicated containers.

navigate and respond to dynamic road conditions in a safe and reliable manner. Failure to meet these real-time requirements could lead to performance degradation and an increased risk of accidents [4]. In the context of software-defined autonomous driving architectures, practitioners have been experimenting with frameworks that simplify the difficult tasks of configuring, tuning, and optimizing the complex chains of architectural interdependencies. In particular, Autoware [5] and the Robot Operating System (ROS) 2 [6] are among the most widely used frameworks.

This paper introduces a microservice architecture—developed and applied to the research vehicle EDGAR [7]—that is explicitly designed for Autoware, an open-source autonomous driving software built on ROS 2. Since the impact of lightweight virtualization technologies on the latency of complex software has—to our knowledge—not yet been considered, this paper investigates the impact of containerization on the end-to-end latency in autonomous driving systems. Specifically, we focus on the end-to-end latency of a real-world autonomous driving architecture based on Autoware. We deploy the architecture on two different platforms (x86 and aarch64) using multiple configurations corresponding to an increasing level of container-based isolation (see Fig. 1). Using standard practice in industry [8], we use the container orchestration tool *k3s* [9] and Docker [10] to manage the functional dependencies among software packages and containers. Overall, the paper makes the following contributions:

- We present the structure and building process of a microservice architecture for autonomous driving software serving as a testbed for future work.
- We perform a comprehensive analysis of the impact of containerization using both specific benchmarks and direct measures on increasingly isolated microservice configurations.
- We quantitatively evaluate real-time metrics, including end-to-end latency, jitter, CPU and memory utilization.

Contrary to the common belief, our results show that containers can achieve lower end-to-end latency and better system utilization than bare Linux configurations. This underlines the challenge of finding the best-suited configuration options in very complex system scenarios and shows the benefit of containerization for future SDV systems. The developed microservice architecture will be contributed open-source to the Autoware Foundation (<https://github.com/autowarefoundation/autoware>).

## II. RELATED WORK

Several papers discuss challenges and advancements in embedded systems and automotive software. Sax *et al.* [11] emphasize the shorter release cycles, increased variants, and software updates in modern vehicles. However, they do not provide any in-depth analysis of particular solutions or tools. The integration of new functionalities increases the complexity of vehicle systems, requiring careful considerations of the architecture and distribution of electronic control units to effectively manage this complexity [12]. Kugele *et al.* [13] discuss elastic service provisioning in intelligent vehicles. The management of different workloads, resource constraints, and changing user requirements is highlighted as a need. Hence, the importance of providing scalable and flexible services that are able to dynamically allocate resources and change performance characteristics based on real-time conditions.

Microservices and service-oriented architectures (SOA) have the potential to improve the flexibility of automotive systems. Lotz *et al.* [14] investigate the feasibility and impact of implementing a microservice architecture for driver assistance systems and demonstrate the reduction of complexity and improvement of software systems. Tamanaka *et al.* [15] present a conceptual framework for a fault-tolerant architecture and highlight the use of microservices and containerization as critical components. In [16], a literature review explores design principles and architectural refinement strategies for microservices. Through a systematic mapping study, Kukulicic *et al.* [17] analyzes the adoption of SOA in automotive software. Functional usability stands out as the most relevant benefit, while issues such as *e.g.*, security, safety, and reliability are identified as challenges. Previous research provides overviews, on a purely theoretical basis, of the challenges and benefits of moving to a microservice architecture (see [18]).

Regarding the performance impact of virtualization and containerization technologies on diverse systems, in [19] the authors introduce a benchmarking suite to assess the resource costs of various virtualization technologies. They compare the performance of hardware native hypervisors, hosted hypervisors, and containers using reference benchmarks. Morabito *et al.* [20] focus on evaluating the performance of containerization on Internet-of-Things edge environments. The strengths and weaknesses of various low-power devices when dealing with container-virtualized instances are highlighted. Notably, both demonstrate that virtualized or containerized systems show acceptable performance compared to bare-metal systems. Felter *et al.* [21] compare the performance of virtual machines (VMs) and Linux containers within cloud computing environments. The study demonstrates that containers outperform or match the performance of VMs in most cases,

emphasizing the potential benefits of using containers in cloud architectures. Similarly, in [22], the authors conduct research on container-based virtualization in high-performance computing, highlighting the low overhead and potential for near-native performance. While all these studies provide valuable insights, they lack experiments on real-world use cases. In the automotive area, Rajan *et al.* [23] explore the technique of bringing virtualization into automotive multicore controllers. The authors evaluate the performance of a virtualized system in terms of core loading, interrupt timing, and task timing parameters. Long *et al.* [24] develop a general benchmark that yields results consistent with the conclusions mentioned previously. Furthermore, they specifically focus on the startup time of microservice-based Autoware automotive applications demonstrating that virtualization and containerization are suitable and viable options. The adoption of these virtualization technologies might be beneficial in the automotive industry. For the development of robot software, ROS 2 is the most widespread framework. [25] presents an exemplary architecture tailored to autonomous driving and the possibilities of using it for high-speed autonomous racing are presented in [26]. The proposed racing architecture is based on microservices where each functional module, *e.g.*, perception, planning, and control, is deployed as a container. Autoware [5] represents the most comprehensive open-source initiative dedicated to ROS 2 for autonomous driving software. In the literature, there are already several different frameworks [27], [28], [29], [30] that allow to measure ROS 2 applications. These are normally based on `ros2_tracing` [31], which instruments trace points into the middleware accordingly. This allows determining callback times as well as end-to-end latencies. In addition, there are benchmark tools that are mostly limited to simple examples where statements can be made about the DDS latency [32], [33] or the entire system performance [34]. For a single-threaded executor system, timing analysis is performed in [35]. Reke *et al.* [25] conduct an end-to-end latency and corresponding jitter analysis for an entire application. In [36] an analysis for Autoware is realized, with a focus on embedded hardware. The authors in [37] analyze the impact of the different system abstraction layers on the end-to-end latency of Autoware. They also tried different Linux scheduling configurations to improve the timing behavior on bare-metal systems. In [38], [39] the focus is only on the influence of the DDS layer. For the ROS 2 autonomous racing software presented in [26], a latency evaluation is carried out in [4]. The focus is on the application layer and vehicle stability impact of timing. However, the influence of the microservice architecture is not evaluated. Based on the current state of the art, it is impossible to find an assessment of the impact of containerization and a corresponding microservice architecture on the end-to-end latency of ROS 2 applications.

## III. MICROSERVICE ARCHITECTURE FOR AN AUTONOMOUS DRIVING SOFTWARE

Following the paradigms of software-definedness, the microservice architecture for autonomous driving software is designed to enhance the modularity of the software, enabling

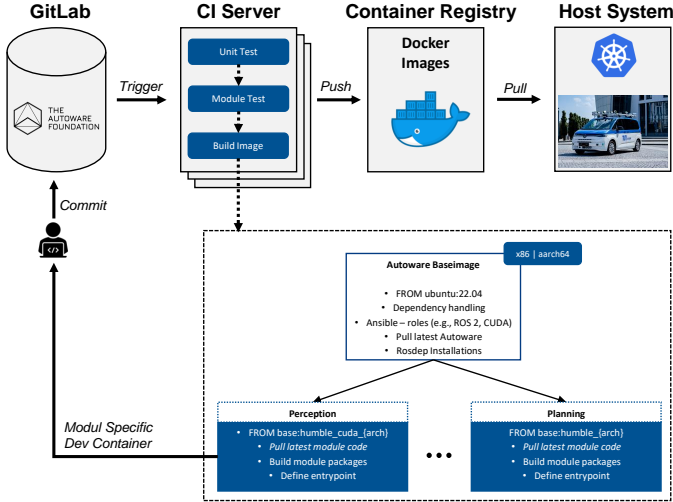


Fig. 2: Schematic of the build and deployment process of the microservice architecture: After committing code changes to the Autoware repository on the CI server, the test procedure and docker image build steps are triggered. The built images are stored in the container registry and can be pulled from the cloud onto the host system. The corresponding module images are based on a base image that contains the necessary basic installations. This module image, in turn, also serves as a container for the development of features.

efficient development and corresponding software deployment. The core of our architecture is a *base image* that forms the basic building block for the individual module containers. Specialized containers implementing dedicated functionalities are derived from the base image. The base image can also be used for development as it has the requirements for building the complete code. The base image includes essential installations such as ROS 2 and optional libraries like `cuda`, `cuDNN`, and `TensorRT`, which are not necessarily required by every specialized module. The advantages of using a single base image are manifold. The configuration and installation of all packages can be centralized using a multi-step process that relies on, e.g., *Ansible* roles, *rosdep* installation, and manual configuration. This also simplifies the management of cross-package dependencies, facilitates *freezing* packages to specific versions, and avoids introducing incompatibilities between (updated) packages and our code.<sup>1</sup> Once configured, the base image rarely needs to be rebuilt. Fig. 2 depicts the build and deployment process of the microservice architecture. We divided the Autoware software into eight dedicated containers based on the functional modules in the software. The containers are sensing, perception, localization, map, planning, control, vehicle, and system. Each container consists of multiple ROS 2 nodes, as shown in Table I. In our architecture, the entire ROS 2 launch structure of Autoware was restructured with the separation of individual modules. The centralized launch package, which listed all packages as dependencies, was split into individual launch packages for each module (with only

<sup>1</sup>Managing dependencies in ROS 2 is particularly complex and manual optimization, as well as package updates, quickly become a daunting task.

TABLE I: Description of the individual services and number of executed ROS 2 nodes for the Autoware microservices.

Service	Nodes	Description
Sensing	48	Collecting and pre-processing of raw sensor data
Perception	49	Object detection, tracking, and prediction of traffic participants
Localization	33	Estimation of vehicle pose, velocity, and acceleration
Map	6	Broadcast semantic and geometric information about the environment
Planning	25	Generation of the trajectory of the ego vehicle
Control	8	Generate control commands to the vehicle
Vehicle	1	Passes control signals to the vehicle and receives vehicle information
System	21	Error monitoring

the needed dependencies). As a result, each module can be built and launched individually. The former central launch package included all launch parameters. In contrast, in our architecture, a separate package was created to contain these launch parameters, which are accessible by the module launch files. Additionally, we integrated the launch parameters to be located outside the containers and mounted during the startup of the respective containers. This approach provides the advantage that changes affecting several module containers, such as the vehicle model, only need to be modified in one location, ensuring consistent parameters for all modules. We developed a continuous integration (CI) pipeline for building custom module containers that ensures compatibility with both `x86` and `aarch64` architectures by using cloud-native hardware resources. The CI pipeline consists of several stages that enable both the creation of the entire software and the targeted creation of individual modules. This approach offers efficiency advantages, eliminating the need to rebuild all containers for each code change. Additionally, it facilitates selective updates and maintenance via a CI pipeline-based multi-stage testing process. Initially, unit tests are conducted, followed by modular tests in which several functions and their interactions are assessed. Due to the modular container structure, a test does not have to be executed repeatedly, but only within the respective container module. We utilize the CI cloud infrastructure to store our built containers in the container registry. The built containers can be seamlessly deployed on both simulation infrastructure and actual vehicles, offering a flexible deployment strategy. Compared to a monolithic architecture, our microservice architecture improves the development and deployment of the software. During development, the software developer only needs to handle the dependencies related to the respective functionality. The building of the software is automatized in the cloud, and the deployment is simplified. This development and deployment workflow of the microservice architecture is successfully used in real vehicle projects [7], [26].

#### IV. EXPERIMENTS

A typical ROS 2 application can be abstracted in several layers ranging from high-level applications to the foundational hardware. We define the layers as depicted in Fig. 3. With the use of containerization, the container runtime adds an additional layer. Positioned above the operating system, this layer facilitates the creation, execution, and management of containers. These executable software packages encapsulate an application and its dependencies. Our study focuses on

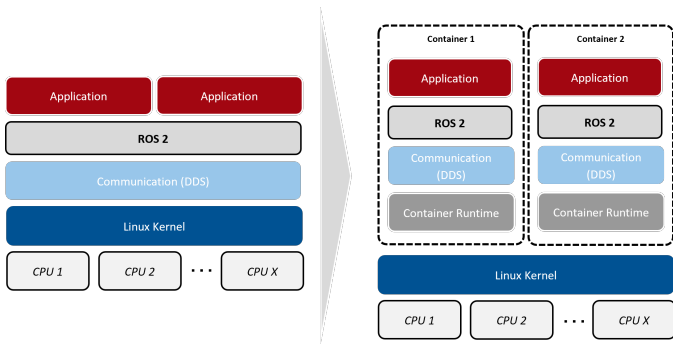


Fig. 3: The abstraction layers of a ROS 2 application executed in bare-metal (left) and in multiple containers (right). Using containerization, an additional layer is introduced. Each container consists of independent individual layers, but shares a common Linux kernel.

TABLE II: Specifications of the computing platforms.

	InoNet Mayflower-B17	ADLINK AVA COM-HPC
CPU	AMD EPYC 7313P (x86)	Ampere Altra Q32-17 (aarch64)
Clock Frequency	16 x 3.0 GHz (max. 3.7 GHz)	32 x 1.5 GHz (max. 1.7 GHz)
RAM	4 x 32 GB	32 GB
GPU	NVIDIA RTX A6000 48 GB	
Disk	Samsung 980 PRO NVMe M.2 SSD - 2 TB	
Kernel	6.2.0-34-generic	
OS	Ubuntu 22.04.3 LTS Jammy Jellyfish	

understanding the influence of containerization on ROS 2 applications. Specifically, our experiments were systematically designed with an increasing complexity:

- **DDS Communication:** This experiment examines the pure communication performance of DDS in isolation.
- **ROS 2:** A publish/subscribe example is introduced to observe the performance implications of DDS and ROS 2.
- **Real-World Autonomous Driving Application:** Incorporates the impact of containerization on the developed microservice architecture.

For each of the experiments, the three increasing-isolation deployments (Fig. 1) have been evaluated. The first scenario (bare-metal) serves as reference point and tests run natively on the system without containerization. In the second scenario we ran the test within a single container. This aims to measure the overheads introduced by containerizations in the first place. The third scenario, multi-container, placed the respective benchmark algorithms in separate containers.

In this section, we first introduce our hardware setup and the specific configurations of our containerization architecture. Afterward, we describe the DDS, ROS 2, and Autoware benchmark setups with their individual metrics.

#### A. Hardware Setup and Software Configurations

All experiments were conducted on two distinct computing platforms (one x86 and one aarch64 (Armv8)), as depicted in Table II. The two platforms are representative of autonomous driving platforms for SDVs [40] and use the same GPU, OS, and Kernel version. On the x86 computing system, we

disabled hyperthreading to minimize potential performance fluctuations. The corresponding experiments are performed with ROS 2 *Humble Hawksbill* with the underlying middleware Eclipse CycloneDDS [41]. We chose Docker (version 24.0.5) as containerization technology due to its advanced GPU integration capabilities, which provide an obvious advantage over alternative solutions like Podman. To orchestrate the microservice architecture, we utilized k3s (version v1.27.3+k3s1) to deploy and manage the containers. We employed the `nvidia-docker2` package to enable GPU support for Docker and the `nvidia-device-plugin` for k3s. The container `pods` are configured in such a way that they communicate over the local host network. No CPU requests or limits are set in the configuration. The standard Linux Completely Fair Scheduler (CFS) is used for every experiment. Despite not being a “true” real-time setup, we are interested in replicating a soft real-time environment that reflects the typical setups for software-defined architectures adopted by the practitioners [1], [42].

#### B. Benchmarks

1) **DDS Communication:** To test DDS communication, we use the `ddsperf` benchmark from the Eclipse CycloneDDS. This benchmark focuses purely on DDS communication, as it skips the ROS 2 abstraction layer. This approach enables us to investigate the influence of containers on pure DDS communication. The experiment uses a straightforward “ping pong” communication pattern to analyze containerization’s impact on DDS performance. This pattern consists of continuously sending a defined message size back and forth between two nodes. In the multi-container scenario, each node is placed in an individual container. CycloneDDS can be configured in two modes: reliable and best-effort. In the best-effort setting, a publisher sends messages without any assurance that the recipient will receive them correctly. Conversely, in reliable mode, the publisher continues sending messages until it receives an acknowledgment from the subscriber indicating successful reception. Given that best-effort is the default setting for most nodes in the Autoware software, we opt for this mode for our study. Another crucial aspect was the variation in message size. Starting at 1 kB, the size was gradually increased by doubling message sizes to analyze the impact on performance across a spectrum of message sizes up to 8 MB. This variation allowed us to assess the scalability and efficiency of DDS communication under different load conditions. Finally, each test was run three times to ensure reproducibility and consistency of results. Each run with a different message size lasted 30 minutes. This time period was chosen, in particular, to ensure that a sufficient number of packets could still be exchanged during the tests with the largest message sizes.

2) **ROS 2:** We used the NVIDIA-ISAAC-ROS `ros2_benchmark` from [34] to evaluate the impact of containerization on simple ROS 2 applications. This benchmark framework is sophisticated and allows testing several example ROS 2 graphs. From the `ros2_benchmark`, we chose the *AprilTag* [43] node as a reference for our evaluation. The benchmark includes a playback node that sends camera data, which is in turn processed by the *AprilTag*

detection node. The benchmark also comprises a data-loader node that loads the *rosbag* *r2b\_storage* data into a buffer and sends it to the playback node. A monitoring node for benchmark-internal evaluations (e.g., CPU monitoring) is also included. In the bare-metal configuration, we run the entire framework without changes to the systems. In the single-container configuration, we put the playback and detector nodes inside the single container, whereas in the multi-container configuration, we separate both nodes into individual containers. We let the benchmark complete a total of 100 runs per each deployment type. Each individual run consists of 5 internal iterations. Eventually, the benchmark outputs a statistical result for the five iterations, which we merge accordingly for the 100 runs. In our experiments, the benchmark tests four different setups in terms of the publishing frequency of the playback node: 10 fps (100 ms), 30 fps (33.3 ms), 60 fps (16.7 ms), and an additional setup where the system is configured to achieve the maximum throughput. With increased framerate, the workload for the system also grows. Therefore different stress levels of the system can be evaluated.

### C. Real-World Autonomous Driving Application

We evaluated the performance impact of containerization on Autoware in the microservice architecture presented in Section III. In the bare-metal setup, the Autoware software is created natively on the system and launched accordingly. In the container environments, Autoware is installed inside of one container. The launch command of the bare-metal variant is defined as an entry point in the container and can then be started with *k3s*. For the microservice architecture, as previously described, each module has its individual launch command defined in the entry point of the container. For all three deployment variants, it is guaranteed that the same software version is compared. We leverage the orchestration framework proposed in [37] to simulate in a closed-loop the deployed Autoware variants using the *AWSIM* environment. The Autoware software is executed standalone on the described compute platforms, and the simulation is executed on a different compute unit. The vehicle is driving on a defined test route in Nishi-Shinjuku in Tokyo, Japan. Traffic participants were removed from the simulation because they cannot be simulated in a reproducible manner. Each experiment is repeated until 100 valid runs can be evaluated. Each test drive takes approximately two minutes to reach the goal pose.

### D. Metrics

It is important to develop metrics at both application and system levels to analyze the impact of containerization. Such metrics provide valuable insights into resource utilization, helping to identify the latency impact induced by containerization. However, benchmarks are often published with their metrics, making it difficult to evaluate all experiments consistently. In the following, we will go into more detail about the metric used for each experiment.

1) *DDS Communication*: The benchmark provides the throughput of packets sent during the test period. In addition, the round trip latency is displayed, which is the time it takes for a message to be sent from the source node to the destination node and back again. The benchmark does not provide the CPU load during the execution. After the tests, we calculate the average round trip time and the average throughput.

2) *ROS 2*: The framework outputs different metrics for each test node. We evaluate the mean end-to-end latency from sending the raw data until the test node generates an output. This metric is calculated internally in the benchmark via tracing points. Also the mean jitter of the corresponding node is measured. Additionally, the framework provides insight into CPU utilization. We evaluate the average CPU utilization over the test runs.

3) *Real-World Autonomous Driving Application*: The complex ROS 2 Autoware setup is evaluated using the data-age (end-to-end latency) metric, shown [44] to be equivalent to the reaction time. It is the average of path durations with the same sensor input. For this, the framework of [29] is used, which can determine the end-to-end latency for Autoware accordingly. The computation is based on *ros2\_tracing*, which places corresponding trace points in the *rclcpp* client library of the ROS 2 middleware. To enable tracing while using the containerized architecture of Autoware, it was necessary to mount specific *LTTng* related file information from the host system into each of the containers. Inside the containers *ros2\_tracing* must be enabled. For the bare-metal and containerized measurements, the tracing session was executed on the host system. The framework computes the total end-to-end latency as well as its individual components:

- The *idle* latency or intra-node communication latency defines the time between a subscription callback and a timer callback of a ROS 2 node.
- The *communication* latency is the time between publishing and receiving a ROS 2 message via a subscription callback. It corresponds approximately to the time needed for the DDS communication.
- The *compute* latency describes the time it takes to process the input from a subscription and publish the corresponding output data to the subsequent node.

Since Autoware consists of a large number of individual

TABLE III: ROS 2 callback signatures for the evaluated computation chain.

Computation Chain
(0) Filter::(PointCloud2,PointIndices)
(1) NDTScanMatcher::(PointCloud2)
(2) EKFLocalizer::(PoseWithCovarianceStamped)
(3) EKFLocalizer::()
(4) StopFilter::(Odometry)
(5) BehaviorPathPlannerNode::(Odometry)
(6) BehaviorPathPlannerNode::()
(7) BehaviorVelocityPlannerNode::(PathWithLaneId)
(8) ObstacleAvoidancePlanner::(Path)
(9) ObstacleVelocityLimiterNode::(Trajectory)
(10) ObstacleStopPlannerNode::(Trajectory)
(11) ScenarioSelectorNode::(Trajectory)
(12) MotionVelocitySmootherNode::(Trajectory)
(13) PlanningValidator::(Trajectory)
(14) Controller::(Trajectory)
(15) Controller::()
(16) VehicleCmdGate::(AckermannControlCommand)

computational chains, we selected a single chain for evaluating latency. This chain, detailed in Table III, was chosen to traverse as many containers as possible for a more accurate assessment of their influence. Furthermore, it represents the critical path with the highest latency in the application. The quality of service setting is configured to “keep last,” operating in best-effort mode with a queue length of 1. To measure the CPU and memory utilization of Autoware, we recorded the process status using Linux `ps`. We recorded the information for all processes every 200 ms. As we are interested in the influence of the containerized ROS 2 application, the processes are correspondingly filtered after the session to ROS 2, Docker, Kubernetes, and Autoware processes.

## V. RESULTS

### A. DDS Communication

The performance benchmark results offer valuable insights into the effects of containerization. As described in Section IV-B1, we used CycloneDDS and measured the round trip latency of specific message sizes and the number of successfully delivered packets. Table IV presents the measured results for both of the compute platforms from message sizes ranging from 1 kB to 8 MB. On both platforms, container-based deployment achieves lower latencies than bare-metal deployment in almost all scenarios. This effect is more evident on `aarch64` and particularly pronounced for small message sizes. On `aarch64`, bare-metal configurations never perform better than containers, while only 2 MB bare-metal messages achieve a lower latency on `x86`. The latency improvement is more evident on `aarch64` than `x86`. For example, for 1 kB, on `aarch64` containers achieve around 15% lower latency (around 8-10% on `x86`), while for 64 kB, the improvement is even higher (18% vs. 8%). Multi-containers can consistently perform better than single-containers. This is the case for large message sizes on `aarch64`, where multi-container setups always perform better than single-container starting from 512 kB message sizes. This trend is not confirmed on `x86`, where single-container setups perform better for large message sizes. In a real application such as the Autoware software stack shown later, the observed message sizes are in the range of 1 kB to 128 kB. This is a range where containerized versions on both systems showed smaller latencies.

### B. ROS 2

We further investigated the combined performance implications of DDS and ROS 2 using `ros2_benchmark`. We present the experimental setup in Section IV-B2. Image data with a size of 0.92 MB is transferred for the data set used. In addition to the pure DDS time, the end-to-end latency now also includes the computation time of the detection algorithm. Therefore, in percentage terms, the DDS time has a much smaller share. Table V shows the measured results of the conducted benchmark. Contrary to `ddsperf`, latency values in `ros2_benchmark` are very close across all setups. Although containerization can achieve slightly lower latency than bare-metal for low fps (10 and 30), bare-metal performs slightly better at 60 fps. Differences in latency are minimal (between

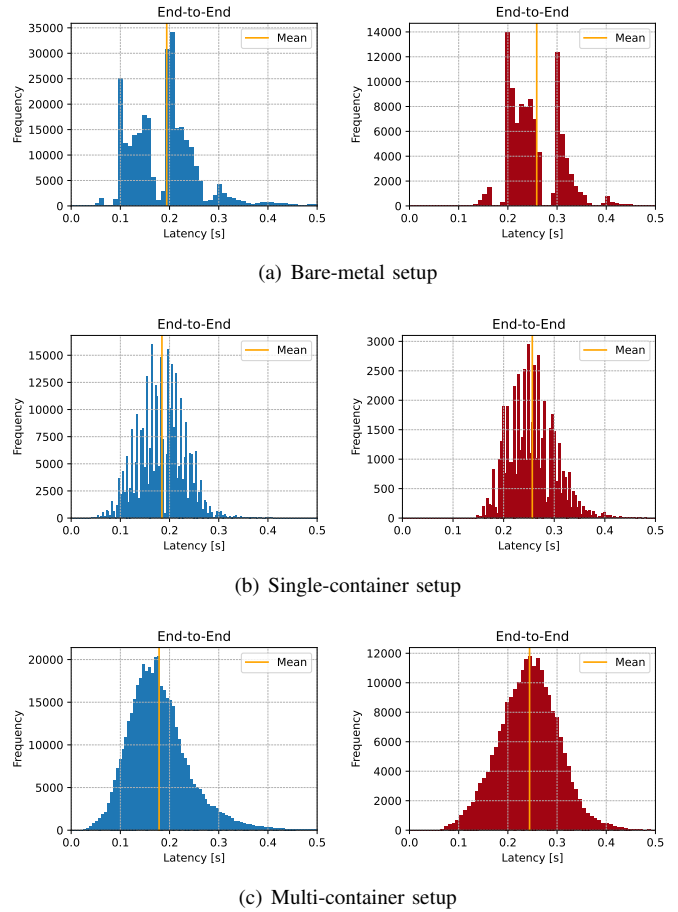


Fig. 4: End-to-end latency histograms (a) using the bare-metal setup, (b) using the single-container setup, and (c) using the multi-container setup (blue=`x86`, red=`aarch64`).

0.4% and 4.6%). Looking at the jitter shows a reduction due to containerization in almost all cases. For the 10 fps setup, the occurring jitter is reduced by 10.3% for single-container and 8.1% for multi-container. At 60 fps by 6.2% and 3.10%. At maximum throughput, only minor differences occurred. An outlier occurs in the multi-container deployment only in the 30 fps setup. The `aarch64` platform also shows this behavior in all setups except for the maximum throughput. At 10 fps, the jitter is reduced by 30.1% and 31.4%. Again, for the other setups, we observe the same behavior. Only the multi-container deployment in the last setup exhibits slightly increased jitter. Overall, we conclude that containerization may lead to a reduction in latency. Additionally, we observe slight differences in CPU utilization. Specifically, for `x86`, the single-container setup demonstrates the lowest utilization compared to other deployments. Conversely, for `aarch64` at lower fps, the bare metal benchmark outperforms the containerized benchmark, while for higher fps, the single-container setup performs the best.

### C. Real-World Autonomous Driving Application

At last, we evaluate the developed multi-container microservice architecture (see Section III) for an autonomous vehicle

TABLE IV: Mean latency and package count for different message sizes for the `ddsperf` benchmark. (BM=bare-metal, SC=single-container, MC=multi-container). Min values across configurations are marked in **bold**.

Message Size	x86						aarch64					
	Mean Latency [ $\mu$ s]			Mean Package Count			Mean Latency [ $\mu$ s]			Mean Package Count		
	BM	SC	MC	BM	SC	MC	BM	SC	MC	BM	SC	MC
1 kB	9.40	8.40	<b>8.29</b>	95183875	106363549	<b>107825676</b>	35.08	<b>29.22</b>	29.76	25466930	<b>30566990</b>	30019061
4 kB	10.67	<b>9.47</b>	9.84	83878392	<b>94467569</b>	90891488	37.47	33.83	<b>33.40</b>	23865466	26422565	<b>26762629</b>
8 kB	11.56	<b>10.07</b>	10.41	77502760	<b>88876974</b>	85965143	43.24	<b>36.87</b>	38.12	20685487	<b>24257601</b>	23468802
16 kB	18.31	16.56	<b>16.07</b>	48958255	54174267	<b>55694824</b>	50.95	<b>40.44</b>	41.66	17573440	<b>2212799</b>	21467480
32 kB	26.69	<b>23.58</b>	23.90	33623951	<b>38049383</b>	37530124	78.29	70.62	<b>69.39</b>	11445854	12690563	<b>12915312</b>
64 kB	40.66	37.52	<b>37.14</b>	22096636	23936672	<b>24175705</b>	126.98	103.47	<b>103.42</b>	7065036	8671388	<b>8672852</b>
128 kB	68.57	68.64	<b>66.15</b>	13104554	13088043	<b>13578698</b>	219.03	<b>191.06</b>	195.60	4099368	<b>4698385</b>	4589767
256 kB	127.44	<b>125.71</b>	127.35	7051885	<b>7148614</b>	7056868	421.19	<b>370.59</b>	374.44	2132447	<b>2423861</b>	2399043
512 kB	270.62	<b>269.02</b>	269.67	3380179	<b>3386639</b>	3381805	1526.63	1359.85	<b>993.96</b>	649670	716643	<b>944632</b>
1 MB	693.24	<b>639.32</b>	707.37	1323072	<b>1427427</b>	1290395	2415.79	2033.56	<b>1868.45</b>	401169	461185	<b>499301</b>
2 MB	<b>1201.41</b>	1302.54	1709.00	<b>761672</b>	704386	532031	5310.44	4662.29	<b>3980.55</b>	181039	205960	<b>242480</b>
4 MB	2665.02	<b>2655.18</b>	3071.55	340980	<b>342625</b>	294714	9279.03	12661.69	<b>6581.91</b>	102937	92376	<b>139366</b>
8 MB	5338.55	<b>5233.07</b>	5532.21	169790	<b>172760</b>	165345	17157.53	24752.60	<b>13326.46</b>	54806	44914	<b>68176</b>

TABLE V: Mean latency, jitter, and CPU utilization for different setups of the `ros2_benchmark`.

Architecture	Setup	Mean Latency [ms]			Mean Jitter [ms]			Mean CPU Util. [%]		
		BM	SC	MC	BM	SC	MC	BM	SC	MC
x86	10 fps	6.13	<b>5.85</b>	5.90	2.33	<b>2.09</b>	2.14	0.96	<b>0.94</b>	0.98
	30 fps	5.99	5.87	<b>5.85</b>	1.90	<b>1.83</b>	2.07	1.60	<b>1.57</b>	1.64
	60 fps	<b>5.76</b>	5.80	5.81	1.62	<b>1.52</b>	1.57	2.78	<b>2.72</b>	2.81
	Max. TP	7.20	<b>7.08</b>	7.26	1.03	<b>1.02</b>	<b>1.02</b>	3.59	<b>3.57</b>	3.63
aarch64	10 fps	17.82	<b>17.74</b>	17.83	1.43	1.00	<b>0.98</b>	<b>1.11</b>	1.13	1.15
	30 fps	17.65	<b>17.58</b>	17.71	1.42	1.40	<b>1.37</b>	<b>2.19</b>	2.24	2.29
	60 fps	<b>22.76</b>	24.73	23.98	1.23	1.19	<b>1.12</b>	2.19	<b>2.15</b>	2.31
	Max. TP	<b>18.03</b>	18.27	18.09	1.61	<b>1.59</b>	1.62	3.57	<b>3.55</b>	3.60

TABLE VI: Consolidated statistics for different metrics for Autoware.

Architecture	KPI	Type	Mean	Std	Skew	Kurtosis	Min	Q25	Q50	Q75	P99	Max
x86	E2E	BM	194.67	84.68	3.01	19.94	31.28	139.84	199.43	223.85	515.87	1437.94
		SC	184.51	<b>47.17</b>	<b>0.18</b>	<b>0.35</b>	35.37	152.59	183.86	213.95	<b>300.79</b>	<b>553.79</b>
		MC	<b>178.73</b>	64.42	1.19	3.59	<b>25.77</b>	<b>136.00</b>	<b>171.30</b>	<b>210.20</b>	383.03	903.45
	Idle	BM	133.81	73.12	3.04	21.88	5.24	87.09	129.82	165.77	406.713	1274.48
		SC	<b>125.92</b>	<b>44.89</b>	<b>0.26</b>	<b>0.44</b>	7.33	95.85	125.71	<b>156.61</b>	<b>237.34</b>	<b>504.89</b>
		MC	127.78	62.05	1.38	4.25	<b>3.75</b>	<b>85.78</b>	<b>118.42</b>	156.91	331.79	841.68
	Communication	BM	8.10	10.97	<b>4.83</b>	<b>43.31</b>	1.30	2.88	4.09	8.26	56.05	402.42
		SC	3.14	<b>1.79</b>	28.52	1372.51	<b>1.16</b>	2.53	2.91	3.42	<b>6.59</b>	<b>136.17</b>
		MC	<b>3.12</b>	3.20	32.91	1559.51	1.27	<b>2.45</b>	<b>2.90</b>	<b>3.32</b>	6.79	298.26
	Computation	BM	52.77	24.39	3.60	56.22	<b>8.56</b>	37.68	<b>52.19</b>	<b>61.28</b>	124.27	623.96
		SC	55.46	<b>17.30</b>	<b>0.02</b>	<b>0.91</b>	13.41	42.75	59.74	65.80	<b>104.70</b>	<b>175.43</b>
		MC	<b>47.83</b>	20.71	0.11	0.36	9.05	<b>32.34</b>	53.72	62.10	105.60	244.08
aarch64	E2E	BM	258.65	71.42	3.71	28.26	120.21	210.70	<b>242.14</b>	301.10	537.49	1389.36
		SC	256.65	<b>46.41</b>	<b>0.60</b>	<b>0.85</b>	144.74	221.71	251.80	284.49	<b>389.62</b>	<b>606.88</b>
		MC	<b>244.11</b>	66.08	1.39	11.44	<b>55.02</b>	<b>202.66</b>	243.77	<b>282.81</b>	405.38	1168.96
	Idle	BM	112.76	54.70	2.09	14.98	<b>6.80</b>	74.95	101.85	151.29	269.51	981.29
		SC	<b>97.75</b>	<b>44.85</b>	<b>0.68</b>	<b>0.95</b>	7.94	<b>64.44</b>	<b>94.20</b>	<b>126.06</b>	<b>223.01</b>	<b>448.31</b>
		MC	117.89	56.62	3.54	35.67	7.01	83.09	112.66	146.26	258.24	1084.05
	Communication	BM	10.67	14.82	<b>10.56</b>	<b>162.57</b>	4.69	6.49	<b>7.18</b>	9.35	69.42	468.11
		SC	8.32	4.27	13.95	279.49	5.61	6.86	7.50	8.64	18.19	<b>146.23</b>
		MC	<b>7.59</b>	<b>3.19</b>	13.08	312.38	<b>3.47</b>	<b>6.35</b>	7.20	<b>8.08</b>	<b>15.20</b>	148.73
	Computation	BM	135.23	27.84	6.85	73.21	90.90	123.59	133.65	<b>140.63</b>	245.06	704.96
		SC	150.58	<b>14.54</b>	1.11	7.64	107.28	141.23	151.91	158.52	199.52	324.74
		MC	<b>118.63</b>	36.90	<b>-0.61</b>	<b>-0.41</b>	<b>24.66</b>	<b>87.86</b>	<b>131.09</b>	147.71	<b>182.57</b>	<b>322.10</b>

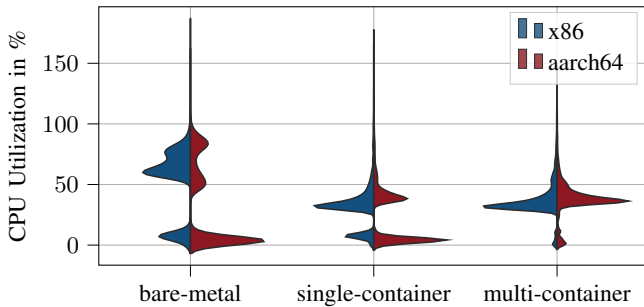


Fig. 5: CPU utilization of the systems for the different deployment variants.

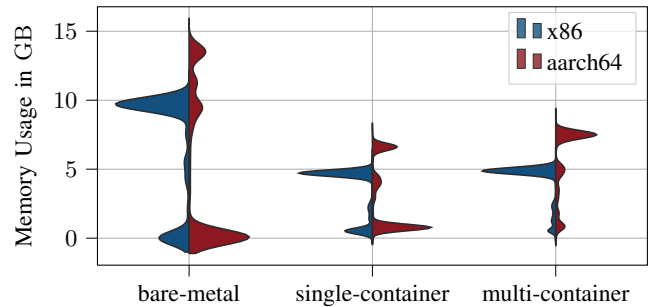


Fig. 6: Memory utilization of the systems for the different deployment variants.

and compare it with the bare-metal execution and the execution within a single container. We further split the end-to-end latency into its components (Idle, DDS Communication, Computation)

to gain a better understanding of each contribution. Fig. 4 shows histograms of the different deployment variants on the respective compute platforms. Table VI presents the detailed

measurement values for the entire experiment. As explained in Section IV-D the latencies are shown for one computation chain from the sensor to the control output.

**E2E Latency.** The histograms of bare-metal deployments (Fig. 4(a)) show a bimodal distribution with ( $x86$ ) high kurtosis value of 19.94 and a long tail visible in the Q75 and P99 values (Table VI). A similar behavior is evident on the  $aarch64$  system. Instead, for single-containers, we see a reduced standard deviation and a significantly lower kurtosis of only 0.35. This is reflected in the histogram that show a compact distribution.

The bare-metal implementation of Autoware on the  $x86$  platform shows an end-to-end latency of 194.67 ms. Instead, for the developed microservice architecture, the mean latency is reduced by 8.1% to 178.73 ms for  $x86$  and by 5.6% to 244.11 ms for  $aarch64$ .

A relatively large maximum value with 1437.94 ms is measured for bare-metal. In the single-container scenario, this maximum value drops to only 553.79 ms. We also see a reduction in the various quantiles. From 515.87 ms in the 99th percentile to 300.79 ms, which is an improvement of 41.7%.

**Idle Latency.** Looking at the idle latency, which describes the time data waits for processing via a timer callback, we observe the following for  $x86$ . The native deployment shows an idle latency of 133.81 ms. Experiments on  $aarch64$  show lower mean (112.76 ms) and maximum values. Executing Autoware in a single-container environment eliminates the two peaks, resulting in a more even distribution for both  $x86$  and  $aarch64$ . The mean latency value decreases to 125.92 ms for  $x86$  and 97.75 ms for  $aarch64$ , indicating improvements of 5.9% and 13.3%, respectively. This improvement extends to quantile values as well. The 99th percentile sees an improvement of 41.6% to 237.34 ms for  $x86$  and 17.3% to 223.01 ms for  $aarch64$ . In the multi-container deployment, the mean idle latency is higher than that of single containers, but still lower than the bare-metal setup for  $x86$ . Conversely, a higher measurement value compared to bare-metal is observed for  $aarch64$  (4.5%). For  $x86$ , quantiles 25 and 50 are lower compared to single containers, but higher for  $aarch64$ . Both systems exhibit increased values for P99 and the maximum.

**Communication Latency.** The communication latency represents the smallest portion of the entire end-to-end latency. On the  $x86$  in the native deployment, this latency has a mean of 8.10 ms, while on the  $aarch64$ , it has a mean of 10.666 ms. However, both systems exhibit maximum values of 402.42 ms and 468.110 ms respectively, resulting in distributions with long right tails. Notably, the distribution of the single-container variant shows a reduced tail: the mean DDS latency improves by 61.2% to 3.14 ms for  $x86$ , with a smaller improvement (22.0%) observed for  $aarch64$ . Both systems also see reductions in their maximum values. In the multi-container variant, DDS values for both systems are similar to those of the single-container setup, with reduced mean and quantile values within a negligible range compared to single-container. Only the maximum value increases slightly to 298.26 ms for  $x86$ , still smaller than in the bare-metal deployment. On the  $aarch64$ ,

the maximum value is only slightly higher than that of the single-container setup.

**Computation Latency.** The computation latency is reduced in the multi-container deployment, resulting in improved mean values (47.83 ms for  $x86$  and 118.63 ms for  $aarch64$ ). These improvements lead to lower values than those observed in both the bare-metal and single-container scenarios. However, we note increased values for P99 and the maximum on the  $x86$  variant. However, the mean values diverge significantly, and the  $aarch64$  system experiences higher values compared to before. The results deviate from a normal distribution, as no clear peak is visible, but rather multiple peaks in both cases.

**CPU and Memory Utilization.** Fig. 5 shows the distributions of CPU usage measured over all runs, *i.e.*, from the execution of the ROS 2 launch file until the vehicle reaches the target position. We observe the same behavior noted by [24] regarding the startup of the software. With bare-metal and single-container, the mean ramp-up phase of the software lasted 13.60 s and 14.4 s, on the  $x86$ . On the  $aarch64$  it is 21.3 s and 23.1 s. Multi-container has a much lower ramp-up phase, where all nodes of the software startup the fastest, 3.8 s for  $x86$  and 4.6 s for  $aarch64$  respectively. This ramp-up phase is clearly visible in both of the plots in the lower part of the graph. As visible in Fig. 5 and Fig. 6, containerized applications, regardless of whether single or multi-container, have a lower CPU and memory utilization. Further analyzing CPU utilization, a significant scatter is observed for bare-metal, ranging from approximately 50% to 90% after node initialization. The variance is notably lower for the single-container setup, with a slight difference in utilization, where  $aarch64$  exhibits slightly higher values compared to  $x86$ . Regarding memory consumption, bare-metal deployments on  $x86$  cluster at around 10 GB RAM, whereas  $aarch64$  displays higher memory consumption with a significantly higher variance. Once again, transitioning the application to a container environment, whether single- or multi-container, leads to a reduction in memory consumption by almost a factor of two for  $x86$ . Similarly, on  $aarch64$ , a drastic decrease in memory consumption is observed. However, both container variants still exhibit higher memory consumption compared to the  $x86$  platform, as seen with bare-metal deployment.

## VI. DISCUSSION

The results of our research uncover unexpected insights into the performance of containerized applications, particularly with respect to end-to-end latency and system utilization. Our results suggest that applications deployed in a container environment have a better latency compared to applications running directly on bare-metal. In the real-world application, end-to-end latency improvements of up to 5.2% were achieved. The developed microservice architecture showed an improvement of 5-8% in the mean. For the maximum values, it was apparent that the single-container had significantly reduced max values. The DDS Communication benchmark showed that for smaller message sizes, containerization produced better results. This margin was considerably lower (almost absent) in the `ros2_benchmark`.

However, in this benchmark, containerization-related jitter was lower than bare-metal.

To better understand the root causes of such behaviors, we have performed several attempts to optimize the bare-metal Linux system to achieve better results than containers. However, the complexity of the applications considered and their internal interaction is so high that it was not possible to have all parameters under control. We tried to improve the latency with different real-time scheduling algorithms and patches. Nevertheless, the Autoware software starved when we utilized the entire cores for the software. Reserving resources for the Linux processes led to a higher latency compared to the presented results.

At their core, containers leverage kernel parameters and settings to isolate processes using namespaces and cgroups. Achieving better performance on bare metal typically involves tuning kernel parameters and settings. Isolation is likely a primary factor contributing to the improved performance of containers in our complex setup. By isolating processes, containers ensure that standard Linux processes do not interfere with those inside the container, thereby facilitating an optimized execution environment.

Interestingly, our results showed that deploying applications in multiple containers enhances the improvements of single container configurations, particularly for average end-to-end latencies. This implies that distributing workloads across multiple containers can optimize the overall system performance, particularly in terms of latency. However, this approach can also result in significantly higher maximum execution times. Such trade-offs must be carefully considered when designing and optimizing a system, especially in real-time environments where small maximum execution time is critical.

A critical factor in this discussion is the role of cgroup scheduling and task assignment to cores within the Linux CFS. Platforms like Kubernetes and Docker use this mechanism to effectively schedule container workloads. *Cshares*, a core component of this system, are influenced by various parameters such as predefined CPU limits and the number of processes or threads within a container. The CFS then allocates resources to *Cshares*, determining how resources are distributed among containers. One of the key advantages of this system is the relative isolation it offers. In a native system, without the protective containerization layer, processes could inadvertently impact each other. For instance, native processes could affect the performance of the Autoware software within the CFS scheduler. Containerization effectively segregates processes, ensuring that each operates within its own domain and remains unaffected by external entities. The inherent mechanisms of cgroup scheduling and its relation with the Linux CFS could play a crucial role in these results.

Another positive effect is that containerization improves latencies by increasing second-order effects such as the locality of data by grouping related tasks on a smaller set of cores, preventing unregulated migrations to distant cores (our systems have 16 and 32 cores respectively). Unregulated migration could be the cause of the bimodal distribution observed for end-to-end latency in Fig. 4(a). However, a number of experiments and testbeds were implemented to confirm this. This reason

could not be confirmed as the sole cause of the performance behavior. Our study also showed another interesting trend. As the complexity of the test cases increased, the number of processes or threads working within the container also increased significantly. This suggests that as the complexity increases, the container environment becomes more densely populated with processes and threads to handle the increased requirements.

In summary, our exploration of the containerized applications domain has confirmed the potential benefits of such an approach, not only in terms of isolation but also in terms of performance optimization. This is also observed in the paper [24], where investigations of the start-up time of nodes to complete launch coincide with our observation of runtime.

## VII. CONCLUSION

We presented a microservice architecture tailored to an open-source software for autonomous driving. Our study provided a comprehensive overview of the continuous integration and development process associated with this architecture. We analyzed multiple metrics for a real-world ROS 2 autonomous driving application based on Autoware and deployed on increasingly isolated container environments. In order to determine the impact of containerization on communication and simple ROS 2 examples, the analysis was complemented with dedicated benchmarks for DDS and ROS 2.

Our findings indicate that the effect of containerization on runtime varies depending on the complexity of the scenario. In simpler scenarios, the impact of containerization was relatively minor, but, in more complex scenarios, such as that of Autoware, the influence—especially on end-to-end latency—was significant. Moreover, both CPU and memory usage were reduced, leading to improved software stability. These effects were observed and validated on two distinct systems: *x86* and *aarch64* compute platforms. This cross-system analysis enhances the generalizability of our results.

While our study shows the positive impact of containerization, the complex interactions between containers, Linux CFS, cgroups and Autoware framework require more detailed investigation to determine the exact contributions of each of the mechanisms involved. However, to our knowledge, this work is the first to provide such in-depth insights into complex real-world autonomous driving setups and highlights the need for more detailed studies in the future.

Looking ahead, there are several directions for further research. One is to explore strategies to optimize node assignment to containers and the impact of static container allocations to CPUs and setting bounds on CPU shares. Another interesting topic is hierarchical scheduling, which should be explored in depth to improve the performance of containerized ROS 2 applications. Furthermore, it is worth considering the generalizability of these results beyond ROS 2 applications.

## ACKNOWLEDGEMENTS

T. Betz, as the first author, was the initiator of the research idea and is responsible for the presented concept and implementation. L. Wen, F. Pan, and A. Knoll contributed to implementation and design of the benchmarks. G. Kaljavesi contributed to the

implementation of the microservice architecture. A. Zuepke, A. Bastoni, and M. Caccamo contributed to the evaluation of the performance impacts and the design of experiments. J. Betz contributed to the conception of the research project and revised the paper critically for important intellectual content. He gave final approval of the version to be published and agrees to all aspects of the work. As a guarantor, he accepts responsibility for the overall integrity of the paper. M. Caccamo was supported by an Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research.

## REFERENCES

- [1] SOAFEE, "SOAFEE: Scalable open architecture for embedded edge," 2022. [Online]. Available: <https://www.soafee.io>
- [2] M. Spencer, "How the SOAFEE architecture brings a cloud-native approach to mixed critical automotive systems," *white paper*, Sept. 2021.
- [3] S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. E. Haque, L. Tang, and J. Mars, "The architectural implications of autonomous driving," in *Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2018, pp. 751–766.
- [4] T. Betz, P. Karle, F. Werner, and J. Betz, "An analysis of software latency for a high-speed autonomous race car—a case study in the indy autonomous challenge," *SAE Int. Journal of Connected and Automated Vehicles*, vol. 6, no. 12-06-03-0018, 2023.
- [5] The Autoware Foundation, "Autoware - the world's leading open-source software project for autonomous driving." [Online]. Available: <https://github.com/autowarefoundation/autoware>
- [6] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm6074, 2022.
- [7] P. Karle, T. Betz, M. Bosk, F. Fent, N. Gehrke, M. Geisslinger, L. Gressenbuch, P. Hafemann, S. Huber, M. Hübner *et al.*, "Edgar: An autonomous driving research platform—from feature development to real-world application," *arXiv preprint arXiv:2309.15492*, 2023.
- [8] J. Arundel and J. Domingus, *Cloud Native DevOps with Kubernetes: building, deploying, and scaling modern applications in the Cloud*. O'Reilly Media, 2019.
- [9] Rancher Labs, "K3s - lightweight kubernetes." [Online]. Available: <https://github.com/k3s-io/k3s/>
- [10] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.
- [11] E. Sax, R. Reussner, H. Guissouma, and H. Klare, *A survey on the state and future of automotive software release and configuration management*. KIT Amsterdam, The Netherlands, 2017.
- [12] W. Haas and P. Langjahr, "Cross-domain vehicle control units in modern e/e architectures," in *Int. Stuttgarter Symposium: Automobil- und Motorentechnik*, 2016, pp. 1619–1627.
- [13] S. Kugele, D. Hettler, and S. M. Shafaei, "Elastic service provision for intelligent vehicle functions," *Int. Conf. on Intelligent Transportation Systems (ITSC)*, pp. 3183–3190, 2018.
- [14] J. Lotz, A. Vogelsang, O. Benderius, and C. Berger, "Microservice architectures for advanced driver assistance systems: A case-study," in *IEEE Int. Conf. on Softw. Archit. Companion (ICSA-C)*, 2019, pp. 45–52.
- [15] G. T. B. Tamanaka, R. V. Aroca, and G. A. de Paula Caurin, "Fault-tolerant architecture and implementation of a distributed control system using containers," in *LARS/SBR/WRE*, 2022, pp. 1–6.
- [16] A. Brogi, D. Neri, J. Soldani, and O. Zimmermann, "Design principles, architectural smells and refactorings for microservices: a multivocal review," *SICS Softw.-Intensive Cyber-Physical Systems*, pp. 3–15, 2019.
- [17] N. Kukulicic, D. Samardzic, A. Bucaioni, and S. Mubeen, "Automotive service-oriented architectures: a systematic mapping study," in *Euromicro Conf. on Softw. Engin. and Adv. Applications (SEAA)*, 2022, pp. 459–466.
- [18] V. Velepucha and P. Flores, "Monoliths to microservices - Migration Problems and Challenges: A SMS," in *Int. Conf. on Information Systems and Softw. Technologies (ICI2ST)*, 2021, pp. 135–142.
- [19] S. Giallorenzo, J. Mauro, M. G. Poulsen, and F. Siroky, "Virtualization costs: benchmarking containers and virtual machines against bare-metal," *SN Computer Science*, vol. 2, no. 5, p. 404, 2021.
- [20] R. Morabito, "Virtualization on internet of things edge devices with container technologies: A performance evaluation," *IEEE Access*, vol. 5, pp. 8835–8850, 2017.
- [21] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and linux containers," in *IEEE Int. Symp. on Performance Analysis of Systems and Software (ISPASS)*, 2015, pp. 171–172.
- [22] M. G. Xavier, M. V. Neves, F. D. Rossi, T. C. Ferreto, T. Lange, and C. A. De Rose, "Performance evaluation of container-based virtualization for high performance computing environments," in *Euromicro Int. Conf. on Parallel, Distributed, and Network-Based Processing (PDP)*, 2013, pp. 233–240.
- [23] A. K. S. Rajan, A. Feucht, L. Gamer, I. Smaili *et al.*, "Hypervisor for consolidating real-time automotive control units: Its procedure, implications and hidden pitfalls," *J. Syst. Archit.*, vol. 82, pp. 37–48, 2018.
- [24] L. Wen, M. Rickert, F. Pan, J. Lin, and A. Knoll, "Bare-metal vs. hypervisors and containers: Performance evaluation of virtualization technologies for software-defined vehicles," in *IEEE Intelligent Vehicles Symp. (IEEE IV)*, Jun 2023.
- [25] M. Reke, D. Peter, J. Schulte-Tigges, S. Schiffer, A. Ferrein, T. Walter, and D. Matheis, "A self-driving car architecture in ros2," in *Int. SAUPEC/RobMech/PRASA Conf.*, 2020, pp. 1–6.
- [26] J. Betz, T. Betz, F. Fent, M. Geisslinger, A. Heilmeyer, L. Hermansdorfer, T. Herrmann, S. Huch, P. Karle, M. Lienkamp *et al.*, "TUM autonomous motorsport: An autonomous racing software for the indy autonomous challenge," *Journal of Field Robotics*, vol. 40, no. 4, pp. 783–809, 2023.
- [27] Z. Li, A. Hasegawa, and T. Azumi, "Autoware\_Perf: A tracing and performance analysis framework for ROS 2 applications," *J. Syst. Archit.*, vol. 123, p. 102341, 2022.
- [28] T. Kuboichi, A. Hasegawa, B. Peng, K. Miura, K. Funaoka, S. Kato, and T. Azumi, "CARET: Chain-Aware ROS 2 Evaluation Tool," in *IEEE Int. Conf. on Embedded and Ubiquitous Computing (EUC)*, 2022.
- [29] T. Betz, M. Schmeller, A. Korb, and J. Betz, "Latency measurement for autonomous driving software using data flow extraction," in *IEEE Intelligent Vehicles Symp. (IEEE IV)*, 2023.
- [30] T. Bläß, A. Hamann, R. Lange, D. Ziegenbein, and B. B. Brandenburg, "Automatic Latency Management for ROS 2: Benefits, Challenges, and Open Problems," in *IEEE Real-Time and Embedded Technology and Applications Symp. (RTAS)*, 2021, pp. 264–277.
- [31] C. Bédard, I. Lütkebohle, and M. Dagenais, "ros2\_tracing: Multipurpose Low-Overhead Framework for Real-Time Tracing of ROS 2," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6511–6518, 2022.
- [32] Apex.AI, "performance\_test." [Online]. Available: [https://gitlab.com/ApexAI/performance\\_test](https://gitlab.com/ApexAI/performance_test)
- [33] "iRobot: ROS2 performance," 2021. [Online]. Available: <https://github.com/irobot-ros/ros2-performance>
- [34] "Nvidia-isaac-ros ros2\_benchmark," 2023. [Online]. Available: [https://github.com/NVIDIA-ISAAC-ROS/ros2\\_benchmark](https://github.com/NVIDIA-ISAAC-ROS/ros2_benchmark)
- [35] H. Teper, M. Günzel, N. Ueter, G. von der Brügggen, and J.-J. Chen, "End-to-end timing analysis in ros2," in *IEEE Real-Time Systems Symp. (RTSS)*, 2022, pp. 53–65.
- [36] S. Kato, S. Tokunaga, Y. Maruyama, S. Maeda, M. Hirabayashi, Y. Kitsukawa, A. Monroy, T. Ando, Y. Fujii, and T. Azumi, "Autoware on board: Enabling autonomous vehicles with embedded systems," in *ACM/IEEE Int. Conf. on Cyber-Physical Systems (ICCP)*, 2018.
- [37] T. Betz, M. Schmeller, H. Teper, and J. Betz, "How Fast is My Software? Latency Evaluation for a ROS 2 Autonomous Driving Software," in *IEEE Intelligent Vehicles Symp. (IEEE IV)*, 2023.
- [38] T. Kronauer, J. Pohlmann, M. Matthé, T. Smejkal, and G. Fettweis, "Latency analysis of ros2 multi-node systems," in *IEEE Int. Conf. on Multisensor Fusion and Integration for Intell. Syst. (MFI)*, 2021, pp. 1–7.
- [39] T. Wu, B. Wu, S. Wang, L. Liu, S. Liu, Y. Bao, and W. Shi, "Oops! It's Too Late. Your Autonomous Driving System Needs a Faster Middleware," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7301–7308, 2021.
- [40] ADLINK, "SOAFEE for software defined vehicles," 2022. [Online]. Available: <https://www.adlinktech.com/en/soafee>
- [41] Eclipse Foundation, "Eclipse Cyclone DDS," 2022. [Online]. Available: <https://cyclonedds.io>
- [42] Indy Autonomous Challenge, 2021. [Online]. Available: <https://www.indyautonomouschallenge.com/>
- [43] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016, pp. 4193–4198.
- [44] M. Günzel, H. Teper, K.-H. Chen, G. von der Brügggen, and J.-J. Chen, "On the equivalence of maximum reaction time and maximum data age for cause-effect chains," in *Euromicro Conf. on Real-Time Systems (ECRTS)*, 2023.