http://dx.doi.org/10.25673/111637

Experimental Validation of a RDMA-based High Availability Concept over a deterministic IP-based network

Thomas Kampa[1], Daniel Grossmann[2]

Abstract: The edge computing paradigm continues to disrupt areas such as the automation domain where hardware-based implementations, i.e., programmable logic controllers (PLCs), have dominated for decades. The shift to virtualization and data center technologies offers both new opportunities and risks. By providing the first experimental validation of Remote Direct Memory Access (RDMA) over a deterministic IP-based network, we converge the IT and OT domains with a high availability approach for virtualized PLCs. Synchronizing the state of two PLC instances via RDMA provides fault tolerance across multiple data centers and exemplifies the potential of data center technologies for the automation industry. The experimental validation demonstrates the excellent match between the communication requirements of RDMA and the characteristics of the deterministic network. Finally, the need for deterministic acyclic communication in the pursuit of IT/OT convergence is illustrated.

Keywords: DetNet; vPLC; Real-time

1 Introduction

IT/OT convergence has been the target of various research efforts in the past and continues to attract work, often with the goal of taking advantage of the mature and highly scalable IT infrastructure in the proprietary environment of the automation domain. A prominent example is the virtualization of workloads with real-time constraints. Often referred to as edge computing, it continues to make inroads into areas where bare-metal deployments were considered the only option.

In our previous work, we presented an approach for this class of applications to meet their real-time requirements in failover scenarios through state synchronization using Remote Direct Memory Access (RDMA) [KEG23]. The concept was applied to a redundant Programmable Logic Controller (PLC) with a hot standby application running in parallel to the active PLC on a different physical infrastructure component, providing fault tolerance in failure scenarios.

The goal of this work is the subsequent experimental validation with PROFINET RT communication and a deterministic IP network with jitter and latency bounds [Ba19]. Our

[1] Technische Hochschule Ingolstadt, AImotion Bavaria, Esplanade 10, 85049 Ingolstadt, Germany; AUDI AG, Auto-Union-Straße 1, 85057 Ingolstadt, Germany, thomas.kampa@audi.de
[2] Technische Hochschule Ingolstadt, AImotion Bavaria, Esplanade 10, 85049 Ingolstadt, Germany, daniel.grossmann@thi.de

previous work only measured the synchronization time as a function of state size, missing the holistic view of the end-to-end process [KEG23]. Furthermore, this work includes a first evaluation for RDMA over a deterministic network with the aim to verify the capabilities of the mentioned network under the load of lower prioritized network traffic, i.e., TCP/IP based communication.

This work is structured as follows: Section 2 provides related work in the topics RDMA, DetNet, and virtualization. Section 3 introduces the concept and requirements for the deterministic network. An experimental validation is performed in Section 4 and discussed in Section 5. Finally, Section 6 concludes this work and gives an outlook on further research opportunities.

2 Related Work

This section provides an introduction to the state of research for RDMA, DetNet, and virtualization in the OT domain.

Through the usage of RDMA, memory synchronization between two hosts is possible without the involvement of the CPU or caches of the respective host, reducing latency and improving throughput [Zh17]. Dropped or out of order packets lead to retransmission, which is why RDMA requires lossless transmission of frames. A possible solution is the Priority Flow Control (PFC) defined in IEEE 802.1Qbb [Mi18]. For RDMA to work over Ethernet, the term Converged Ethernet (CE) was coined to include properties such as priorization of RDMA over lower priority traffic. Today, CE poses the basis of current RDMA over Converged Ethernet (RoCE) implementations [Mi18][Zi23].

However, the use of PFC can lead to a network-wide deadlock. To mitigate this issue, a current research trajectory are improved RoCE Network Interface Cards (ICN) [Mi18]. Previous work has already pointed out that traffic uncertainty poses the biggest challenge in simultaneously balancing congestion control on one hand and achieving high throughput and deterministic latency on the other [Zh21].

DetNet, that is currently being standardized by the IETF, could be a suited technology to ensure deterministic behavior and congestion prevention for routed environments [IE22]. In our previous work we proposed RDMA over DetNet (RoDN) as the next natural evolution step of RoCE in the IT/OT convergence area [KEG23]. DetNet is promised to provide deterministic upper and lower bounds for packets and zero congestion due to planned resource allocation. Its characteristics constitute a perfect match with the properties of CE required for RoCE. The devices used in the validation are deterministic IP routers that have been proven to have comparable characteristics and support IE protocols [Ba19].

While RDMA is extensively being used in data centers, so far it has not reached the OT domain. This might change with the introduction of edge computing and the virtualization of critical services such as PLCs. The need for high availability and resilience in the automation

world has been historically immense, which is why the same properties will be required from a virtualized consolidated architecture as outlined in our previous work [KEG23].

3 Methods

This section introduces the high availability concept of a stateful real-time application with high availability requirements in the OT context, motivates the need for a deterministic IP-based network to connect data centers, and illustrates the communication requirements for this use case.

3.1 Synchronization Concept

In our previous work, we introduced a concept for high availability of vPLCs through state synchronization based on RDMA [KEG23]. The concept was applied to a redundant PLC with a hot standby application running in parallel to the active PLC on a different physical infrastructure component, providing fault tolerance in failure scenarios. The state of each application is synchronized every task cycle to keep the state of both instances consistent. Using RDMA over UDP-based synchronization reduced the average synchronization time by up to 99.39% for a modified software PLC from CODESYS. Fig. 1 shows a timing diagram of the synchronization process. It is important to note that the whole IEC task that is being synchronized must be completed before synchronization is performed. Also, the entire state is synchronized, even those variables that are not marked as persistent, to allow seamless failover.

3.2 RDMA over DetNet

In our previous work, we motivated the need for network-based scheduling of synchronization tasks in an architecture with consolidated vPLCs [KEG23]. Centralized management of the synchronization of multiple applications with their respective backups may be necessary to avoid burst scenarios and high retransmission rates, especially with UDP-based RoCE. Purely application-triggered synchronization is not desirable due to the lack of a holistic view of network and host synchronizations. A first step in this direction is to evaluate the suitability of RDMA over DetNet.

From a compatability perspective, RoCE should work without modifications for any given Ethernet-based DetNet implementation because it relies on UDP as the transport protocol. Prioritizing RDMA over lower-priority traffic is typically done by utilizing the priority bits in the Ethernet / IP header. An implementation based on holistic network-based scheduling is desirable, but out of scope for this work due to limitations in the configuration of deterministic IP routers and the timing of synchronization between vPLC instances.

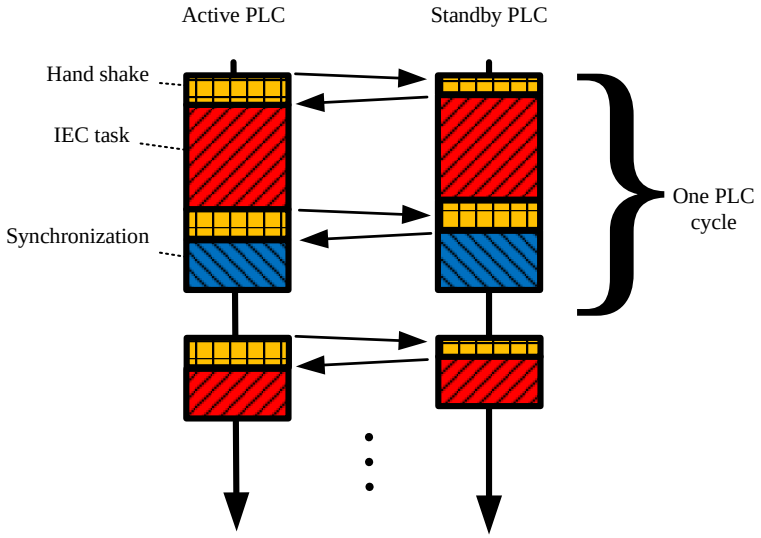


Fig. 1: Timing diagram of two PLCs, an active PLC and a standby PLC, synchronizing their states and cyclically computing a task.

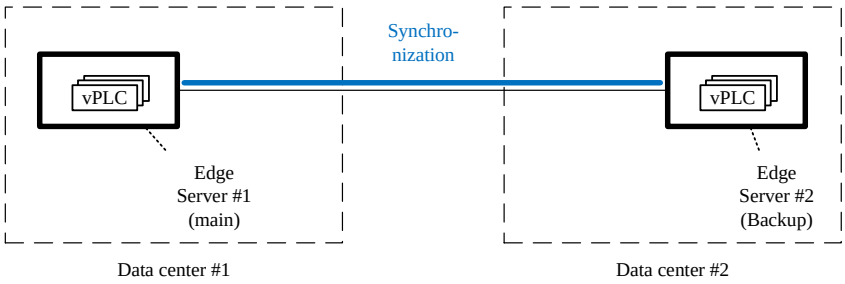
4 Experimental Validation

This section includes a description of the experimental setup, the test cases as well as the result.

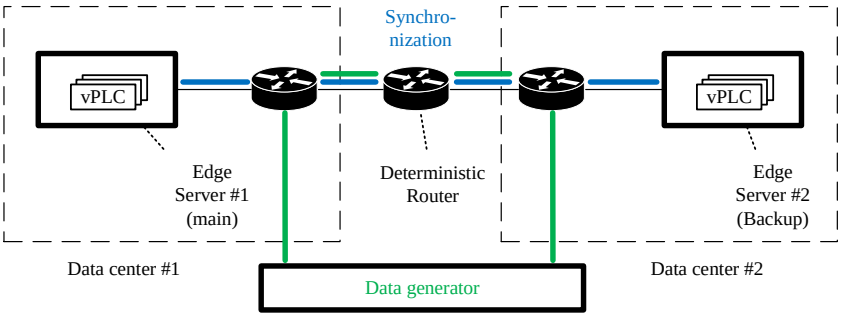
4.1 Setup

Three different test setups are defined to evaluate the suitability of a deterministic IP-based network for RDMA and PROFINET RT. Fig. 2a gives a baseline for synchronization between two hosts with no network nodes in between. Fig. 2b includes three deterministic IP routers to mimic inter-data center communication. Fig. 2c evaluates vPLCs synchronizing over one hop and communicating over three hops with their respective I/Os. We consciously avoided running RDMA in parallel to IE traffic on the same link, since we believe that in real deployments both traffic classes would transverse different dedicated physical paths.

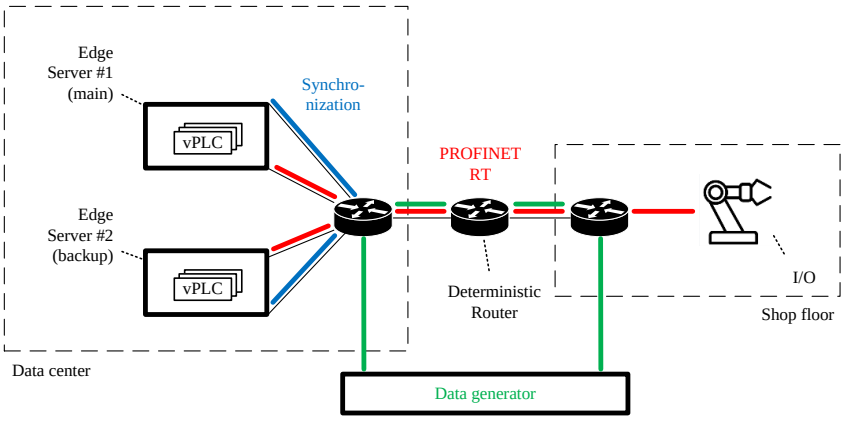
To evaluate the prioritization of PROFINET RT and RDMA-based traffic, best-effort traffic in form of UDP-packets generated by iPerf3, noted as data generator in the respective setups and generating enough traffic to fully utilize the whole bandwidth of the link between the routers. The uplinks of the edge servers have a 1 Gbit/s interface for RDMA and PROFINET, whereas the connection between the routers have a bandwidth of 10 Gbit/s.



(a) Test setup with two virtualized PLCs for high availability and no network in between.



(b) Test setup with two virtualized PLCs for high availability and a deterministic IP-based network connecting both for fault tolerance across data centers.



(c) Test setup with two virtualized PLCs for high availability and a deterministic IP-based network connecting both with PROFINET I/Os.

Fig. 2: Test cases with different network topologies and areas of interest.

Since DetNet is still in the standardization process, this work utilized deterministic IP networking routers that have been proven to provide determinism for real-time Industrial Ethernet (IE) traffic such as PROFINET [Ba19].

4.2 Results

Tab. 1 displays the values of the synchronization time as a function of state size and amount of network nodes in between, i.e., Fig. 2a equals zero network nodes and Fig. 2b equals three network nodes. It is apparent that the addition of three deterministic routers only increases the synchronization time by high two to low three-digit microseconds, whereas the Standard Deviation (SD) almost remains unchanged. It is important to note that the measurement had a millisecond accuracy limitation due to an implementation in the CODESYS vPLC.

The conducted validation after Fig. 2c resulted in no watchdog timers expiring with an update time of 1 ms and a watchdog timer of 3 ms.

Network nodes	State Size	Min	Max	Average	SD
0 network nodes	100 KB	0.00	2.00	0.957	0.205
3 network nodes	100 KB	1.00	3.00	1.019	0.136
0 network nodes	1 MB	8.00	10.00	8.978	0.146
3 network nodes	1 MB	9.00	10.00	9.056	0.231
0 network nodes	10 MB	89.00	90.00	89.224	0.417
3 network nodes	10 MB	89.00	90.00	89.274	0.446

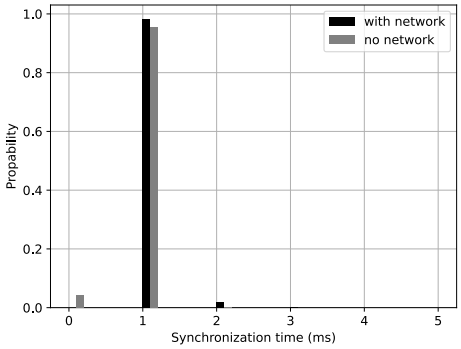
Tab. 1: Statistical analysis of the state synchronization times of the CODESYS PLC with either none or three network nodes in between. All values are in milliseconds if not described otherwise.

5 Discussion

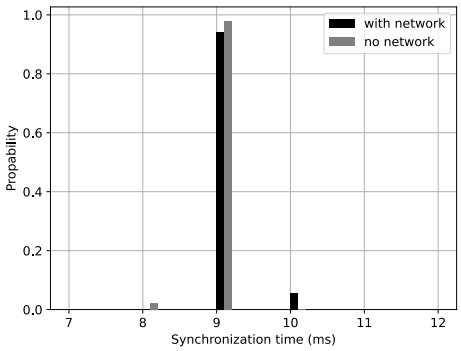
The following section discusses the presented results and provides further observations.

5.1 Meeting Real-Time Requirements of vPLCs

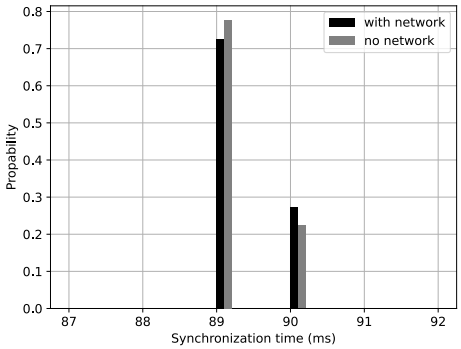
The envisioned concept is suitable for applications running on virtual machines and containers and could be an enabler for the virtualization of real-time critical applications such as control functions in the automation and process industries. As displayed in Section 4.2 the routed environment that mimics a data center connection fulfills the requirements in terms of real time and utilization. Moreover, PROFINET traffic gets prioritized over best-effort traffic and meets its stringent real-time requirements.



(a) State size = 100 KB



(b) State size = 1 MB



(c) State size = 10 MB

Fig. 3: Synchronization time with and without the deterministic network after 2a) and b).

5.2 Failover Times

Failover times, i.e., the time it takes for the standby PLC to take over from the failed active one, were not part of the conducted experiments. This is mostly due to the high variety of scenarios that one might construct, depending on setup and configuration. Possible things to consider are:

- The frequency of the heartbeat and the related timeout after which a PLC is deemed to be lost.
- The implementation of the redundancy mechanism, i.e., does the heartbeat only check in between tasks or continuously.
- The protocol involved, e.g., stateful (PROFINET) and stateless (EtherCAT) protocols.
- Whether the I/Os in the field require a reconnection, e.g., PROFINET S1 vs S2 devices.

We expect the involved failover time to be in the high double digit millisecond range or there might be no failover at all and a seamless takeover happens through the standby PLC.

5.3 Resiliency Options

Depending on the requirements of the application, different deployment options can be realized to achieve varying degrees of resiliency.

The concept presented in this work allows seamless failover between two hosts with no loss of state or interruption of the process, if the application and the respective I/Os allow it. However, these benefits come with a price to pay: The state is being synchronized every single PLC task cycle, leading to high bandwidth requirements for even small state sizes if the synchronization time is desired to be kept low. Moreover, a second PLC instance is running in a hot-standby mode at all times, leading to increased resources utilization and additional licenses from the PLC vendors.

Another, less costly option, provides the saving of the state into a database or share, which in turn could also be conducted with RDMA. This way, a new vPLC would have to be spun up and synchronized with the last known state before restarting operations. While this might safe an additional PLC running in parallel, it might lead to loss of data in the split seconds following the failure of the primary PLC, leading to possible increased recovery times.

Finally, only parts of the state could be saved and stored every cycle, reducing the load on the network and the time for saving even further. This could either be only the data that has changed since the last save, a method that we called partial synchronization in our previous

work [KEG23]. Another prime target could be persistent data, which is often written into a persistent storage in case of a power outage of the PLC. This albeit small change could already reduce the synchronization time mutliple times, increasing performance and reducing network load.

5.4 Network Utilization Pattern

PLC task cycles can vary in their duration time due to co-routines, varying hardware utilization and other external effects, especially in a virtualized environment. The synchronization of the state is conducted after completing a full task as described in Fig. 1. Varying task cycle times do not effect the network utilization due to PLC - I/O communication as shown in Fig. 4. However, converging the communication of the PLCs with their respective I/Os and the synchronization of PLCs in between creates an remarkable network utilization pattern which is illustrated in the lower portion of Fig. 4. This illustration shows that acyclic communication patterns need to coexist with cyclic communication patterns, both requiring deterministic real-time properties.

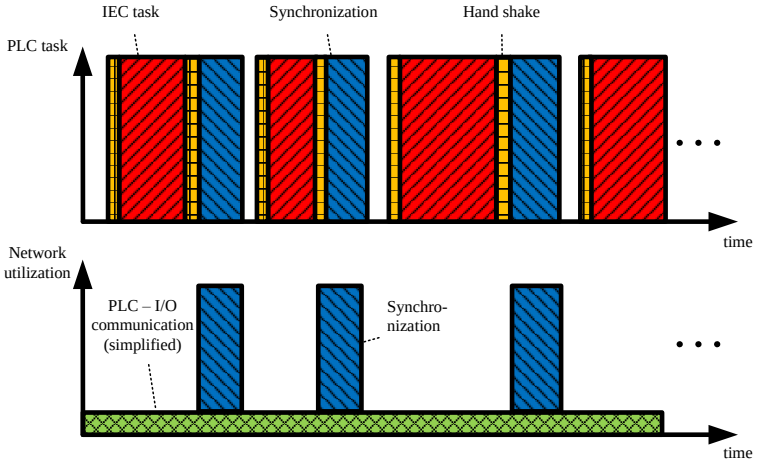


Fig. 4: How variable IEC task duration can lead to acyclic communication patterns with real-time constraints in synchronization scenarios.

6 Conclusion

In this work, we have performed an experimental validation of RDMA over a deterministic IP-based network. A high-availability concept for vPLCs is used as an evaluation example to highlight the requirements for this specific combination of technologies. The results prove that deterministic networking complements RDMA well, even under high network

utilization, and poses a suitable alternative to pure RoCE to guarantee packet delivery and enable deterministic behavior.

Future research may focus on evaluating other resiliency options and improving state synchronization as well as evaluating the failover time across different deployment options. In addition, acyclic communication patterns need more attention in the automation domain as they may emerge as another class of traffic with similar deterministic requirements as cyclic real-time traffic.

Acknowledgment

The presented concept was developed within the cluster “Digital Production” of the Research Institute AImotion Bavaria of the Technische Hochschule Ingolstadt. The project is funded by AUDI AG. We thank Dr. Zhe Lou and Sushrey Sunil Sawant for providing the Deterministic IP networking setup.

References

[Ba19] Badar, A.; Lou, D. Z.; Graf, U.; Barth, C.; Stich, C.: Intelligent Edge Control with Deterministic-IP based Industrial Communication in Process Automation. In (Lutfiyya, H., Hrsg.): 15th International Conference on Network and Service Management: 1st International Workshop on Analytics for Service and Application Management (AnServApp 2019) : International Workshop on High-Precision Networks Operations and Control, Segment Routing and Service Function Chaining (HiPNet + SR/SFC 2019) : October 21-25 2019, Halifax, Canada. IEEE, [Piscataway, NJ], 2019, ISBN: 9783903176249.

[IE22] IETF: Deterministic Networking (detnet), 2022, URL: https://datatracker.ietf.org/wg/detnet/documents/, Stand: 09. 10. 2023.

[KEG23] Kampa, T.; El-Ankah, A.; Grossmann, D.: High Availability for virtualized Programmable Logic Controllers with Hard Real-Time Requirements on Cloud Infrastructures. In: 2023 IEEE 21st International Conference on Industrial Informatics (INDIN). IEEE, 2023.

[Mi18] Mittal, R.; Shpiner, A.; Panda, A.; Zahavi, E.; Krishnamurthy, A.; Ratnasamy, S.; Shenker, S.: Revisiting network support for RDMA. In: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. ACM, New York, NY, USA, 2018.

[Zh17] Zhang, Y.; Gu, J.; Lee, Y.; Chowdhury, M.; Shin, K. G.: Performance Isolation Anomalies in RDMA. In: Proceedings of the Workshop on Kernel-Bypass Networks. 2017.

[Zh21] Zhang, J.; Zhang, Y.; Guan, Z.; Wan, Z.; Xia, Y.; Pan, T.; Huang, T.; Tang, D.; Lin, Y.: HierCC: Hierarchical RDMA Congestion Control. In: 5th Asia-Pacific Workshop on Networking (APNet 2021). S. 29–36, 2021.

[Zi23] Zilong Wang; Layong Luo; Qingsong Ning; Chaoliang Zeng; Wenxue Li; Xinchen Wan; Peng Xie; Tao Feng; Ke Cheng; Xiongfei Geng; Tianhao Wang; Weicheng Ling; Kejia Huo; Pingbo An; Kui Ji; Shideng Zhang; Bin Xu; Ruiqing Feng; Tao Ding; Kai Chen; Chuanxiong Guo: SRNIC: A Scalable Architecture for RDMA NICs. In. S. 1–14, 2023, ISBN: 978-1-939133-33-5, URL: https://www.usenix.org/conference/nsdi23/presentation/wang-zilong.