

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

## The Journal of Systems &amp; Software

journal homepage: [www.elsevier.com/locate/jss](http://www.elsevier.com/locate/jss)

# Benchmarking scalability of stream processing frameworks deployed as microservices in the cloud<sup>☆</sup>

Sören Henning<sup>a,\*</sup>, Wilhelm Hasselbring<sup>b</sup><sup>a</sup> JKU/Dynatrace Co-Innovation Lab, LIT CPS Lab, Johannes Kepler University Linz, Linz, Austria<sup>b</sup> Software Engineering Group, Kiel University, 24098 Kiel, Germany

## ARTICLE INFO

## Keywords:

Stream processing  
Microservices  
Benchmarking  
Scalability

## ABSTRACT

**Context:** The combination of distributed stream processing with microservice architectures is an emerging pattern for building data-intensive software systems. In such systems, stream processing frameworks such as Apache Flink, Apache Kafka Streams, Apache Samza, Hazelcast Jet, or the Apache Beam SDK are used inside microservices to continuously process massive amounts of data in a distributed fashion. While all of these frameworks promote scalability as a core feature, there is only little empirical research evaluating and comparing their scalability.

**Objective:** The goal of this study is to obtain evidence about the scalability of state-of-the-art stream processing framework in different execution environments and regarding different scalability dimensions.

**Method:** We benchmark five modern stream processing frameworks regarding their scalability using a systematic method. We conduct over 740 h of experiments on Kubernetes clusters in the Google cloud and in a private cloud, where we deploy up to 110 simultaneously running microservice instances, which process up to one million messages per second.

**Results:** All benchmarked frameworks exhibit approximately linear scalability as long as sufficient cloud resources are provisioned. However, the frameworks show considerable differences in the rate at which resources have to be added to cope with increasing load. There is no clear superior framework, but the ranking of the frameworks depends on the use case. Using Apache Beam as an abstraction layer still comes at the cost of significantly higher resource requirements regardless of the use case. We observe our results regardless of scaling load on a microservice, scaling the computational work performed inside the microservice, and the selected cloud environment. Moreover, vertical scaling can be a complementary measure to achieve scalability of stream processing frameworks.

**Conclusion:** While scalable microservices can be designed with all evaluated frameworks, the choice of a framework and its deployment has a considerable impact on the cost of operating it.

## 1. Introduction

Over the last decade, microservices became a frequently adopted software architecture pattern for building scalable, cloud-native software systems (Hasselbring and Steinacker, 2017; Soldani et al., 2018; Fritzsche et al., 2019; Knoche and Hasselbring, 2019). More recently, a shift toward adopting distributed stream processing techniques in microservice architectures can be observed (Laigner et al., 2021). In such systems, different microservices communicate with each other through asynchronous messages, which are sent via scalable messaging systems such as Apache Kafka (Bellemare, 2020). Especially data-intensive applications and big data analytics systems are increasingly designed as

microservices building upon frameworks that process continuous data streams in a scalable manner (Davoudian and Liu, 2020).

Combining microservice architectures and large-scale stream processing leads to a new way of operating stream processing frameworks. Whereas traditional big data systems rely on a single, heavyweight stream processing platform, which runs several data analysis jobs, microservice architectures allow running stream processing frameworks inside individual microservices, embedded as a library. This allows for choosing a suitable, usually more lightweight stream processing framework for each microservice. From an operational perspective, such microservices are deployed and scaled using established microservice orchestration tooling such as Kubernetes. Internally, the stream

<sup>☆</sup> Editor: Raffaella Mirandola.

\* Corresponding author.

E-mail addresses: [soeren.henning@jku.at](mailto:soeren.henning@jku.at) (S. Henning), [hasselbring@email.uni-kiel.de](mailto:hasselbring@email.uni-kiel.de) (W. Hasselbring).

<https://doi.org/10.1016/j.jss.2023.111879>

Received 27 February 2023; Received in revised form 18 June 2023; Accepted 12 October 2023

Available online 24 October 2023

0164-1212/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

processing frameworks perform the necessary coordination among microservice instances to (re)partition data streams for state locality and, thus, enable scalability.

While there are several open-source stream processing frameworks promoting scalability as a core feature, there is only little empirical research evaluating and comparing their scalability. However, the need for systematic scalability evaluations has been recognized (van Dongen and van den Poel, 2020; Hesse et al., 2021). Benchmarking is a well-established method in software engineering to assess and compare the quality of software systems and services (Hasselbring, 2021). As such, it is used both in research and engineering (Kounev et al., 2020) to choose among competing software solutions, to evaluate the quality of new ones, or to assure quality levels over time.

In previous work, we presented and empirically evaluated a scalability benchmarking method for cloud-native applications in general (Henning and Hasselbring, 2022a) and stream processing frameworks in particular (Henning and Hasselbring, 2021a). Further, we presented specific benchmarks for stream processing frameworks (Henning and Hasselbring, 2021b). In this work, we use the presented benchmarking method and the benchmarks to experimentally evaluate the scalability of stream processing frameworks, particularly suited to be used within microservices. Specifically, we address the following research questions:

**RQ1: How do different stream processing frameworks deployed as microservices compete regarding their scalability?** Scalability is a main driver for adopting microservices and stream processing-based architectures. The choice of a stream processing framework is thus likely to have a crucial impact on the scalability of the overall system.

**RQ2: Can the observed performance limitations of Apache Beam's abstraction layer be overcome with recently proposed performance optimization configurations?** Our results show that Apache Beam scales with significantly higher resource demands, which has also been reported in related work for an older version of the framework (Hesse et al., 2019). Since a couple of configuration settings have been proposed for improved performance, we evaluate whether such configurations can improve scalability.

**RQ3: How do stream processing frameworks scale with increasing computational work performed inside the microservice?** The most common way to assess scalability is by evaluating how a system can handle increasing external load as addressed by RQ1. Likewise, however, it can be important to assess how a framework scales with increasing the complexity of the performed computation, for example, to improve the quality of its results.

**RQ4: Can vertical scaling be a complementary measure to achieve scalability of stream processing frameworks?** A common way to scale microservices is to increase the number of instances deployed, which may require additional underlying cluster nodes (horizontal scaling). Within the bounds of the individual cluster nodes, microservices can also be scaled by increasing their provided resources, such as CPU cores or memory (vertical scaling). We benchmark both alternatives regarding their scalability on a single large node.

**RQ5: Do scalability results differ between public and private clouds?** Public clouds and private clouds are both representative environments for operating microservice-based systems. We evaluate whether our scalability benchmark results differ between both environments.

**RQ6: Can scalability limits be raised by using larger clusters?** To investigate whether observed scalability limits are due to the frameworks or just due to high utilization of the underlying cluster resources, we repeat selected experiments with clusters of different sizes.

We address these research questions by conducting over 740 h of scalability experiments on Kubernetes clusters in Google Cloud and in a private cloud environment. We benchmark the frameworks Apache Flink, Apache Kafka Streams, Hazelcast Jet, and Apache Beam with the Flink and the Samza runners, for which we deploy up to 110 simultaneously running instances, which process up to one million messages per second. We provide a replication package and the collected data of all experiments as supplemental material (Henning and Hasselbring, 2023), allowing other researchers to repeat and extend our work.

*Outline.* The remainder of this work starts by discussing the relation of microservices and stream processing as well as the foundations of scalability benchmarking in Section 2. Section 3 introduces the stream processing frameworks benchmarked in this study and Section 4 gives an overview of the cloud environments for our experiments. Section 5 describes our experimental setup and Section 6 presents and discusses the results of our experiments. Afterward, Section 7 discusses threats to validity, followed by a discussion of related work in Section 8. Finally, Section 9 concludes this work.

## 2. Background

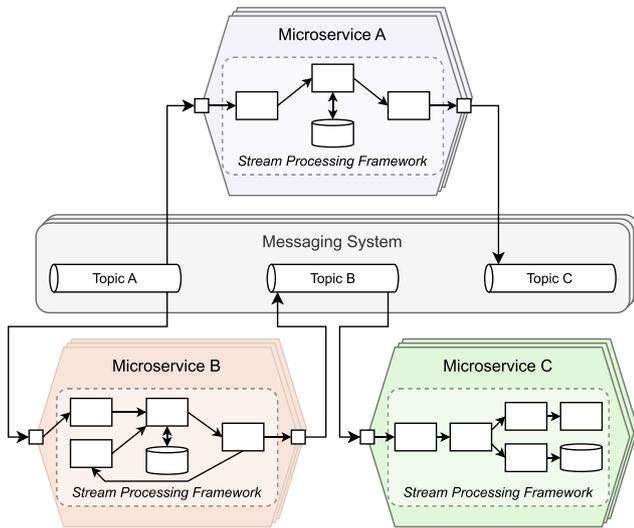
In the following, we first discuss the concepts of stream processing within microservice architectures before outlining our Theodolite scalability benchmarking method used for this research.

### 2.1. Stream processing within microservice architectures

At the time of this research, combining microservice architectures and distributed stream processing is covered only superficially in scientific literature. While there are a couple of case studies reporting on stream processing-based microservices, research is still lacking a systematic evaluation of this new architectural style. On the other hand, some textbooks for practitioners (Bellemare, 2020; Stopford, 2018) have recently been published, which also serve as references for this work. Despite the lack of systematic studies, stream processing within microservices is named an emerging trend (Davoudian and Liu, 2020; Fragkoulis et al., 2023; Karabey Aksakalli et al., 2021) and the need for further research on this topic is recognized (Katsifodimos and Fragkoulis, 2019; Laigner et al., 2021). Fig. 1 shows an exemplary architecture with microservices applying stream processing frameworks.<sup>1</sup> The following attributes can often be observed for such microservices and are particularly relevant to this study.

*Self-contained and loosely coupled.* In microservice architectures, a software system is composed of multiple small services that are built around business capabilities (Hasselbring, 2018). Individual microservices run in their own processes, may use different technology stacks, and communicate via lightweight, fault-tolerant mechanisms over the network. For microservices employing distributed stream processing techniques, this means each service can use its own stream processing framework. In contrast to traditional big data stream processing systems running on top of resource management systems such as Apache Hadoop YARN or Apache Mesos, this also leads to smaller jobs and a single job per stream processing cluster.

<sup>1</sup> In practice, one can typically observe a combination of different communication methods, such as additional HTTP-based communication (Karabey Aksakalli et al., 2021). Depending on the microservice-specific use case, such architectures may also include microservices relying on other mechanisms to process data streams, such as simple producers/consumers or Function-as-a-Service deployments (Bellemare, 2020).



**Fig. 1.** Example microservice architecture of a software system, in which microservices communicate via asynchronous messages and process these messages using stream processing frameworks. With such frameworks, the dataflow inside microservices is modeled as directed graphs of (potentially stateful) operators. In this study, we benchmark different state-of-the-art stream processing frameworks with task samples, representing microservices such as those depicted in this example architecture.

**Independently scalable.** In contrast to monolithic systems, individual microservices can be scaled independently due to their loose coupling. In fact, scalability has been reported as one of the most important drivers for and benefits of adopting microservice architectures in several systematic literature reviews (Pahl and Jamshidi, 2016; Li et al., 2021; Kratzke and Quint, 2017; Soldani et al., 2018; Laigner et al., 2021), interview studies (Taibi et al., 2017; Fritzsche et al., 2019; Knoche and Hasselbring, 2019; Laigner et al., 2021; Zhou et al., 2023), and experience reports (Balalaie et al., 2016b; Hasselbring and Steinacker, 2017; Bucchiarone et al., 2018). Typically, horizontal duplication, data partitioning, and function decomposition are distinguished as methods for scaling microservice architectures. From an operation’s perspective, also microservices adopting stream processing frameworks can be scaled by horizontal duplication. Internally, however, duplicating services leads to data partitioning, meaning that each instance of a service handles only messages with certain keys. The actual data partitioning as well as the necessary state management and fault tolerance is managed by modern stream processing frameworks.

**Cloud-native deployment.** Microservice architectures are a pattern, particularly suited for building cloud-native applications (Balalaie et al., 2016a; Gannon et al., 2017; Pahl et al., 2018). They are mostly deployed as containers in public or private cloud environments, with Kubernetes (Burns et al., 2016) being the de-facto standard orchestration tool for cloud-native applications (Cloud Native Computing Foundation, 2022). With Kubernetes, microservice deployments including, for example, resource restrictions or numbers of replicas, are defined purely declaratively.

**Asynchronous communication via messaging system.** Microservices that process and consume streams of messages often employ log-based messaging systems for their communication Karabey Aksakalli et al. (2021). To eventually reach consistency among individual microservices, the log must be durable, append-only, fault-tolerant, partitioned, and it must support sequential reads (Kleppmann et al., 2019). Probably the most prominent messaging system fulfilling these properties is Apache Kafka (Kreps et al., 2011; Wang et al., 2015), which is intensively used in industry.<sup>2</sup> Log-based messaging systems such as Kafka are required

by modern stream processing frameworks to provide strict processing guarantees and fault tolerance while scaling out (Fragkoulis et al., 2023).

**Adoption of distributed stream processing techniques.** Modern stream processing frameworks are designed to run in a distributed fashion on commodity hardware in order to scale with massive amounts of data (Fragkoulis et al., 2023). Besides high throughput, these systems focus on low latency, fault tolerance, and coping with out-of-order streams. Modern stream processing frameworks process data in jobs, where a job is defined as a dataflow graph of processing operators. They can be started with multiple instances (e.g., on different compute nodes, containers, or with multiple threads). For each job, each instance processes only a portion of the data. Whereas isolated processing of data records is not affected by the assignment of data portions to instances, processing that relies on previous data records (e.g., aggregations over time windows) requires the management of state. Similar to the MapReduce programming model, keys are assigned to records and the stream processing frameworks guarantee that all records with the same key are processed by the same instance. Hence, no state synchronization among instances is required. When a processing operator changes the record key and a subsequent operator performs a stateful operation, the stream processing framework splits the dataflow graph into subgraphs that can be processed independently by different instances. We refer to the recent surveys of Fragkoulis et al. (2023) and Margara et al. (2022) for detailed information on state-of-the-art stream processing models and patterns.

## 2.2. Scalability benchmarking with theodolite

Our study builds upon our Theodolite scalability benchmarking method for cloud-native applications (Henning and Hasselbring, 2022a). According to the ACM SIGSOFT Empirical Standard for benchmarking as a software engineering research method (Ralph et al., 2021; Hasselbring, 2021), we briefly summarize the quality, metric, measurement method, and task samples used in this study.<sup>3</sup>

**Quality.** The quality evaluated in this study is scalability. Scalability is defined as “the ability of [a] system to sustain increasing workloads by making use of additional resources” (Herbst et al., 2013). It should not be confused with elasticity, which describes how fast or how precise a system (automatically) adapts to varying workloads (Herbst et al., 2013; Lehrig et al., 2015).

**Metric.** The Theodolite scalability benchmarking method provides two alternative scalability metrics, the *resource demand* metric and the *load capacity* metric. In this study, we focus on the *resource demand* metric. It is a function, mapping the load intensity on a system under test (SUT) to the minimal amount of resources that must be provisioned for the SUT such that the SUT fulfills all specified service-level objectives (SLOs). In Section 5.3, we describe how load, resources, and SLOs are defined in this study.

**Measurement method.** To measure scalability according to the metric, we chose discrete subsets of the load and resource domains and run isolated performance experiments for different load and resource combinations to assess whether the specified SLOs can be fulfilled. Experiment duration, warm-up periods, and the number of repetitions are configurable and have to be adjusted to the context. By using appropriate search strategies, not all combinations of load and resources have to be executed.

<sup>3</sup> <https://acmsigsoft.github.io/EmpiricalStandards/docs/?standard=Benchmarking>.

<sup>2</sup> <https://kafka.apache.org/powered-by>.

**Task samples.** Our Theodolite benchmarking method is not restricted to specific task samples, but allows also using existing benchmarks for cloud-native applications. Based on real-world use cases for Industrial Internet of Things analytics (Henning et al., 2021), we proposed four task samples of different complexity for stream processing frameworks in previous work (Henning and Hasselbring, 2021b). As we discuss in Section 8, these are the only benchmarks focusing on modern stream processing frameworks deployed as cloud-native microservices. In Section 5.2, we describe how we configure these task samples for our evaluation.

### 3. Evaluated stream processing frameworks

In their textbook, Bellemare (2020) distinguishes between lightweight and heavyweight stream processing frameworks. Lightweight frameworks are embedded as a programming library into the source code of independently deployable components such as microservices. This way, the stream processing framework does not require any specific way to build or deploy the microservice. This allows the service to also perform other tasks beyond stream processing such as providing a REST API. Individual instances of a service discover each other (e.g., via features of the messaging system or Kubernetes) and perform the necessary coordination internally. Heavyweight frameworks on the contrary are provided as deployable software systems, which can be configured by one or more stream processing jobs to be executed. They are typically designed as a master-worker architecture. In this study, we mainly focus on lightweight frameworks. All modern stream processing frameworks can be deployed containerized on commodity hardware with Kubernetes. In the following, we give a brief overview of frameworks, particularly suited to be deployed as microservices and which we decided to benchmark in this study. For a detailed comparison of the framework's features, see the works of, for example, Hesse and Lorenz (2015), Fragkoulis et al. (2023), and van Dongen (2021).

**Apache Flink.** Originating from a scientific research project (Alexandrov et al., 2014), Apache Flink (Carbone et al., 2015) has been extensively used, evaluated, and extended in research and became increasingly popular in industry. It offers one of the most elaborated dataflow models, providing precise control of time and state (Carbone et al., 2015, 2017; Akidau et al., 2021). Moreover, Flink provides different abstraction layers and a rich feature set regarding the integration with external systems. Flink clearly falls into the category of heavyweight frameworks (Bellemare, 2020). Its deployment consists of one or – for fault-tolerance – more coordinating *JobManagers* and a scalable amount of *TaskManagers*. Although heavyweight, we consider Flink in this study due to its widespread adoption and since we observe recent trends to more lightweight deployments of Flink.

**Apache Kafka Streams.** Kafka Streams (Sax et al., 2018; Wang et al., 2021) is a stream processing framework built on top of Apache Kafka. It is available as a Java library and, thus, aligns with the idea of incorporating stream processing in standalone microservices. Compared to most other stream processing frameworks, it has a restricted set of features, in particular, concerning the integration with external systems. Kafka Streams only supports Kafka topics as data sources and sinks. Integrating external systems always requires transferring data via Kafka topics.

**Apache Samza.** Similar to Kafka Streams, Apache Samza (Noghabi et al., 2017) can be embedded as a library in standalone applications. Individual instances of the same application use Apache Zookeeper and Apache Kafka for coordination, data transfer, and fault tolerance. Although still maintained, Samza is sometimes considered a predecessor of Kafka Streams (Kleppmann and Krepis, 2015). In this study, we use Samza as a runner for Apache Beam pipelines (see below), which allows implementing more complex use cases (Zhang et al., 2020).

**Hazelcast Jet.** Hazelcast Jet (Gencer et al., 2021) is a stream processing framework built on top of the Hazelcast IMDG distributed, in-memory object store. It can be embedded into Java applications and does not have any dependencies on an external system. Instead, individual instances discover each other, form a cluster, and handle coordination and data replication internally. Hazelcast Jet differs from other frameworks in its execution model, which is based on a concept similar to coroutines and cooperative threads (Gencer et al., 2021). With the release of Hazelcast 5.0 in 2021, Hazelcast Jet has been merged with Hazelcast IMDG into one unified product.

**Apache Beam.** Apache Beam is not a stream processing system by itself, but instead, an SDK to implement stream processing jobs in a uniform model, which can be executed by several modern stream processing systems. Apache Beam implements Google's Dataflow model (Akidau et al., 2015), which is also internally used by Google's cloud service *Google Cloud Dataflow*. A stream processing job implemented with Apache Beam is executed by a so-called runner. Runners can be seen as adapters for the actual stream processing systems. Besides a runner for Google Cloud Dataflow, Apache Beam provides also runners for several other systems, including the aforementioned Apache Flink, Apache Samza, and Hazelcast Jet. Previous research found that using Apache Beam as an abstraction layer comes with a significant negative impact on performance (Hesse et al., 2019).

**Other stream processing frameworks.** Apache Spark (Zaharia et al., 2016), Apache Storm (Toshniwal et al., 2014), and the successor of the latter, Apache Heron (Kulkarni et al., 2015), are considered heavyweight frameworks and, thus, fit less into the context of microservices (Bellemare, 2020). Spark differs from the frameworks discussed before in that it processes data streams in "micro-batches". Storm and Heron provide less sophisticated programming models and weaker fault tolerance mechanisms (Fragkoulis et al., 2023). Moreover, there are several cloud services available for stream processing, which, however, are out of scope of our study.

### 4. Evaluated cloud platforms

We run our experiments in a private and a public cloud, representing two common execution environments for microservice-based software systems. On both platforms, we operate Kubernetes clusters. For the public cloud infrastructure, we chose Google Cloud, representing one of the largest cloud vendors with a Kubernetes offering that provides multiple configuration options. As private cloud infrastructure, we use the Software Performance Engineering Lab (SPEL) at Kiel University. Table 1 summarizes the configuration of our Kubernetes clusters for all experiments.

**Google Cloud.** The Google Cloud clusters consist of *e2-standard* virtual machines. This machine type is promoted by Google as cost-optimized and suitable for general purposes such as microservices. In the course of our experiments, we use clusters with different numbers and sizes of such VMs. All clusters are set up and managed with the Google Kubernetes Engine (GKE) service and run in the *us-central1* region.

**Private cloud (SPEL).** In the private cloud, a self-operated Kubernetes is installed via Kubespray on five bare metal nodes, connected via 10 Gb Ethernet. Each node is equipped with two 16-core CPUs and 384 GB of memory.

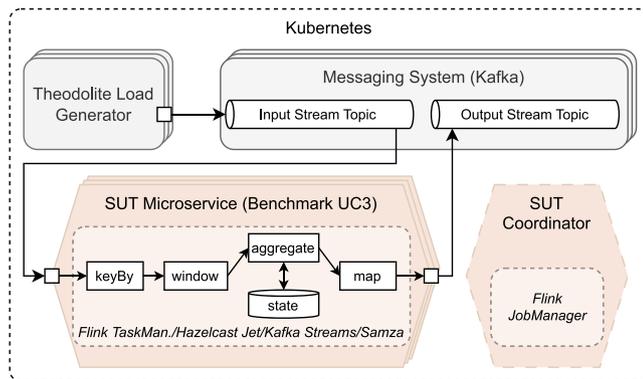
### 5. Experimental setup

We run our experiments with our Theodolite scalability benchmarking framework<sup>4</sup> (version 0.9). Theodolite is installed as a Kubernetes Operator inside the Kubernetes clusters and controls the execution of

<sup>4</sup> <https://www.theodolite.rocks/>

**Table 1**  
Configuration of Kubernetes clusters used for our evaluation, running at Google Cloud and a private cloud (SPEL).

	Sections 6.1–6.3 and 6.5	Section 6.4	Section 6.5	Section 6.6
Cloud platform	Private cloud	Private cloud	Google Cloud	Google Cloud
Nodes	5	5	5	8–12
CPU cores	2 × 16	2 × 16	32	16
RAM	384 GB	384 GB	128 GB	64 GB
Machine type	Intel Xeon Gold 6130	Intel Xeon Gold 6130	e2-standard-32	e2-standard-16
Kubernetes	1.23.7	1.23.7	1.24.11-gke.1000	1.24.11-gke.1000
Kafka brokers	5	3	5	4



**Fig. 2.** Benchmark deployment illustrated for the case of benchmark UC3. Our Theodolite load generator sends messages with a constant, but configurable frequency to the input stream topic in Kafka. From there, our SUT microservice consumes the messages, aggregates them and sends the results to the output stream topic in Kafka. The actual processing inside the SUT microservice is implemented with a stream processing framework. For each evaluated stream processing framework, we implemented dedicated microservice. The implementation of Apache Flink requires an additional coordinating instance running the Flink JobManager. The deployment for the other task samples look similar, however, with different dataflow architectures (Henning and Hasselbring, 2021b) and potentially different input and output streams.

benchmarks according to our scalability benchmarking method (Henning and Hasselbring, 2022b). Unless stated differently, we run 5 Kafka brokers, one on each node, with Kafka version 3.2 and 100 partitions per Kafka topic. All implementations of the benchmarks and the benchmarking tool itself are available as open-source software.<sup>5</sup> In our replication package (Henning and Hasselbring, 2023), we provide the declarative Theodolite files used for executing the benchmarks and the collected data of all experiments along with analysis scripts, allowing other researchers to repeat and extend our work. In the following, we summarize the configuration of the benchmarked stream processing frameworks, the selected benchmark task samples, and the benchmarking method. Fig. 2 illustrates our benchmark deployment.

### 5.1. Configuration of stream processing frameworks

We benchmark the stream processing frameworks Apache Beam (version 2.35) with the Flink (version 1.13) and the Samza runner (version 1.5), Apache Flink (version 1.13), Hazelcast Jet (version 4.5), and Apache Kafka Streams (version 3.1). For a fair comparison, we evaluate all frameworks with mostly their default configuration.

We enable committing read offsets to Kafka in all frameworks. This allows us to monitor the consumer lag via Kafka metrics, which is required to evaluate our lag trend SLO. Enabling offset committing is also often done in production deployments to increase observability. We set the commit interval to 5 s for all frameworks, which is the default configuration of Kafka consumers once offset committing is enabled. Kafka Streams has a default commit interval of 30 s as in

Kafka Streams, the commit interval also controls fault tolerance (comparable to the checkpointing interval in other frameworks). For our experiments with Apache Beam and the Flink runner, we enable the *FasterCopy* option as we further discuss and evaluate in Section 6.2. Each microservice instance runs as a container in its own Kubernetes Pod with an additional sidecar container, exposing monitoring data specific to the stream processing framework. Per default, we configure 1 CPU and 4 GB of memory for each Pod, which is a common ratio of CPU and memory of cloud VMs. For Flink, we always have one additional Pod running the JobManager (see Fig. 2).

### 5.2. Configuration of task samples

As benchmark task samples, we use dataflow architectures representing typical use cases for analyzing IIoT sensor data streams. The task samples consume messages representing simulated power consumption measurements and produce messages with aggregation results. Our previous publication (Henning and Hasselbring, 2021b) describes our four benchmark task samples, named UC1–UC4, in detail. Unless otherwise stated, we use the following configuration of our benchmark dataflow architectures.

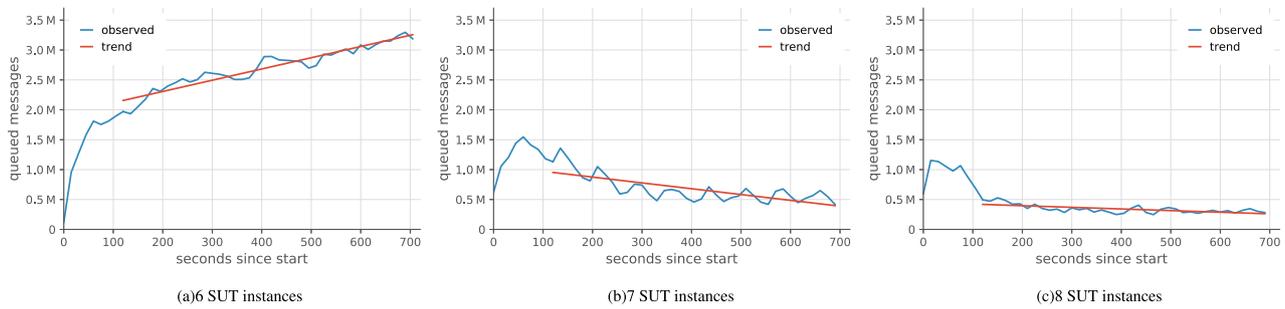
- Benchmark UC1 is configured to write each incoming message as a log statement to the standard output stream to simulate a database write operation (i.e., simulating a side effect in the dataflow architecture).
- Benchmark UC2 aggregates incoming messages over tumbling windows of one minute. Any out-of-order records arriving after the window has been closed are discarded.
- Benchmark UC3 aggregates records by their hour of day attribute over a time window of three days with a slide period of one day. That means each incoming record belongs to three time windows. Early results (i.e., before the end of the time window has passed) are emitted every 5 s. For Kafka Streams, such emission cannot explicitly be configured. However, Kafka Streams continuously forwards aggregation results based on the configured *commit interval* (which is 5 s as well).
- Benchmark UC4 aggregates incoming messages in nested groups of sensors. In contrast to previous work, we benchmark a simplified version of benchmark UC4, which omits the feedback loop. This allows for better predictability of the message volume and, hence, more comparable results.

### 5.3. Configuration of the benchmarking method

Our Theodolite benchmarking method assesses scalability regarding a configurable load type, a resource type, and SLOs.

**Load type.** If not stated differently, we evaluate scalability in regards to increasing the number of simulated sensors as load type. In benchmark UC4, this is indirectly controlled by increasing the number of nested groups, with each group containing 4 sub-groups or sensors. This means for  $n$  nested groups, we simulate  $4^n$  sensors. Each simulated sensor generates one measurement per second. When addressing RQ 3, we evaluate scalability in regards to increasing the size of time windows for which data is aggregated.

<sup>5</sup> <https://github.com/cau-se/theodolite>.



**Fig. 3.** Illustration of the consumer lag trend metric for an exemplary benchmark execution (Theodolite’s UC3 benchmark implemented with Kafka Streams and 50 000 messages/s) with different numbers of SUT instances. Independent of the number of instances, we can observe a variable lag. However, computing a trend line (without considering the measurements from an initial warmup period), reveals that for 6 instances, the number of queued messages will steadily increase over time. Providing 7 instances leads to a decrease in queued messages after the warmup period, while 8 instances yields an almost constant trend line.

**Resource type.** Unless otherwise stated, we use the number of instances as resource type. When addressing RQ 4, we additionally use the CPU and memory resources of a single instance as resource type.

**SLOs.** In all benchmark executions, we use an SLO based on our consumer lag trend metric (Henning and Hasselbring, 2021b). The consumer lag trend describes how many messages are queued in the messaging system, which have not been processed. Our consumer lag trend metric describes the average increase (or decrease) of the lag per second. It can be measured by monitoring the lag and computing a trend line using linear regression. The slope of this line is the lag trend. Fig. 3 illustrates the concept of the lag trend. For our experiments, we consider the SLO to be fulfilled if the lag trend does not increase by more than 1% of the generated message volume. For Apache Samza, we set this threshold to 5% of the generated message volume since we observe a permanent slightly increase independent of the provided resources.

In certain cases, we observed that under high load the consumer lag does not substantially increase, but records were discarded due to lateness. In most stream processing frameworks, operations on time windows still accept out-of-order records for a configurable amount of time. If this time has elapsed, records are discarded and not further processed. Thus, records are still consumed from the messaging system, not causing a consumer lag increase, but results become incorrect. To detect these cases, our benchmarks UC2 and UC4 (which aggregate data in short windows) contain a second SLO, which requires that no more than 1% of the generated messages are discarded. For the Apache Beam implementation with the Samza runner, no metrics concerning the number of dropped records are provided. With Beam’s Flink runner, these metrics are only unreliably available.<sup>6</sup> This means that for these two SUTs, we cannot definitely be sure whether the determined resource demand for UC2 and UC4 is sufficient to process all records successfully, but still indicates a lower bound.

**Additional configuration.** According to the previous experimental evaluations of our benchmarking method (Henning and Hasselbring, 2022a), we run our experiments with benchmark UC1–UC3 for a duration of 5 min while considering the first 2 min as warm-up period. As benchmark UC4 shows a higher variability in the results, we run its experiments for 10 min including a 4-min warm-up period. We repeat all our experiments 3 times, which we experimentally observed to be a good trade-off between overall execution time (or costs in the public cloud) and statistical grounding (Henning and Hasselbring, 2022a). We provide a replication package (Henning and Hasselbring, 2023) allowing to further repeat our experiments. We quantify scalability with the resource demand metric and use the *linear search* strategy in combination with the *lower bound restriction* (Henning and Hasselbring, 2022a).

<sup>6</sup> We asked a corresponding question regarding the metrics of both runners at Beam’s mailing list, but did not receive an answer.

## 6. Experimental results

In this section, we discuss and present the results of our benchmark executions. Section 6.1 addresses RQ 1 by running baseline experiments for each benchmark and framework. Section 6.2 addresses RQ 2 and investigates the impact of recently proposed performance optimization configurations for Apache Beam. Section 6.3 addresses RQ 3 and evaluates how different stream processing frameworks scale when increasing the duration of window aggregations. Section 6.4 addresses RQ 4 by evaluating how stream processing frameworks scale on a single node. Section 6.5 addresses RQ 5 and compares our baseline results in the private cloud with those of a public cloud. Finally, Section 6.6 addresses RQ 6 and repeats the same experiments for benchmark UC3 in Kubernetes clusters of different sizes.

### 6.1. Baseline comparison of frameworks

For our baseline experiments (RQ 1), we use the private cloud environment (see Table 1). We benchmark load intensities between 100 000 and 1 000 000 simulated sensors (and, thus, generated messages per second) for benchmark UC1 and UC2, 10 000 to 100 000 simulated sensors for UC3, and 5 to 9 nested groups (1024–262 144 generated messages per second) for benchmark UC4.

Fig. 4 shows the resource demand results for all evaluated frameworks and benchmarks. The results for benchmark UC4 (in the following figures as well) are visualized with an exponential scale with base 4 at the horizontal axis since the number of generated messages grows exponentially with a linear increase in the number of nested groups. We can observe that in almost all experiments, Flink, Hazelcast Jet, and Kafka Streams, have a considerably lower resource demand than the Beam deployments. Only Hazelcast Jet in UC4 and the Beam Flink runner for low loads in UC1 are exceptions to this. As in some cases, the generated load intensities were too high for the Beam deployments, we repeat the corresponding experiments with lower load intensities (see Figs. 5 and 8).

Despite some outliers (see Beam/Flink in UC1 and Beam/Samza in UC4), all frameworks show linear scalability, yet with different rates. Whereas both SUTs based on Beam show the steepest increase in required resources, the results of Flink, Kafka Streams, and Hazelcast Jet vary depending on the benchmark. In UC1 (see Fig. 4(a)), all frameworks behave similarly, with resource demands increasing slightly steeper for Hazelcast Jet compared to Kafka Streams and for Kafka Streams compared to Flink. For UC2 (see Fig. 4(b)), we see a clear ranking with Hazelcast Jet showing the best results, followed by Kafka Streams and Flink. For UC3 (see Fig. 4(c)), Hazelcast Jet appears to be even more superior. A single Jet instance is sufficient for all evaluated load intensities. On the other hand, Flink requires up to 10 TaskManagers and Kafka Streams up to 18 instances. Overall, Kafka Streams’ resource demand for UC3 increases at a steeper rate compared to Flink. To further inspect the scalability of Hazelcast Jet

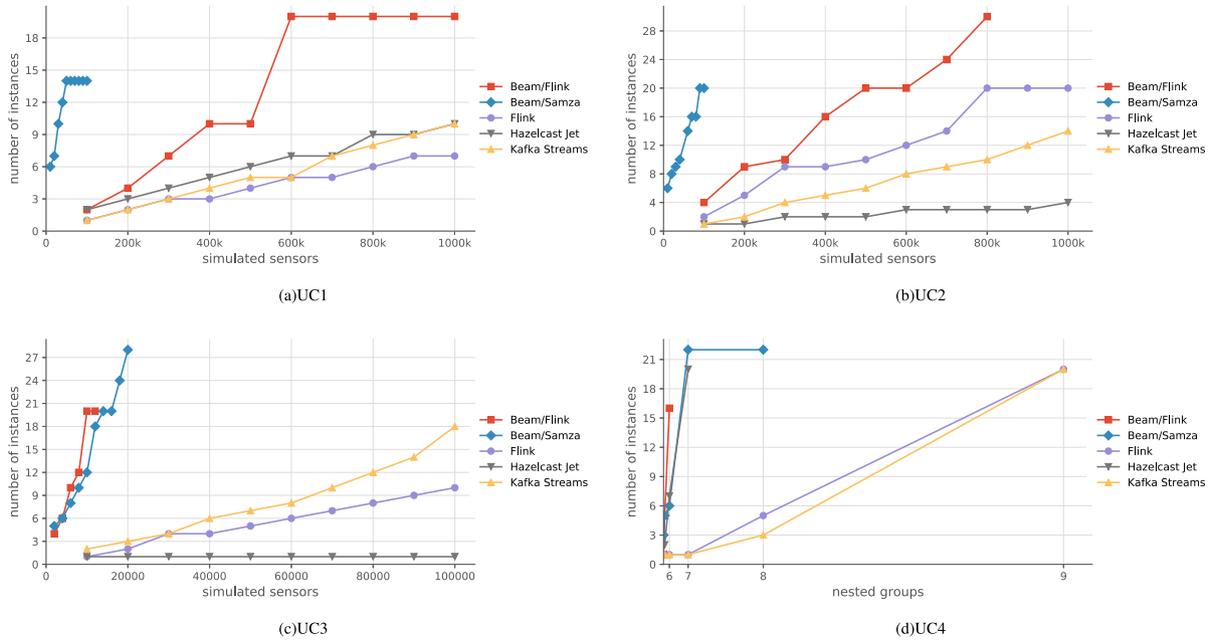


Fig. 4. Scalability benchmark results according to our resource demand metric for the stream processing frameworks Apache Beam (with the Flink and Samza runners), Apache Flink, Hazelcast Jet, and Apache Kafka Streams.

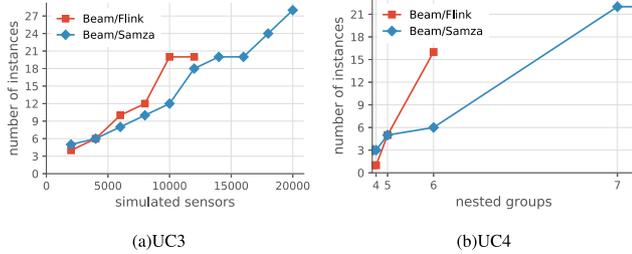


Fig. 5. Repetition of scalability experiments shown in Figs. 4(c)–4(d) with lower load intensities for Apache Beam with the Flink and the Samza runners.

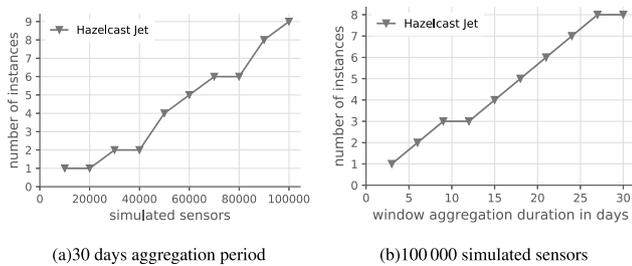


Fig. 6. Repetition of scalability experiments with Hazelcast Jet and benchmark UC3. (a) evaluates scalability regarding the number of simulated sensors with a 30 days aggregation period in contrast to the 3 days period in Fig. 4(c). (b) evaluates scalability regarding the aggregation period with a constant load of 100000 simulated sensors in contrast to 10000 simulated sensors in Fig. 9(a).

for UC3, we repeat these experiments with an aggregation duration of 30 days in contrast to 3 days as used in the other experiments. Fig. 6(a) shows that Hazelcast Jet also scales linearly in this case. With UC4 (see Fig. 4(d)), we observe a comparable increase in resource demand for Kafka Streams and Flink. In contrast to the other benchmarks, Hazelcast Jet shows a significantly higher resource demand. Up to 30 instances are not able to handle load from more than 7 nested sensor groups.

For the frameworks used with Apache Beam, we observe a significantly steeper increase in resource demand of Samza compared to Flink

in UC1 (cf. Fig. 4(a)) and UC2 (cf. Fig. 4(b)). For benchmark UC3 (see Fig. 5(a)), both frameworks scale at similar rates with Samza requiring slightly fewer instances. For benchmark UC4 (see Fig. 5(b)), it appears that the resource demand of Flink increases at a steeper rate as well. For lower load intensities (less than 5 nested groups), Flink requires fewer instances than Samza.

Worth mentioning is also the significant difference between native Apache Flink and the Flink runner of Apache Beam. In almost all experiments, the resource demand of Apache Beam with Flink is at least twice as high. For the more compute-intensive benchmarks UC3 and UC4, it is tremendously higher. The performance overhead of using Apache Beam as an abstraction layer has also been observed in related research (Hesse et al., 2019).

**RQ1:** All frameworks appear to be linearly scalable, however, with different resource usage. Depending on the task sample, Flink, Hazelcast Jet, or Kafka Streams have the smallest resource demand. In particular for the task samples of medium complexity, Hazelcast Jet’s performance is outstanding. Apache Beam implementations, executed by Samza or Flink, are inferior, independent of the use case.

## 6.2. Impact of Apache Beam configuration

As we have seen in Section 6.1, Apache Flink and Apache Samza in combination with Apache Beam have a significantly higher resource demand compared to the other evaluated frameworks. In this section, we address RQ2 and take a closer look at the scalability of the Apache Beam SUTs and evaluate how scalability is affected by different configuration options. Again, we use the private cloud environment (see Table 1). In the following, we first look at the Apache Flink runner and, afterward, at the Apache Samza runner.

### 6.2.1. Apache Flink

In their master’s thesis, Spæren (2021) investigates possible reasons for the performance overhead of the Flink runner found by Hesse et al. (2019). They discovered unnecessary serialization and deserialization

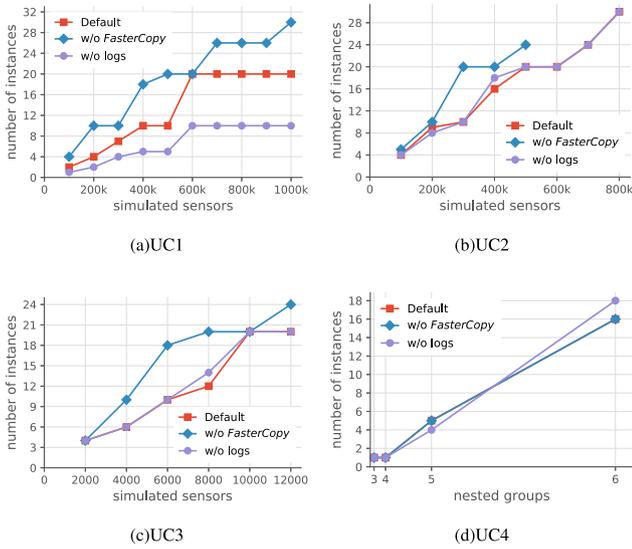


Fig. 7. Scalability benchmark results for different configurations of Apache Beam with the Apache Flink runner.

between operators and introduced the *FasterCopy* option, which disables these copy operations. This option is integrated in Beam since version 2.26. While the stream processing application must fulfill some requirements to run with the *FasterCopy* option, [Bensien \(2021\)](#) found that the Theodolite benchmarks fulfill these requirements. As additionally this option might become the default in future releases, we decided to turn on this option in our benchmark implementations by default. In this section, we evaluate how enabling and disabling *FasterCopy* affects scalability.

Additionally, we observed that Beam’s Kafka consumers generate a lot of log messages if not configured differently. This contrasts with the other frameworks. As extensive logging can actually have an impact on performance (see the following Section 6.4), we evaluate whether disabling all logging results in lower resource demand.

Fig. 7 shows our results of running the scalability benchmarks with the *FasterCopy* option disabled and logging disabled compared to the experiments from Section 6.1. We can see that enabling *FasterCopy* results in significantly lower resource demands for UC1–UC3 (see Figs. 7(a)–7(c)). This is in line with the performance improvements reported by [Spæren \(2021\)](#). For benchmark UC4 (see Fig. 7(d)), enabling *FasterCopy* seems not to have an effect on the resource demand. A possible explanation is that the dataflow architecture of UC4 involves more data transfer among instances and, hence, actually requires serialization and deserialization between operators.

We can observe that disabling all logging only has a small impact on the resource demand of benchmark UC2–UC4 (see Figs. 7(b)–7(d)), but significantly reduces the resource demand of benchmark UC1 (see Fig. 7(a)). The latter is expected since benchmark UC1 logs each incoming message to simulate side effects such as writing records to a database. We can conclude that the extensive logging of Beam’s Kafka consumer contributes very little to the overhead introduced by Apache Beam.

### 6.2.2. Apache Samza

In a blog post, software engineers at LinkedIn [Zhang et al. \(2020\)](#) report how they tremendously improved the performance of Beam’s Samza runner. Primarily, this was achieved by exporting Beam metrics more efficiently. Moreover, the authors observed that performance could further be improved when disabling the Beam metrics entirely. Although this might not be an option in production ([Zhang et al., 2020](#)), we are still interested in how much performance could be further improved when disabling all Beam metrics.

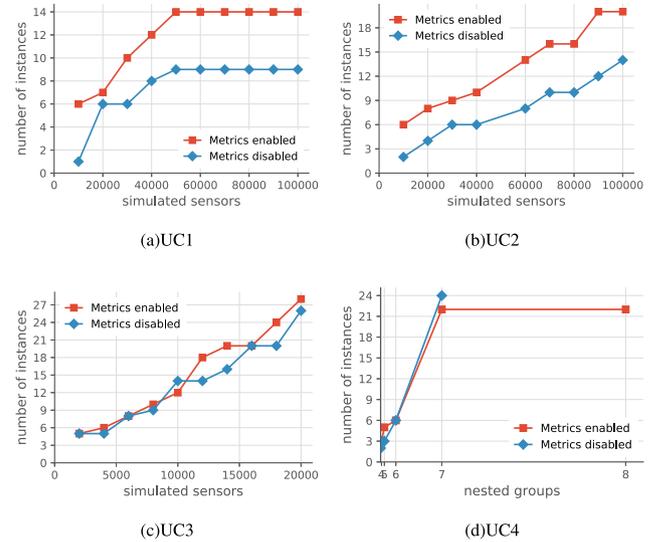


Fig. 8. Scalability benchmark results for different configurations of Apache Beam with the Apache Samza runner.

Fig. 8 shows our results for benchmarking scalability with Beam metrics enabled and disabled. We can observe that independent of the benchmark, disabling metrics results in a similar linear increase in resource demand, yet at a lower level. However, also with metrics disabled, the resource demand of Apache Beam with the Samza runner is considerable higher compared to most other frameworks (with metrics not disabled).

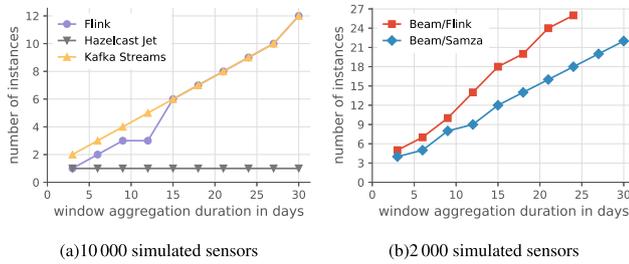
Worth mentioning is also the bachelor’s thesis of [Bensien \(2021\)](#), who benchmarked Beam with the Samza runner in version 2.22 using Theodolite’s UC1 benchmark. This Beam version did not include the performance improvements by [Zhang et al. \(2020\)](#) (released in Beam version 2.27). [Bensien](#) observed a resource demand more than twice as high with metrics enabled, compared to disabling them.

**RQ2:** The previously discovered negative impact on performance of using Apache Beam as abstraction layer persists. Possible configuration options for improving performance (Flink’s *FasterCopy*, disabling metrics, and disabling logging) have no significant impact on our results.

### 6.3. Scaling the window aggregation duration

In Section 6.1, we configure benchmark UC3 with a window duration of 3 days to compute an average daily course. This is a trade-off to still benchmark generated data volume of reasonable size. However, it is likely that in practice, larger time windows are required to obtain more reasonable results. Therefore, in this section, we evaluate, how different stream processing frameworks scale with increasing benchmark UC3’s window duration. We increase the window duration from 3 days to 30 days, while keeping the number of simulated sensors and, thus, the incoming message rate constant. For the benchmark deployment depicted in Fig. 2, this means the message volume and the state processed by the *aggregate* operation increases. Again, we use the private cloud environment (see Table 1). This evaluation is an example of benchmarking scalability with respect to the work performed for each incoming message in contrast to scaling the load at the framework and, thus, addresses RQ3.

Fig. 9 shows the results of these experiments. According to our previous results in Section 6.1, we simulate 10000 sensors for our



**Fig. 9.** Scalability benchmark results of all stream processing frameworks when increasing the duration of aggregation windows and, thus the number of windows maintained simultaneously.

experiments with Flink, Hazelcast Jet, and Kafka Streams (see Fig. 9(a)) and 2000 sensors for the Beam SUTs (see Fig. 9(b)). We can observe that again all frameworks scale approximately linearly. Remarkable is again the performance of Hazelcast Jet, which only requires a single instance, independently of the window size. We repeat these experiments with a higher load of 100 000 sensors. As shown in Fig. 6(b), Hazelcast Jet also scales approximately linearly in this case. In contrast to scaling with the number of sensors (see Fig. 4(c)), Kafka Streams and Flink scale now with about the same rate of resource demand increase. Similar to the results shown in Fig. 5(a), Samza's resource demand increases less steeply compared to Beam's Flink runner.

**RQ3:** Similar to scaling the load intensity the microservice is subject to, all frameworks show linear scalability when scaling the computational work performed inside the microservice. Again, however, the increase in resource demand differs noticeably between frameworks with Hazelcast Jet being superior for the studied case, followed by Flink and Kafka Streams, and significantly higher resource demand of the Beam-based deployments.

#### 6.4. Scaling on a single node

In this section, we address RQ4 and evaluate whether vertical scaling can be a complementary measure to achieve scalability of stream processing frameworks. We, therefore, scale our SUT vertically by varying both the number of instances, which are all deployed to the same cluster node, as well as the CPU and memory provided for a single SUT pod.

For the experiments of this section, we slightly modify our experimental setup (see Table 1). We only deploy 3 Kafka brokers, which run on three different Kubernetes nodes. Of the other two nodes, we dedicate one to run the load generators and one to run the SUT instances. Although we only run 3 Kafka brokers and all load generator instances on the same node, we can confirm that the configured load is still successfully generated.

To not fully utilize the SUT node, which also runs some infrastructure and monitoring components, we deploy up to 20 instances with one CPU core each or up to 20 CPU cores for a single instance. For scaling with the number of instances, we keep the same configuration as in the previous experiments. For scaling with the resources per pod, we only refer to the number of CPU cores, but scale memory proportionally. (However, in all our experiments we never observed fully utilized pod memory.) In order to utilize multiple CPU cores, most stream processing frameworks have to be configured accordingly. For Flink, we scale the number of task slots of the TaskManager equally to the number of CPU cores. In Kafka Streams, we equally scale the

number of threads to the number of CPU cores. For Samza, the documentation is inconsistent regarding scaling a standalone application on a single node. We decided to equally scale the *container thread pool* size to the number of CPU cores. Hazelcast Jet does not require additional configuration as an instance configures its cooperative thread pool automatically according to the number of CPU cores provided.

For most experiments, we generate the same load intensities as in the first experiment. However, we use the smaller load intensities from Section 6.2 for the Beam experiments and the 30 days window for Hazelcast Jet with benchmark UC3 as introduced in Section 6.3.

Fig. 10 shows the results of our experiments with a single node. Almost all frameworks show approximately linear scalability when scaling the number of instances. We observe that Beam with the Samza runner does not scale with increasing the CPU resources per pod. Hence, we assume that scaling the container thread pool is not the right option to increase capacity on a single node. Whether other configuration options exist remains unclear. Further, we can observe that no framework is able to process load intensities higher than 200 000 messages per second with a single instance for benchmark UC1. The reason for this is that we simulate database writes by printing all incoming records to the standard output stream. Running a single instance of the frameworks causes all threads to write to the same stream, which is synchronized and becomes the bottleneck of our evaluation.

Kafka Streams seems to be more efficient when scaling with the number of instances, compared to scaling with the number of cores. For Flink and Hazelcast Jet, scaling with the number of cores is more efficient for the more complex dataflows in UC3 and UC4, while with benchmark UC2 both types of scaling yield similar results. Remarkable are the results for scaling with the number of cores with Beam and the Flink runner. In contrast to native Flink, Beam with the Flink runner seems not to be scalable with respect to the number of cores. Moreover, the Kafka Streams implementation of benchmark UC3 scales with neither increasing the number of instances nor with increasing the number of cores. As it scales linearly when running on multiple nodes (see Fig. 5(a)), our results suggest that underlying hardware resources become exhausted. However, from manually observing system-level metrics, we cannot observe anything conspicuous.

**RQ4:** Despite some exceptions for specific task samples, stream processing frameworks allow scaling them on a single node, making vertical scaling a complementary measure to achieve scalability of stream processing frameworks. For Flink and Hazelcast Jet, it might be more resource efficient to scale with the CPU cores per microservice instance instead of scaling the number of instances. However, caution should be exercised when accessing shared resources.

#### 6.5. Comparing scalability in public and private clouds

Section 6.1 presented our baseline results of benchmarking scalability in the private cloud Kubernetes cluster. We repeat these experiments in an identically configured Kubernetes cluster in the Google cloud to address RQ5 (see Table 1). As in the previous section, we generate the same load intensities as in the first experiment, yet using the smaller load intensities from Section 6.2 for the Beam experiments and the 30 days window for Hazelcast Jet with benchmark UC3.

Fig. 11 compares the scalability results in the private and in the public cloud for each benchmark and stream processing framework. In general, we can observe that the results for both clouds are similar and all frameworks show linear scalability, independent of the cloud platform. We can see a tendency that for higher loads (Flink, Hazelcast Jet, and Kafka Streams), the resource demand in the Google cloud

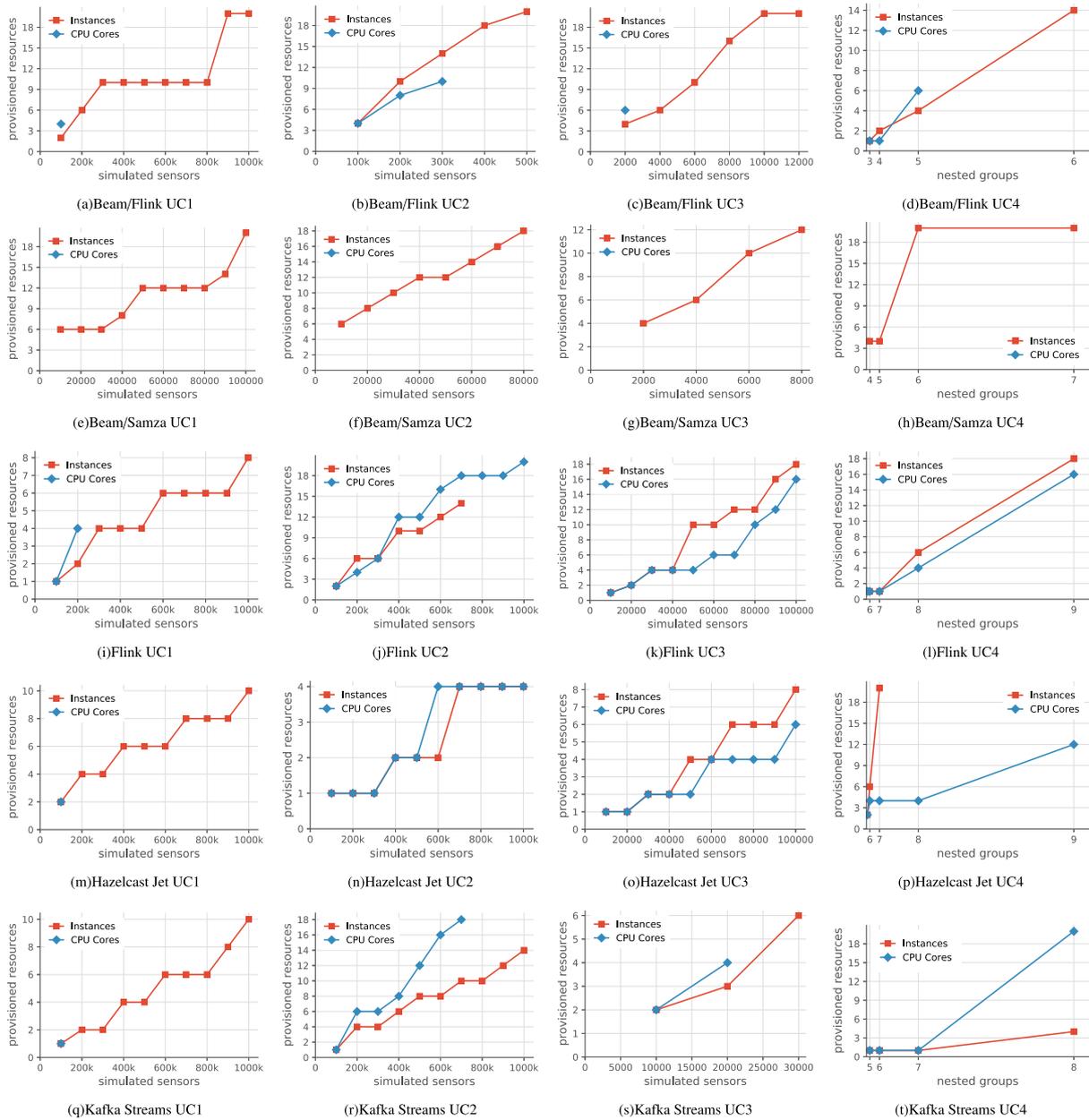


Fig. 10. Scalability of stream processing frameworks on a single node.

increases at a slightly steeper rate. We expect that this is due to the fact that we used Google’s general-purpose *E2* virtual machines with potentially less powerful resources. Our benchmark UC3 execution of Hazelcast stands out as its resource demand increases at a lower rate for the Google cloud. Further experiments would be required to investigate whether this combination of framework and task sample is special in some way.

**RQ5:** Our previous results apply independently of whether a public or private cloud environment is used. In our experiments, resource demand increases steeper in the public cloud, yet we expect this to be due to the specific machine types selected.

### 6.6. Scaling the cluster size

In the previous experiments, we deployed up to 30 SUT instances since for larger numbers we observed interference of the SUT, load generation, messaging system, and infrastructure components causing unstable results. In this section, we address RQ6 and evaluate whether stream processing frameworks can be scaled further when also increasing the underlying computing resources. To obtain a more stable infrastructure, we modify our experimental setup in this section (see Table 1). We compose our Kubernetes cluster of two node groups in Google Cloud: The first consists of four 16-core machines, which run the load generators, four Kafka brokers, and additional benchmarking infrastructure. The second node group only runs the SUTs. We evaluate two sizes of this node group, namely four and eight 16-core machines. As we observed linear scalability for all frameworks and independent of the benchmark, we focus on benchmark UC3 in these experiments.

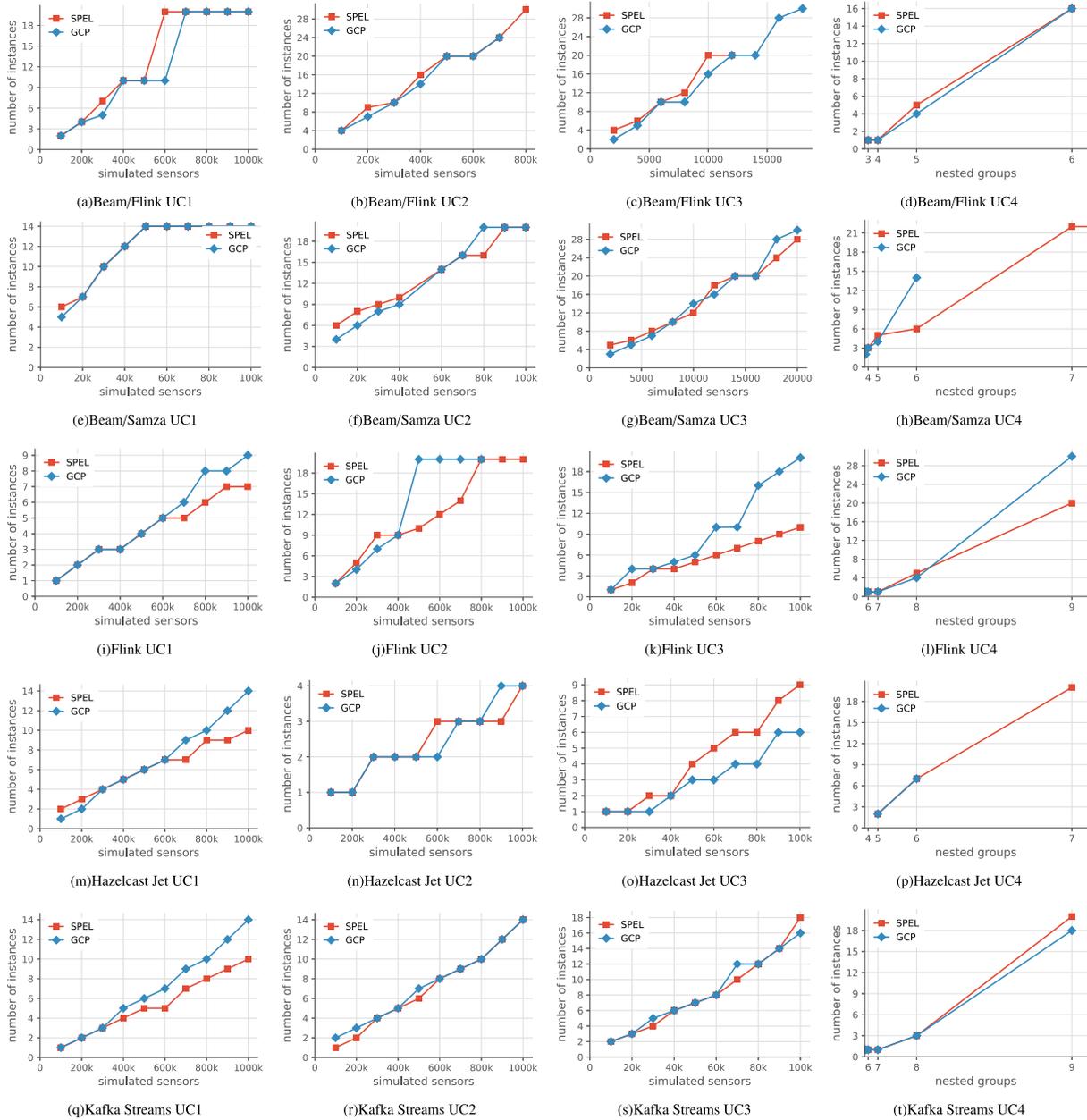


Fig. 11. Scalability of stream processing frameworks in our private cloud (SPEL) compared to the Google cloud (GCP).

We simulate up to 1 000 000 sensors for Flink, Hazelcast Jet, and Kafka Streams and up to 100 000 sensors for the Beam SUTs. For the cluster with four SUT nodes, we deploy up to 55 instances, while for the cluster with eight SUT nodes, we deploy up to 110 instances. For all frameworks, we increase the number of instances in steps of five.

Fig. 12 shows our scalability benchmark results for larger clusters. For Beam/Flink, Flink, and Kafka Streams, we can observe that the maximum processable load increases when using a cluster of eight nodes instead of four nodes. For Hazelcast Jet, already the four-node cluster is able to process the highest generated load. For Beam with the Samza runner, only 20 000 messages per second can be processed, independent of the cluster size. We observe frequent crashes with large Beam/Samza deployments, but due to the many issues observed in our Samza experiments, do not further investigate this. An important observation is that for all frameworks, fewer SUT instances are required when using the larger Kubernetes cluster.

**RQ 6:** Observed scalability limits seem to be caused by utilized hardware and not by the streaming frameworks themselves. Hence, limits can be raised by using larger clusters.

### 7. Threats to validity

Despite careful research design, there exist threats and limitations to the validity of our evaluation, which we report below.

*Threats to internal validity.* Cloud platforms in general allow only making little assumptions regarding the underlying hardware or software infrastructure (Bermbach et al., 2017). Cloud-native containerized deployments as in this study further abstract this. We consciously chose a representative, Kubernetes-based execution environment, which, however, limits control of possible influences on our result. To obtain

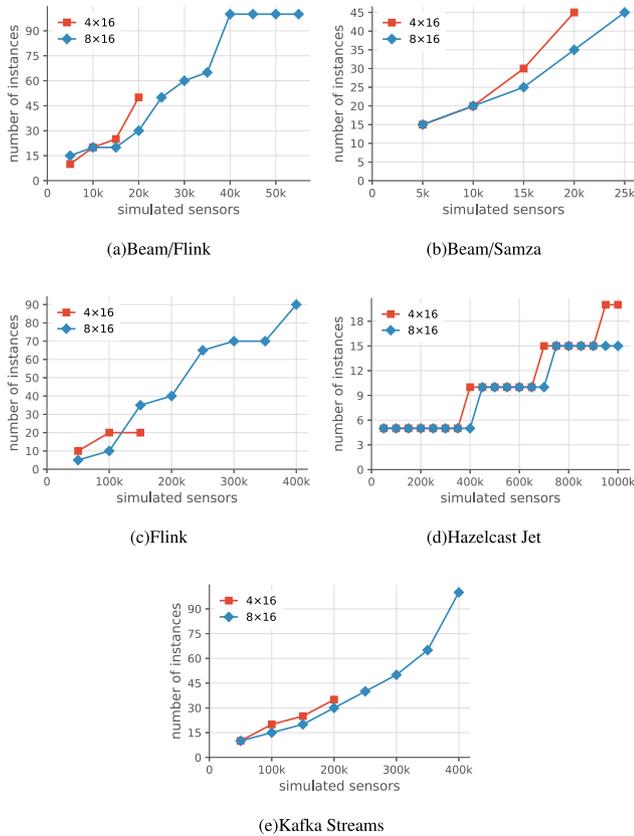


Fig. 12. Scalability benchmark results for benchmark UC3 in Google Cloud for different cluster sizes (machines  $\times$  cores).

statistically grounded benchmarking results, we build these evaluations upon the results of our benchmarking method's evaluation (Henning and Hasselbring, 2022a). Nevertheless, we only found that the selected configuration options provide good estimates. Hence, resource demands obtained in these evaluations should also only be considered estimates. Repeating individual experiments more often and for longer time periods as well as using our *full search* strategy is likely to produce very similar, but not necessarily identical results. Moreover, we evaluate a larger set of SUTs, load types, resource types, SLOs, and benchmarks in this study compared to our method's evaluation. We address this limitation by carefully observing the benchmark execution but do not conduct as extensive experiments as in our method's evaluation. We refer to our previous publication for further discussion on balancing statistical grounding and time-efficient execution for scalability benchmarking (Henning and Hasselbring, 2022a).

**Threats to external validity.** We conduct all experiments in this evaluation with our Theodolite benchmark task samples. As discussed in previous work (Henning and Hasselbring, 2021b), these benchmarks represent relevant use cases. However, we cannot directly generalize our findings to arbitrary other applications, where other frameworks, configuration options, or deployment options might perform differently than our experiments. We conduct all experiments in this evaluation in our private cloud environment and in Google Cloud. For the private cloud, we use comparatively large bare metal nodes, while for Google Cloud we use large virtual machines. Although our evaluations in Section 6.5 show only small deviations between both environments, we cannot necessarily conclude that we would obtain the same results in other cloud environments. We deliberately assess all frameworks primarily using their default configurations. This should mitigate any bias stemming from different levels of our experience with the different

frameworks. However, it does not allow us to make any conclusions about the potential performance and scalability improvements when fine-tuned for specific scenarios. When choosing between different stream processing frameworks, practitioners might also consider other attributes beyond scalability, such as end-to-end latency, the ability to recover from failures, or ease of use and development.

## 8. Related work

Over the last few years, several benchmarks for stream processing frameworks have been proposed and stream processing benchmarking studies have been conducted. The differentiation between benchmarks and experimental studies applying them is sometimes blurry. Many publications that present benchmarks perform also an experimental study with them. On the other hand, many experimental studies utilize existing benchmarks, but modify them. Nevertheless, we structure this section into two parts: First, we give an overview of stream processing benchmarks to justify our benchmark selection for this study. Second, we discuss related stream processing benchmarking studies. For a systematic mapping of the literature on performance in stream processing, we refer to the recent study by Vogel et al. (2023).

### 8.1. Related work on stream processing benchmarks

Besides the Theodolite stream processing benchmarks used in this study, several other benchmarks for stream processing frameworks have been proposed. Table 2 summarizes characteristics of the discussed benchmarks.

StreamBench (Lu et al., 2014) is one of the earliest benchmarks for modern stream processing frameworks. While originally only implemented for Spark and Storm, it has later been used to benchmark Apache Apex, Beam, Flink, and Samza as well (Hesse et al., 2019; Qian et al., 2016). As its name suggests, SparkBench (Li et al., 2015) is a benchmark tailored to Apache Spark. The Yahoo Streaming Benchmark (YSB) (Chintapalli et al., 2016) is frequently used and adapted in research (Lopez et al., 2016; Yang et al., 2017; Karakaya et al., 2017; Nasiri et al., 2019; Zeuch et al., 2019; Chu et al., 2020; van Dongen and van den Poel, 2020). Worth highlighting is the work of Shahverdi et al. (2019), who extend YSB with implementations for the frameworks Kafka Streams and Hazelcast Jet. As discussed in Section 3, these frameworks are particularly suited to be deployed as microservices. RIoT Bench (Shukla et al., 2017) provides four application benchmarks for Storm composed of 27 small task samples. Nasiri et al. (2019) adopt RIoT Bench for Flink and Spark. Karimov et al. (2018) present a benchmark with two task samples, derived from a real industrial context, yet without providing open-source implementations.

More recently, DSPBench (Bordin et al., 2020), OSPBench (van Dongen and van den Poel, 2020), and ESPBench (Hesse et al., 2021) have been proposed. DSPBench contains 15 benchmarks, which re-sample typical stream processing applications, derived from reviewing the literature. OSPBench provides benchmarks for analyzing traffic sensor data. Besides evaluations of latency, throughput, and resource usage, van Dongen and van den Poel used OSPBench to also evaluate scalability (van Dongen and van den Poel, 2021a) and fault recovery (van Dongen and van den Poel, 2021b). In contrast to most other benchmarks, OSPBench provides implementations for the rather new framework Kafka Streams, which is also evaluated in this study. The Enterprise Stream Processing Benchmark (ESPBench) builds upon the Senska benchmark (Hesse et al., 2018). It is special in the sense that it integrates a relational database management system. In contrast to most other benchmarks, ESPBench's task samples are implemented with Apache Beam. While Hesse et al. (2021) only perform evaluations with Spark, Flink, and Hazelcast Jet, we expect that also other Beam runners can be used to run the benchmark.

The Nexmark benchmark (Tucker et al., 2010) has originally been proposed as the *Niagara Extension to the XMark benchmark* addressed to

**Table 2**  
Overview of the characteristics and implementations of stream processing benchmarks (Henning, 2023).

Benchmark	Task samples	Open source	Messaging			Stream processing framework						Cloud-native			
			Kafka	Others	None	Flink	Spark	Storm	Samza	Kafka Streams	Hazelcast Jet	Others	Database	Containers	Kubernetes
Theodolite (Henning and Hasselbring, 2021b)	4	✓	✓			✓			✓ <sup>a</sup>	✓	✓ <sup>a</sup>	<sup>b</sup>	✓	✓	
Beam Nexmark (Apache Software Foundation, 2022) <sup>c</sup>	13	✓	✓	✓		✓ <sup>a</sup>	✓ <sup>a</sup>		?	?	?				
ESPBench (Hesse et al., 2021)	5	✓	✓			✓ <sup>a</sup>	✓ <sup>a</sup>		?	✓ <sup>a</sup>	?	✓			
OSPBench (van Dongen and van den Poel, 2020)	5	✓	✓			✓	✓			✓			✓	✓	
DSPBench (Bordin et al., 2020)	5	✓	✓			✓	✓	✓				✓			
Shahverdi et al. (2019)	1	✓	✓			✓	✓	✓		✓		✓			
Karimov et al. (2018)	2			✓		✓	✓	✓							
RIoT Bench (Shukla et al., 2017)	4 <sup>d</sup>	✓		✓			✓	✓				✓			
YSB (Chintapalli et al., 2016)	1	✓	✓			✓	✓	✓				✓			
SparkBench (Li et al., 2015)	10	✓		✓			✓	✓							
StreamBench (Lu et al., 2014)	7		✓				✓	✓							
Linear Road (Arasu et al., 2004)	5			✓											✓

<sup>a</sup> Using Apache Beam.

<sup>b</sup> Optional.

<sup>c</sup> The Beam Nexmark benchmarks are based on the Nexmark paper (Tucker et al., 2010).

<sup>d</sup> RIoT Bench's 4 application benchmarks are composed of 27 microbenchmarks.

first-generation stream processing systems (see the survey of Fragkoulis et al. (2023) for a discussion of first and second-generation stream processing systems). The Apache Beam community adapted and extended Nexmark with implementations for Beam to benchmark the performance of different runners (Apache Software Foundation, 2022). Documentation and benchmark results are provided for the direct runner as well as for the Flink, the Spark, and the Google Cloud Dataflow runners. However, running the benchmark with other runners should be possible as well. Recently, there seems to be an effort to implement the Nexmark task samples with other frameworks in an open-source project.<sup>7</sup> However, currently this project only provides implementations for Apache Flink. Moreover, Gencer et al. (2021) implemented the Nexmark benchmark for their performance evaluation of Hazelcast Jet.

Worth mentioning is also the Linear Road benchmark presented by Arasu et al. (2004). Although published years before all modern stream processing frameworks considered in this work have been released, it is still used in research (Zhang et al., 2017; Zeuch et al., 2019; Sax, 2020) and compared to newer benchmarks (Bordin et al., 2020; Hesse et al., 2021). Pagliari et al. (2020) and Garcia et al. (2022b,a) present approaches to generate benchmarks.

From Table 2, we can see that a lot of open-source benchmarks have been proposed. Apart from the Theodolite benchmarks, none of these benchmarks is particularly addressed to scalability. Often originating in data management research, many benchmarks are defined as “queries” over data streams (Tucker et al., 2010; Karimov et al., 2018; Hesse et al., 2021). Most benchmarks include a messaging system as a middleware component between workload generation and stream processing framework. In the vast majority of cases, this is Apache Kafka. Karimov et al. (2018) exclude such a system to not let it become the benchmark's bottleneck. Our Theodolite benchmarks purposely include Kafka to represent more realistic microservice deployments (Henning and Hasselbring, 2021b). Flink, Spark, and Storm are by far the most supported frameworks. Only a few benchmarks exist for Samza, Kafka Streams, and Hazelcast Jet, which are frameworks particularly suited to be deployed as microservices. Our Theodolite benchmarks are the only ones providing implementations for all of them. While some benchmarks include an interaction with a database in their setup, others do not. With the Theodolite benchmarks, a database can optionally be used as we did in a previous study (Pfandzelter et al., 2022). Besides our

Theodolite benchmarks, there is only one other benchmark (OSPBench) that is provided as container images to be used in a cloud-native setting. No other benchmark provides Kubernetes manifests.

## 8.2. Related work on stream processing benchmarking

Table 3 provides an overview of experimental performance evaluation and benchmarking studies. It indicates the applied benchmark, the evaluated stream processing, and information regarding the experiment setup and method. The latter includes whether the respective study was performed in a cloud environment, in a distributed fashion with multiple instances of the framework deployed. Moreover, it shows whether the benchmarks have been executed with different resource amounts and different load intensities and whether different resource amounts and load intensities are evaluated in isolated experiments. In previous work, we argued that scalability should be evaluated with isolated experiments for different combinations of load and resources (Henning and Hasselbring, 2021a, 2022a).

We can observe that there is no established stream processing benchmark. Only YSB is used in several studies. However, YSB can be considered a micro-benchmark (Bermbach et al., 2017) and, hence, is less suited to benchmark entire microservices. Except for the preliminary evaluation of our Theodolite benchmarks (Henning and Hasselbring, 2021b), there is no benchmarking study addressed to stream processing frameworks employed within microservice architectures.

Flink, Spark, and Storm are by far the most frequently benchmarked frameworks. Kafka Stream, Hazelcast Jet, and Samza, which are particularly suited to be deployed as microservices, are only benchmarked in a few studies and there is no study benchmarking all of them.

9 out of 20 studies report on experiments in public or private clouds. Except for this and our previous study (Pfandzelter et al., 2022), there are no evaluations in Kubernetes. Likewise, there are no further studies evaluating scalability with a systematic approach as we do in this study. Vikash et al. (2020), Nasiri et al. (2019), Karakaya et al. (2017), and van Dongen and van den Poel (2021a) explicitly evaluate scalability, however, without testing different load intensities against different resource amounts in isolated experiments. Nasiri et al. (2019) conduct independent evaluations of scaling load and computing resources and, thus, address another aspect than our study. Our previous study (Pfandzelter et al., 2022) applies our Theodolite method as well, but benchmarks scalability with respect to costs and is addressed to comparing stream processing deployments against Function-as-a-Service offerings. While this work particularly addresses scalability,

<sup>7</sup> <https://github.com/nexmark/nexmark>.

**Table 3**  
Overview of employed benchmarks, frameworks, and experimental setup of stream processing benchmarking studies (Henning, 2023).

Publication	Benchmark	Framework	Execution
	Theodolite (Henning and Hasselbring, 2021b)		
	Beam Nexmark (Apache Software Foundation, 2022)		
	ESPBench (Hesse et al., 2021)		
	OSPBench (van Dongen and van den Poel, 2020)		
	DSPBench (Bordin et al., 2020)		
	Shahverdi et al. (2019)		
	Karimov et al. (2018)		
	RIoTbench (Shukla et al., 2017)		
	YSB (Chintapalli et al., 2016)		
	SparkBench (Li et al., 2015)		
	StreamBench (Lu et al., 2014)		
	Linear Road (Arasu et al., 2004)		
	Others		
	Flink		
	Spark		
	Storm		
	Samza		
	Kafka Streams		
	Hazelcast Jet		
	Others		
	Cloud environment		
	Distributed		
	Different resource amounts		
	... in isolated experiments		
	Different load intensities		
	... in isolated experiments		
This work	✓	✓	✓
Pfandzelter et al. (2022)	✓	✓ <sup>a</sup>	✓
Hesse et al. (2021)		✓ <sup>a</sup>	✓
van Dongen and van den Poel (2021a)	✓	✓ <sup>a</sup>	✓
van Dongen and van den Poel (2021b)	✓	✓	✓
Bordin et al. (2020)		✓	✓
Chu et al. (2020)		✓	✓
Vikash et al. (2020)		✓	✓
van Dongen and van den Poel (2020)	✓	✓	✓
Nasiri et al. (2019)		✓	✓
Shahverdi et al. (2019)	✓	✓	✓
Zeuch et al. (2019)		✓	✓
Karimov et al. (2018)		✓	✓
Truong et al. (2018)		✓	✓
Karakaya et al. (2017)		✓	✓
Shukla et al. (2017)		✓	✓
Yang et al. (2017)		✓	✓
Chintapalli et al. (2016)		✓	✓
Lopez et al. (2016)		✓	✓
Qian et al. (2016)		✓	✓
Lu et al. (2014)		✓	✓

<sup>a</sup> Using Apache Beam.

there are other recent studies that benchmark qualities such as throughput (Bordin et al., 2020; Chu et al., 2020; van Dongen and van den Poel, 2020), latency (Bordin et al., 2020; van Dongen and van den Poel, 2020; Hesse et al., 2021), and fault-tolerance (van Dongen and van den Poel, 2021b).

### 9. Conclusions

In this study, we benchmark the scalability of distributed stream processing frameworks particularly suited to be used within microservices. We experimentally evaluate the frameworks Apache Flink, Apache Kafka Streams, Hazelcast Jet, and Apache Beam with the Apache Flink runner and the Apache Samza runner in a private cloud environment and in the Google cloud. We find that all frameworks show linear scalability for most use cases, however with partially significantly different resource demands. There is no clear superior framework. Instead, depending on the use case Flink, Hazelcast Jet, or Kafka Streams show the lowest increase in resource demand. Using Apache Beam as abstraction layer still comes with a significant negative impact on performance, leading to a significantly steeper increase in resource demand, independent of the use case. We observe our results regardless of scaling load on a microservice, scaling the computational work performed inside the microservice, the selected cloud environment, and whether the microservice is scaled over multiple nodes or on a single node. The latter means that vertical scaling distributed stream processing frameworks can – to some extent – also complement horizontal scaling. In summary we found that while scalable microservices can be designed with all evaluated frameworks, the choice of

a framework and its deployment has a considerable impact on the cost of operating it. This emphasizes also a key benefit of designing stream processing applications using microservice-based architectures: the ability for each service to select the most suitable stream framework based on its specific use case, requirements for scalability, and other relevant considerations. Our findings leave room for future research on exploring the multi-dimensional space of cloud deployment options and finding an optimal one. Moreover, our work paves the way for deeper analysis of the underlying reasons for our observed results by benchmarking further task samples, deployments, and configuration options.

### CRedit authorship contribution statement

**Sören Henning:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Wilhelm Hasselbring:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

We provide a replication package and the collected data of our experiments as supplemental material, such that other researchers may repeat and extend our work.

## Acknowledgments

This work is supported by the Google Cloud Research Credits program under award no. 211920173. We would like to thank JKU and Dynatrace for co-funding one of the authors (S. Henning) during the revision of this article.

## References

- Akidau, T., Begoli, E., Chernyak, S., Hueske, F., Knight, K., Knowles, K., Mills, D., Sotolongo, D., 2021. Watermarks in stream processing systems: Semantics and comparative analysis of Apache Flink and Google Cloud Dataflow. *Proc. VLDB Endow.* 14 (12), 3135–3147. <http://dx.doi.org/10.14778/3476311.3476389>.
- Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Mohtezuma, R.J., Lax, R., McVeety, S., Mills, D., Perry, F., Schmidt, E., Whittle, S., 2015. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proc. VLDB Endow.* 8 (12), 1792–1803. <http://dx.doi.org/10.14778/2824032.2824076>.
- Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J.-C., Hueske, F., Heise, A., Kao, O., Leich, M., Leser, U., Markl, V., Naumann, F., Peters, M., Rheinländer, A., Sax, M.J., Schelter, S., Höger, M., Tzoumas, K., Warneke, D., 2014. The stratosphere platform for big data analytics. *VLDB J.* 23 (6), 939–964. <http://dx.doi.org/10.1007/s00778-014-0357-y>.
- Apache Software Foundation, 2022. Nexmark benchmark suite. URL <https://beam.apache.org/documentation/sdks/java/testing/nexmark/>.
- Arasu, A., Cherniack, M., Galvez, E., Maier, D., Maskey, A.S., Ryvkina, E., Stonebraker, M., Tibbetts, R., 2004. Linear road: A stream data management benchmark. In: *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30. VLDB '04, VLDB Endowment*, pp. 480–491.
- Balalaie, A., Heydarnoori, A., Jamshidi, P., 2016a. Microservices architecture enables DevOps: Migration to a cloud-native architecture. *IEEE Softw.* 33 (3), 42–52. <http://dx.doi.org/10.1109/ms.2016.64>.
- Balalaie, A., Heydarnoori, A., Jamshidi, P., 2016b. Migrating to cloud-native architectures using microservices: An experience report. In: Celesti, A., Leitner, P. (Eds.), *Advances in Service-Oriented and Cloud Computing*. Springer International Publishing, pp. 201–215.
- Bellemare, A., 2020. *Building Event-Driven Microservices*, first ed. O'Reilly Media, Inc.
- Bensien, J.R., 2021. *Scalability Benchmarking of Stream Processing Engines with Apache Beam Bachelor's thesis*. Kiel University.
- Bermbach, D., Wittern, E., Tai, S., 2017. *Cloud Service Benchmarking: Measuring Quality of Cloud Services from a Client Perspective*, first ed. Springer Publishing Company, Incorporated, <http://dx.doi.org/10.1007/978-3-319-55483-9>.
- Bordin, M.V., Griebler, D., Mencagli, G., Geyer, C.F.R., Fernandes, L.G.L., 2020. DSPBench: A suite of benchmark applications for distributed data stream processing systems. *IEEE Access* 8, 222900–222917. <http://dx.doi.org/10.1109/ACCESS.2020.3043948>.
- Bucchiarone, A., Dragoni, N., Dustdar, S., Larsen, S.T., Mazzara, M., 2018. From monolithic to microservices: An experience report from the banking domain. *IEEE Softw.* 35 (3), 50–55. <http://dx.doi.org/10.1109/MS.2018.2141026>.
- Burns, B., Grant, B., Oppenheimer, D., Brewer, E., Wilkes, J., 2016. Borg, Omega, and Kubernetes. *Commun. ACM* 59 (5), 50–57. <http://dx.doi.org/10.1145/2890784>.
- Carbone, P., Ewen, S., Fóra, G., Haridi, S., Richter, S., Tzoumas, K., 2017. State management in Apache Flink®: Consistent stateful distributed stream processing. *Proc. VLDB Endow.* 10 (12), 1718–1729. <http://dx.doi.org/10.14778/3137765.3137777>.
- Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., Tzoumas, K., 2015. Apache Flink: Stream and batch processing in a single engine. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* 36 (4).
- Chintapalli, S., Dagit, D., Evans, B., Farivar, R., Graves, T., Holderbaugh, M., Liu, Z., Nusbaum, K., Patil, K., Peng, B.J., Poulosky, P., 2016. Benchmarking streaming computation engines: Storm, Flink and Spark Streaming. In: 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPSW). pp. 1789–1792. <http://dx.doi.org/10.1109/IPDPSW.2016.138>.
- Chu, Z., Yu, J., Hamdull, A., 2020. Maximum sustainable throughput evaluation using an adaptive method for stream processing platforms. *IEEE Access* 8, 40977–40988. <http://dx.doi.org/10.1109/ACCESS.2020.2976738>.
- Cloud Native Computing Foundation, 2022. CNCF annual survey 2021. URL <https://www.cncf.io/reports/cncf-annual-survey-2021>.
- Davoudian, A., Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Comput. Surv.* 53 (5), <http://dx.doi.org/10.1145/3408314>.
- Fragkoulis, M., Carbone, P., Kalavri, V., Katsifodimos, A., 2023. A survey on the evolution of stream processing systems. <http://dx.doi.org/10.48550/arxiv.2008.00842>.
- Fritzsche, J., Bogner, J., Wagner, S., Zimmermann, A., 2019. Microservices migration in industry: Intentions, strategies, and challenges. In: 2019 IEEE International Conference on Software Maintenance and Evolution (ICSME). pp. 481–490. <http://dx.doi.org/10.1109/ICSME.2019.00081>.
- Gannon, D., Barga, R., Sundaresan, N., 2017. Cloud-native applications. *IEEE Cloud Comput.* 4 (5), 16–21. <http://dx.doi.org/10.1109/MCC.2017.4250939>.
- Garcia, A.M., Griebler, D., Schepke, C., Fernandes, L.G.L., 2022a. Evaluating microbatch and data frequency for stream processing applications on multi-cores. In: 2022 30th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP). pp. 10–17. <http://dx.doi.org/10.1109/PDP55904.2022.00011>.
- Garcia, A.M., Griebler, D., Schepke, C., Fernandes, L.G., 2022b. SPBench: a framework for creating benchmarks of stream processing applications. *Computing* <http://dx.doi.org/10.1007/s00607-021-01025-6>.
- Gencer, C., Topolnik, M., Đurina, V., Demirci, E., Kahveci, E.B., Gürbüz, A., Lukáš, O., Bartók, J., Gierlach, G., Hartman, F., Yilmaz, U., Doğan, M., Mandouh, M., Fragkoulis, M., Katsifodimos, A., 2021. Hazelcast Jet: Low-latency stream processing at the 99.99<sup>th</sup> percentile. *Proceedings of the VLDB Endowment* 14 (12), 3110–3121. <http://dx.doi.org/10.14778/3476311.3476387>.
- Hasselbring, W., 2018. Software architecture: Past, present, future. In: Gruhn, V., Striemer, R. (Eds.), *The Essence of Software Engineering*. Springer, pp. 169–184. [http://dx.doi.org/10.1007/978-3-319-73897-0\\_10](http://dx.doi.org/10.1007/978-3-319-73897-0_10).
- Hasselbring, W., 2021. Benchmarking as empirical standard in software engineering research. In: *Evaluation and Assessment in Software Engineering. EASE '21, Association for Computing Machinery*, pp. 457–462. <http://dx.doi.org/10.1145/3463274.3463361>.
- Hasselbring, W., Steinacker, G., 2017. Microservice architectures for scalability, agility and reliability in E-commerce. In: *Proceedings of the IEEE International Conference on Software Architecture Workshops*. pp. 243–246. <http://dx.doi.org/10.1109/ICSAW.2017.11>.
- Henning, S., 2023. Scalability benchmarking of cloud-native applications applied to event-driven microservices. In: *Kiel Computer Science Series, (2023/2), Department of Computer Science, Faculty of Engineering, Kiel University*, <http://dx.doi.org/10.21941/kcss/2023/2>.
- Henning, S., Hasselbring, W., 2021a. How to measure scalability of distributed stream processing engines? In: *Companion of the ACM/SPEC International Conference on Performance Engineering. ICPE '21, Association for Computing Machinery*, pp. 85–88. <http://dx.doi.org/10.1145/3447545.3451190>.
- Henning, S., Hasselbring, W., 2021b. Theodolite: Scalability benchmarking of distributed stream processing engines in microservice architectures. *Big Data Res.* 25, 100209. <http://dx.doi.org/10.1016/j.bdr.2021.100209>.
- Henning, S., Hasselbring, W., 2022a. A configurable method for benchmarking scalability of cloud-native applications. *Empir. Softw. Eng.* 27 (6), <http://dx.doi.org/10.1007/s10664-022-10162-1>.
- Henning, S., Hasselbring, W., 2022b. Demo paper: Benchmarking scalability of cloud-native applications with Theodolite. In: 2022 IEEE International Conference on Cloud Engineering (IC2E). pp. 275–276. <http://dx.doi.org/10.1109/IC2E55432.2022.00037>.
- Henning, S., Hasselbring, W., 2023. Replication package for: Benchmarking scalability of stream processing frameworks deployed as microservices in the cloud. <http://dx.doi.org/10.5281/zenodo.7497280>.
- Henning, S., Hasselbring, W., Burmester, H., Möbius, A., Wojcieszak, M., 2021. Goals and measures for analyzing power consumption data in manufacturing enterprises. *J. Data Inf. Manage.* 3 (1), 65–82. <http://dx.doi.org/10.1007/s42488-021-00043-5>.
- Herbst, N.R., Kounev, S., Reussner, R., 2013. Elasticity in cloud computing: What it is, and what it is not. In: *International Conference on Autonomic Computing. ICAC '13, USENIX*, pp. 23–27.
- Hesse, G., Lorenz, M., 2015. Conceptual survey on data stream processing systems. In: 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS). pp. 797–802. <http://dx.doi.org/10.1109/ICPADS.2015.106>.
- Hesse, G., Matthies, C., Glass, K., Huegle, J., Uflacker, M., 2019. Quantitative impact evaluation of an abstraction layer for data stream processing systems. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 1381–1392. <http://dx.doi.org/10.1109/ICDCS.2019.00138>.
- Hesse, G., Matthies, C., Perscheid, M., Uflacker, M., Plattner, H., 2021. ESPBench: The enterprise stream processing benchmark. In: *Proceedings of the ACM/SPEC International Conference on Performance Engineering. ICPE '21, Association for Computing Machinery*, pp. 201–212. <http://dx.doi.org/10.1145/3427921.3450242>.

- Hesse, G., Reissaus, B., Matthies, C., Lorenz, M., Kraus, M., Uflacker, M., 2018. Senska – towards an enterprise streaming benchmark. In: Nambiar, R., Poess, M. (Eds.), *Performance Evaluation and Benchmarking for the Analytics Era*. Springer International Publishing, pp. 25–40. [http://dx.doi.org/10.1007/978-3-319-72401-0\\_3](http://dx.doi.org/10.1007/978-3-319-72401-0_3).
- Karabey Aksakalli, I., Çelik, T., Can, A.B., Tekinerdoğan, B., 2021. Deployment and communication patterns in microservice architectures: A systematic literature review. *J. Syst. Softw.* 180, 111014. <http://dx.doi.org/10.1016/j.jss.2021.111014>.
- Karakaya, Z., Yazici, A., Alayyoub, M., 2017. A comparison of stream processing frameworks. In: 2017 International Conference on Computer and Applications (ICCA). pp. 1–12. <http://dx.doi.org/10.1109/COMAPP.2017.8079733>.
- Karimov, J., Rabl, T., Katsifodimos, A., Samarev, R., Heiskanen, H., Markl, V., 2018. Benchmarking distributed stream data processing systems. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE). pp. 1507–1518. <http://dx.doi.org/10.1109/ICDE.2018.00169>.
- Katsifodimos, A., Fragkoulis, M., 2019. Operational stream processing: Towards scalable and consistent event-driven applications. In: *Advances in Database Technology - 22nd International Conference on Extending Database Technology*. OpenProceedings.org, pp. 682–685. <http://dx.doi.org/10.5441/002/edbt.2019.86>.
- Kleppmann, M., Beresford, A.R., Svingen, B., 2019. Online event processing. *Commun. ACM* 62 (5), 43–49. <http://dx.doi.org/10.1145/3312527>.
- Kleppmann, M., Kreps, J., 2015. Kafka, Samza and the Unix philosophy of distributed data. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* 36 (4).
- Knoche, H., Hasselbring, W., 2019. Drivers and barriers for microservice adoption – A survey among professionals in Germany. *Enterpr. Model. Inf. Syst. Archit. (EMISAJ) – Int. J. Concept. Model.* 14 (1), 1–35. <http://dx.doi.org/10.18417/emisa.14.1>.
- Kounev, S., Lange, K.-D., von Kistowski, J., 2020. *Systems Benchmarking: For Scientists and Engineers, first ed.* Springer Publishing Company, Incorporated.
- Kratzke, N., Quint, P.-C., 2017. Understanding cloud-native applications after 10 years of cloud computing - A systematic mapping study. *J. Syst. Softw.* 126, 1–16. <http://dx.doi.org/10.1016/j.jss.2017.01.001>.
- Kreps, J., Narkhede, N., Rao, J., 2011. Kafka: A distributed messaging system for log processing. In: *Proceedings of the International Workshop on Networking Meets Databases*.
- Kulkarni, S., Bhagat, N., Fu, M., Kedigehalli, V., Kellogg, C., Mittal, S., Patel, J.M., Ramasamy, K., Taneja, S., 2015. Twitter Heron: Stream processing at scale. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15, Association for Computing Machinery, pp. 239–250. <http://dx.doi.org/10.1145/2723372.2742788>.
- Laigner, R., Zhou, Y., Salles, M.A.V., Liu, Y., Kalinowski, M., 2021. Data management in microservices: State of the practice, challenges, and research directions. *Proc. VLDB Endow.* 14 (13), 3348–3361. <http://dx.doi.org/10.14778/3484224.3484232>.
- Lehrig, S., Eikerling, H., Becker, S., 2015. Scalability, elasticity, and efficiency in cloud computing: A systematic literature review of definitions and metrics. In: *Proceedings of the 11th International ACM SIGSOFT Conference on Quality of Software Architectures*. QoSA '15, pp. 83–92. <http://dx.doi.org/10.1145/2737182.2737185>.
- Li, M., Tan, J., Wang, Y., Zhang, L., Salapura, V., 2015. SparkBench: A comprehensive benchmarking suite for in memory data analytic platform spark. In: *Proceedings of the 12th ACM International Conference on Computing Frontiers*. CF '15, Association for Computing Machinery, <http://dx.doi.org/10.1145/2742854.2742783>.
- Li, S., Zhang, H., Jia, Z., Zhong, C., Zhang, C., Shan, Z., Shen, J., Ali Babar, M., 2021. Understanding and addressing quality attributes of microservices architecture: A systematic literature review. *Inf. Softw. Technol.* 131, 106449. <http://dx.doi.org/10.1016/j.infsof.2020.106449>.
- Lopez, M.A., Lobato, A.G.P., Duarte, O.C.M.B., 2016. A performance comparison of open-source stream processing platforms. In: 2016 IEEE Global Communications Conference (GLOBECOM). pp. 1–6. <http://dx.doi.org/10.1109/GLOCOM.2016.7841533>.
- Lu, R., Wu, G., Xie, B., Hu, J., 2014. Stream Bench: Towards benchmarking modern distributed stream computing frameworks. In: 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. pp. 69–78. <http://dx.doi.org/10.1109/UCC.2014.15>.
- Margara, A., Cugola, G., Felicioni, N., Cilloni, S., 2022. A model and survey of distributed data-intensive systems. <http://dx.doi.org/10.48550/arxiv.2203.10836>.
- Nasiri, H., Nasehi, S., Goudarzi, M., 2019. Evaluation of distributed stream processing frameworks for IoT applications in smart cities. *J. Big Data* 6 (52), <http://dx.doi.org/10.1186/s40537-019-0215-2>.
- Noghabi, S.A., Paramasivam, K., Pan, Y., Ramesh, N., Bringham, J., Gupta, I., Campbell, R.H., 2017. Samza: Stateful scalable stream processing at LinkedIn. *Proc. VLDB Endow.* 10 (12), 1634–1645. <http://dx.doi.org/10.14778/3137765.3137770>.
- Pagliari, A., Huet, F., Urvoy-Keller, G., 2020. NAMB: A quick and flexible stream processing application prototype generator. In: 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). pp. 61–70. <http://dx.doi.org/10.1109/CCGrid49817.2020.00-87>.
- Pahl, C., Jamshidi, P., 2016. Microservices: A systematic mapping study. In: *Proceedings of the 6th International Conference on Cloud Computing and Services Science - Volume 1*. In: CLOSER 2016, SciTePress, pp. 137–146. <http://dx.doi.org/10.5220/0005785501370146>.
- Pahl, C., Jamshidi, P., Zimmermann, O., 2018. Architectural principles for cloud software. *ACM Trans. Internet Technol.* 18 (2), <http://dx.doi.org/10.1145/3104028>.
- Pfandzelter, T., Henning, S., Schirmer, T., Hasselbring, W., Bermbach, D., 2022. Streaming vs. Functions: A cost perspective on cloud event processing. In: 2022 IEEE International Conference on Cloud Engineering (IC2E). pp. 67–78. <http://dx.doi.org/10.1109/IC2E55432.2022.00015>.
- Qian, S., Wu, G., Huang, J., Das, T., 2016. Benchmarking modern distributed streaming platforms. In: 2016 IEEE International Conference on Industrial Technology (ICIT). pp. 592–598. <http://dx.doi.org/10.1109/ICIT.2016.7474816>.
- Ralph, P., bin Ali, N., Baites, S., Bianculli, D., Diaz, J., Dittrich, Y., Ernst, N., Felderer, M., Feldt, R., Filiari, A., de França, B.B.N., Furia, C.A., Gay, G., Gold, N., Graziotin, D., He, P., Hoda, R., Juristo, N., Kitchenham, B., Lenarduzzi, V., Martínez, J., Melegati, J., Mendez, D., Menzies, T., Moller, J., Pfahl, D., Robbes, R., Russo, D., Saarimäki, N., Sarro, F., Taibi, D., Siegmund, J., Spinellis, D., Staron, M., Stol, K., Storey, M.-A., Taibi, D., Tamburri, D., Torchiano, M., Treude, C., Turhan, B., Wang, X., Vegas, S., 2021. Empirical standards for software engineering research. <http://dx.doi.org/10.48550/arXiv.2010.03525>, Version 0.2.0.
- Sax, M.J., 2020. Performance Optimizations and Operator Semantics for Streaming Data Flow Programs (Ph.D. thesis). Humboldt-Universität zu Berlin, <http://dx.doi.org/10.18452/21424>.
- Sax, M.J., Wang, G., Weidlich, M., Freytag, J.-C., 2018. Streams and tables: Two sides of the same coin. In: *Proc. International Workshop on Real-Time Business Intelligence and Analytics*. BIRTE '18, pp. 1–10. <http://dx.doi.org/10.1145/3242153.3242155>.
- Shahverdi, E., Awad, A., Sakr, S., 2019. Big stream processing systems: An experimental evaluation. In: 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW). pp. 53–60. <http://dx.doi.org/10.1109/ICDEW.2019.00-35>.
- Shukla, A., Chaturvedi, S., Simmhan, Y., 2017. RIoT Bench: An IoT benchmark for distributed stream processing systems. *Concurr. Comput.: Pract. Exper.* 29 (21), e4257. <http://dx.doi.org/10.1002/cpe.4257>.
- Soldani, J., Tamburri, D.A., Van Den Heuvel, W.-J., 2018. The pains and gains of microservices: A systematic grey literature review. *J. Syst. Softw.* 146, 215–232. <http://dx.doi.org/10.1016/j.jss.2018.09.082>.
- Spæren, T., 2021. *Performance Analysis and Improvements for Apache Beam Master's thesis*. University of Oslo.
- Stopford, B., 2018. *Designing Event-Driven Systems, first ed.* O'Reilly Media, Inc..
- Taibi, D., Lenarduzzi, V., Pahl, C., 2017. Processes, motivations, and issues for migrating to microservices architectures: An empirical investigation. *IEEE Cloud Comput.* 4 (5), 22–32. <http://dx.doi.org/10.1109/MCC.2017.4250931>.
- Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J.M., Kulkarni, S., Jackson, J., Gade, K., Fu, M., Donham, J., Bhagat, N., Mittal, S., Ryabov, D., 2014. Storm@twitter. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD '14, Association for Computing Machinery, pp. 147–156. <http://dx.doi.org/10.1145/2588555.2595641>.
- Truong, T.M., Harwood, A., Sinnott, R.O., Chen, S., 2018. Performance analysis of large-scale distributed stream processing systems on the cloud. In: 2018 IEEE 11th International Conference on Cloud Computing (CLOUD). pp. 754–761. <http://dx.doi.org/10.1109/CLOUD.2018.00103>.
- Tucker, P., Tufte, K., Papadimos, V., Maier, D., 2010. NEXMark – A benchmark for queries over data streams (draft). URL <https://datalab.cs.pdx.edu/niagara/pstream/nexmark.pdf>.
- van Dongen, G., 2021. *Open Stream Processing Benchmark: An Extensive Analysis of Distributed Stream Processing Frameworks (Ph.D. thesis)*. Ghent University.
- van Dongen, G., van den Poel, D., 2020. Evaluation of stream processing frameworks. *IEEE Trans. Parallel Distrib. Syst.* 31 (8), 1845–1858. <http://dx.doi.org/10.1109/TPDS.2020.2978480>.
- van Dongen, G., van den Poel, D., 2021a. Influencing factors in the scalability of distributed stream processing jobs. *IEEE Access* 9, 109413–109431. <http://dx.doi.org/10.1109/ACCESS.2021.3102645>.
- van Dongen, G., van den Poel, D., 2021b. A performance analysis of fault recovery in stream processing frameworks. *IEEE Access* 9, 93745–93763. <http://dx.doi.org/10.1109/ACCESS.2021.3093208>.
- Vikash, Mishra, L., Varma, S., 2020. Performance evaluation of real-time stream processing systems for Internet of Things applications. *Future Gener. Comput. Syst.* 113, 207–217. <http://dx.doi.org/10.1016/j.future.2020.07.012>.
- Vogel, A., Henning, S., Ertl, O., Rabiser, R., 2023. A systematic mapping of performance in distributed stream processing systems. In: 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). pp. 293–300. <http://dx.doi.org/10.1109/SEAA60479.2023.00052>, In press..

- Wang, G., Chen, L., Dikshit, A., Gustafson, J., Chen, B., Sax, M.J., Roesler, J., Blee-Goldman, S., Cadonna, B., Mehta, A., Madan, V., Rao, J., 2021. Consistency and completeness: Rethinking distributed stream processing in Apache Kafka. In: Proceedings of the 2021 International Conference on Management of Data. In: SIGMOD/PODS '21, Association for Computing Machinery, pp. 2602–2613. <http://dx.doi.org/10.1145/3448016.3457556>.
- Wang, G., Koshy, J., Subramanian, S., Paramasivam, K., Zadeh, M., Narkhede, N., Rao, J., Kreps, J., Stein, J., 2015. Building a replicated logging system with Apache Kafka. Proc. VLDB Endow. 8 (12), 1654–1655. <http://dx.doi.org/10.14778/2824032.2824063>.
- Yang, S., Jeong, Y., Hong, C., Jun, H., Burgstaller, B., 2017. Scalability and state: A critical assessment of throughput obtainable on big data streaming frameworks for applications with and without state information. In: Euro-Par 2017: Parallel Processing Workshops. In: Lecture Notes in Computer Science, vol. 10659, pp. 141–152. [http://dx.doi.org/10.1007/978-3-319-75178-8\\_12](http://dx.doi.org/10.1007/978-3-319-75178-8_12).
- Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., 2016. Apache Spark: A unified engine for big data processing. Commun. ACM 59 (11), 56–65. <http://dx.doi.org/10.1145/2934664>.
- Zeuch, S., Monte, B.D., Karimov, J., Lutz, C., Renz, M., Traub, J., Breß, S., Rabl, T., Markl, V., 2019. Analyzing efficient stream processing on modern hardware. Proc. VLDB Endow. 12 (5), 516–530. <http://dx.doi.org/10.14778/3303753.3303758>.
- Zhang, S., He, B., Dahlmeier, D., Zhou, A.C., Heinze, T., 2017. Revisiting the design of data stream processing systems on multi-core processors. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE). pp. 659–670. <http://dx.doi.org/10.1109/ICDE.2017.119>.
- Zhang, Y., Xia, B., Wu, K., Liu, X., 2020. Building a better and faster Beam Samza runner. URL <https://engineering.linkedin.com/blog/2020/building-a-better-and-faster-beam-samza-runner>.
- Zhou, X., Li, S., Cao, L., Zhang, H., Jia, Z., Zhong, C., Shan, Z., Babar, M.A., 2023. Revisiting the practices and pains of microservice architecture in reality: An industrial inquiry. J. Syst. Softw. 195, 111521. <http://dx.doi.org/10.1016/j.jss.2022.111521>.