



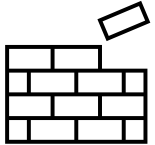

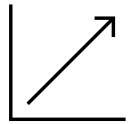

Machine Learning-based Deep Packet Inspection at Line Rate for RDMA on FPGAs

Maximilian J. Heer, Benjamin Ramhorst,
Gustavo Alonso

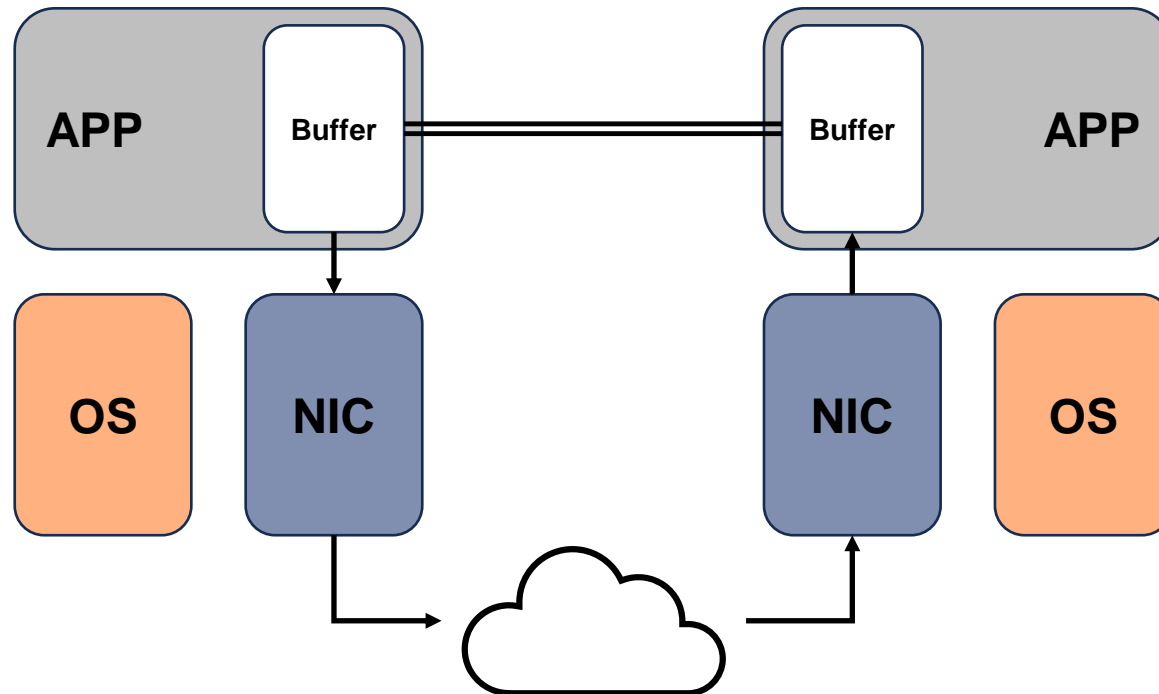
{maximilian.heer, benjamin.ramhorst, alonso}@inf.ethz.ch



Outline

- What is RDMA / RoCE? 
- Deep Packet Inspection for RDMA on FPGAs 
- Architectural Design 
- ML-based DPI 
- Evaluation and network performance 
- Conclusion & Future Work 

What is RDMA / RoCE?



RDMA = Remote Direct Memory Access via the network.

Performance relies on three principles:

- Host-Bypassing
- Zero-Copy
- Polling

High throughput

Low latency

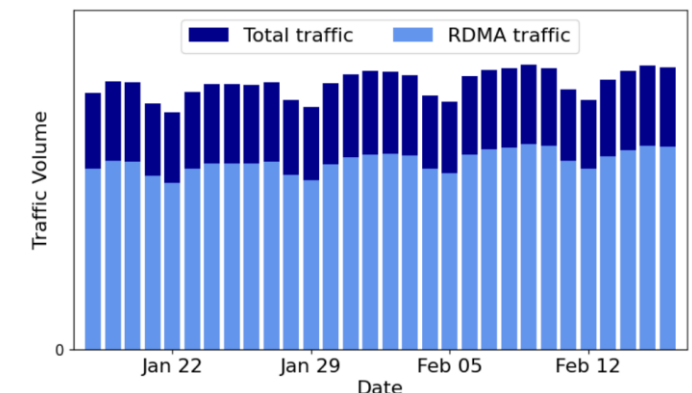
Low CPU util.

Asynchronous ops

Comes from HPC-environments
using the InfiniBand-Hardware

RoCE = RDMA over Converged Ethernet

Combines the advantages of
RDMA with the ease of use of
Ethernet connections



Graphic taken from „Empowering Azure Storage with RDMA“ (W. Bai et al., NSDI 23, <https://www.usenix.org/conference/nsdi23/presentation/bai>)

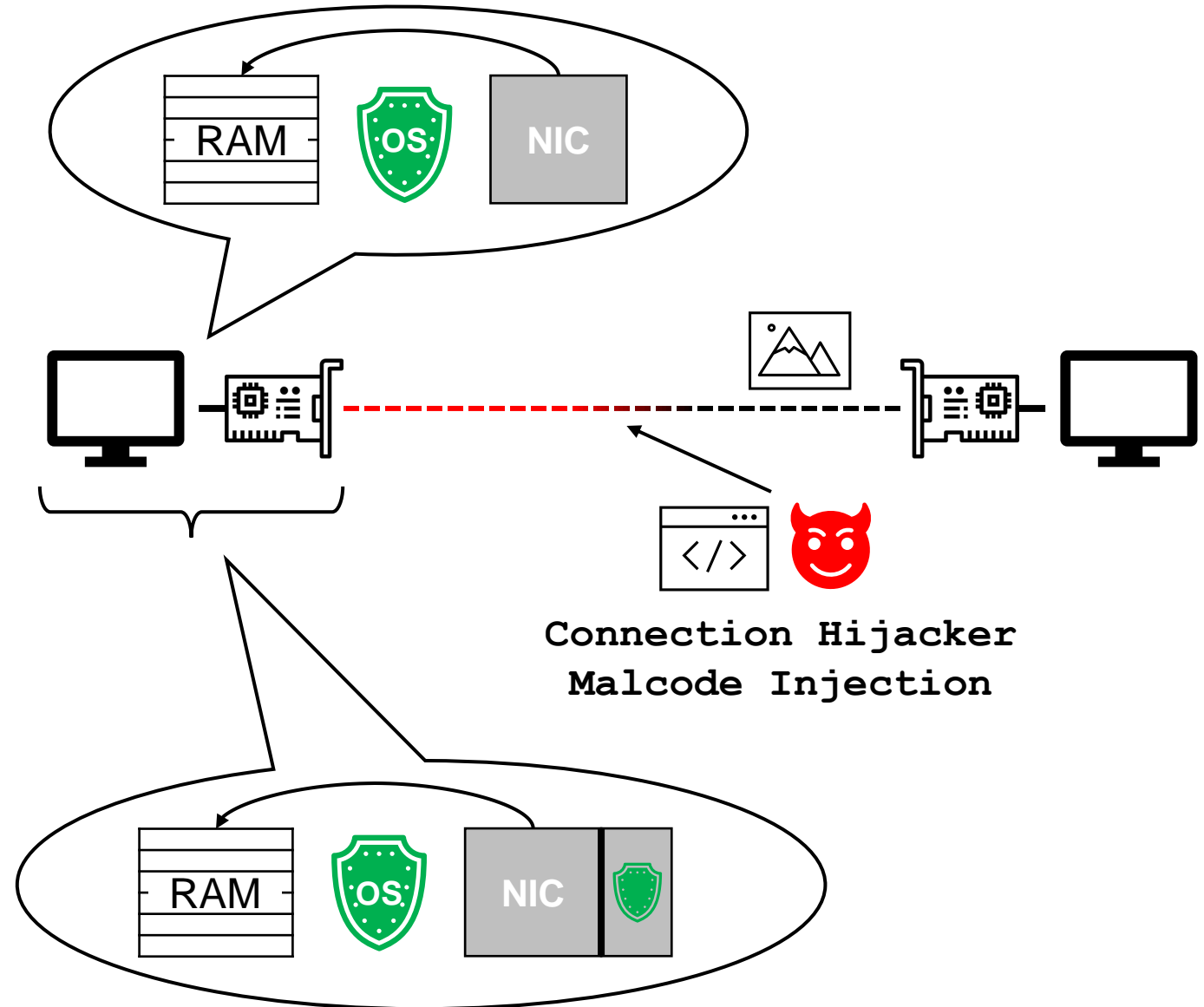
1. Host Bypassing includes bypassing of all OS-enabled security and access-control mechanisms.

2. The RDMA-standard lacks encryption and cryptographic authentication.

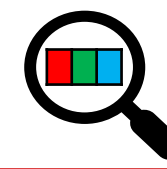
3. Existing RDMA-security features are weak, so that RDMA-connections are easy to hijack and spoof.

Solution:

Integrate security and access control into the **SmartNIC** to maintain performance and gain safety!



Deep Packet Inspection (DPI)



Packet Inspection: Networking Gear (NIC) extracts information from incoming packets and makes a routing decision based on the gathered information (i.e. forward or drop).

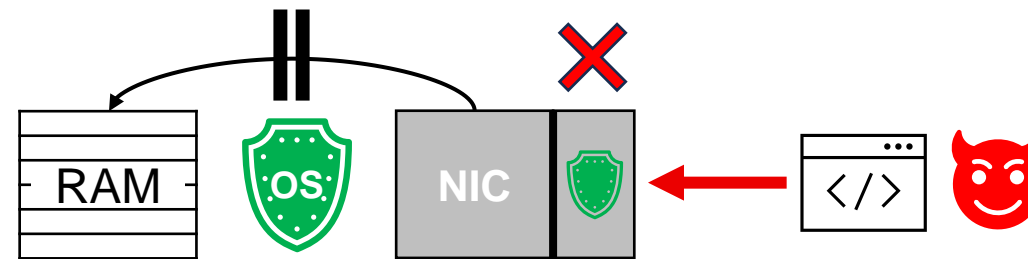


Headers (*Standard Packet Inspection*)

- ✓ Structured information, easy to check pre-defined header fields
- ✓ Allows for meta-checks
- ✗ Not very helpful in detecting elaborated attack schemes

Payload (*Deep Packet Inspection*)

- ✓ Expressive information, especially in advanced attack schemes
- ✓ Allows for huge variety of meta-checks
- ✗ Unstructured information, hard to analyze in real-time without performance-burden

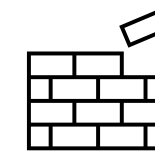


ML-DPI-model in NIC analyzes incoming traffic payload for executable code.
If detected, packet is dropped.

Main challenges:

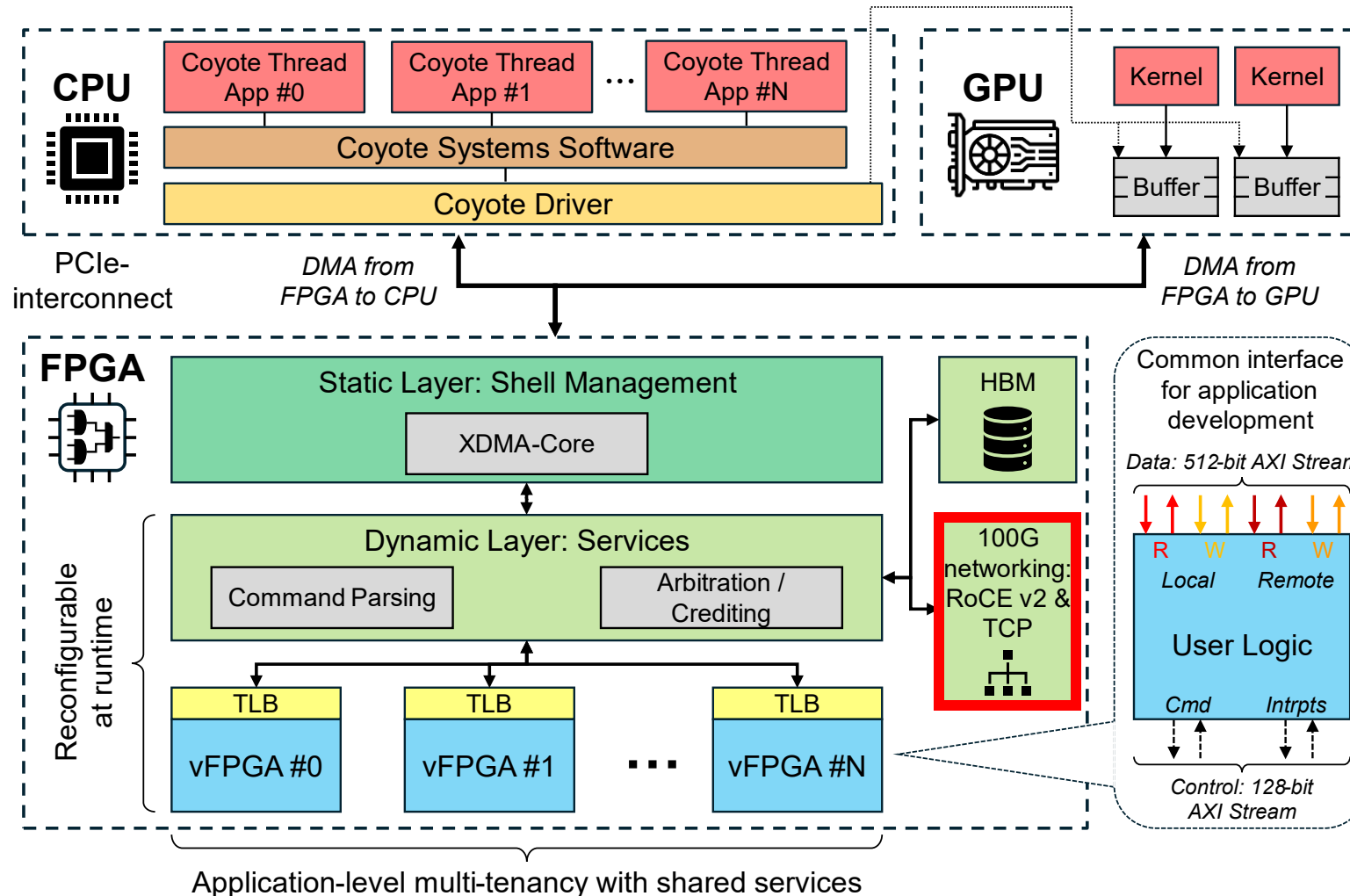
- DPI must happen at **line rate** to prevent performance loss
- DPI must have a **high accuracy** to add security and prevent unnecessary drops through false negatives

Balboa: RDMA-Stack taken from COYOTE



Coyote: Our advanced, open-source FPGA-shell with abstractions for multi-tenancy, memory-management and networking

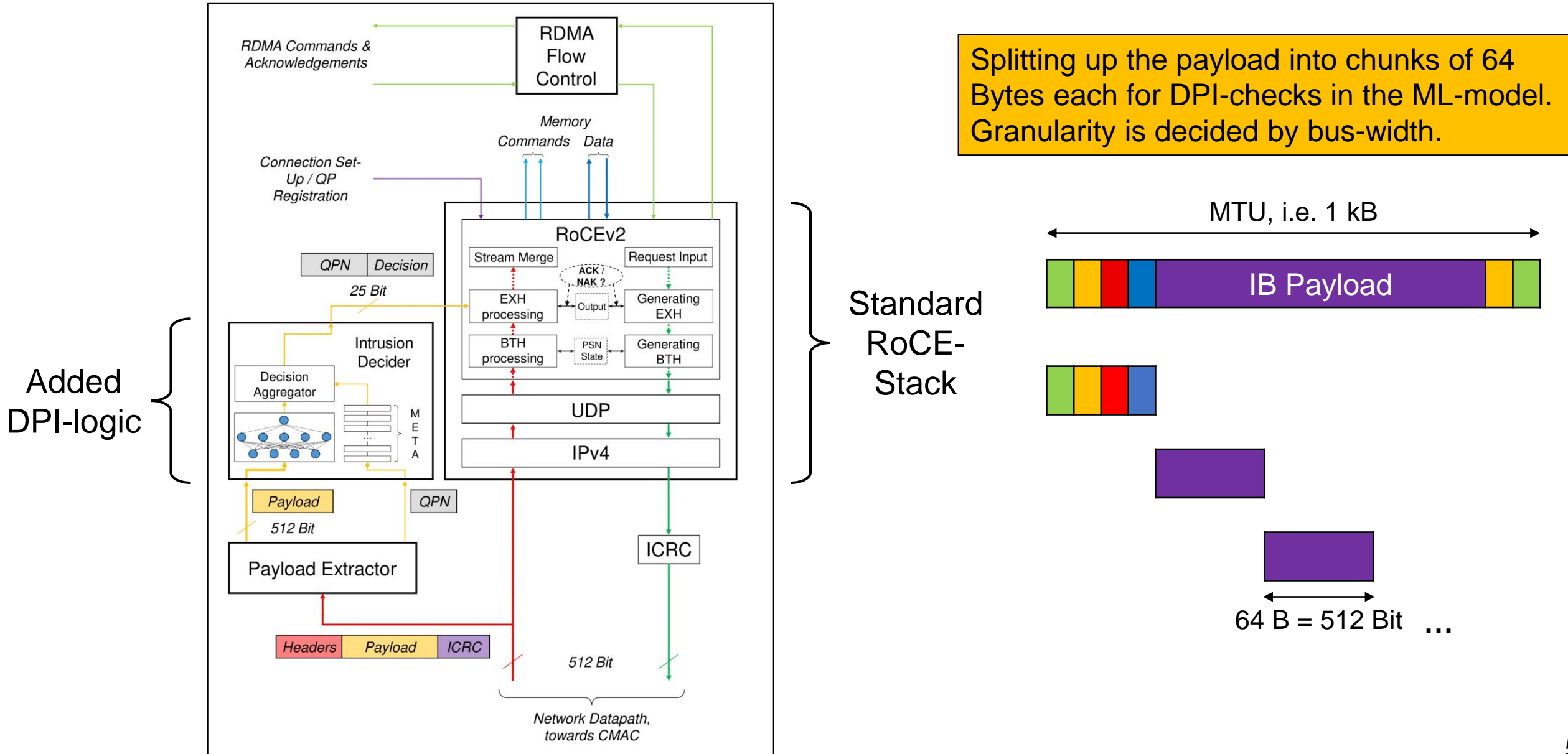
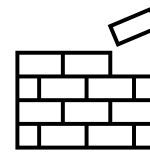
- GitHub-Repo: <https://github.com/fpgasystems/Coyote/tree/tutorial>
- Paper: <https://capra.cs.cornell.edu/latte25/paper/3.pdf>



BALBOA is a RDMA / RoCE-stack for FPGAs, which...

- ... implements RDMA READ, WRITE and SEND
- ... is fully RoCE-v2 compliant and allows communication with standard NICs (Mellanox etc.).
- ... supports 100 Gbit/s networking and provides low-latency comparable to ASIC-based NICs.
- ... comes with an API in the style of InfiniBand Verbs.

Hardware architecture



Key challenges with ML-DPI



- High operating frequency:
 - **250 MHz (4ns)** required to achieve **100Gbps** throughput
- Low initiation interval:
 - Model must accept one new input each clock cycle (**ii = 1cc**)
- Reconfigurability:
 - Hardware can be upgraded, accommodating for new threats
- High dimensionality:
 - Each input vector has **512 dimensions**



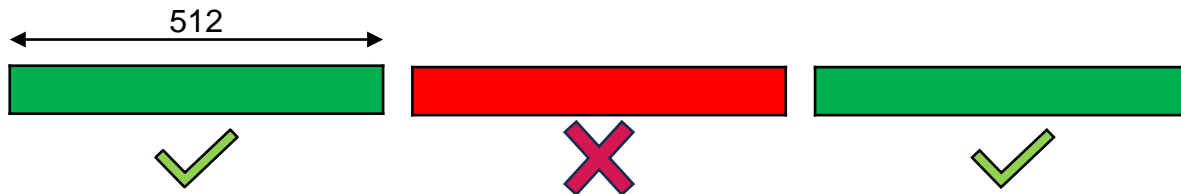
- To achieve reconfigurability:
→ use hls4ml
- To achieve low initiation interval:
→ unroll all multiplications
- To mitigate high parallelism:
→ Increase Vitis partition factor
- To meet 250MHz and fit within the resource budget:
→ Symbolic regression and precisions < 10 bits

Quantized neural networks

- 3-layer fully connected neural network
- 32 neurons in the 1st layer, to reduce dimensionality without sacrificing accuracy
- Remaining layers have 64 units with quantized ReLU activation
- One output with sigmoid activation
- Minimizing binary cross-entropy with Adam and LR reduction, early stopping

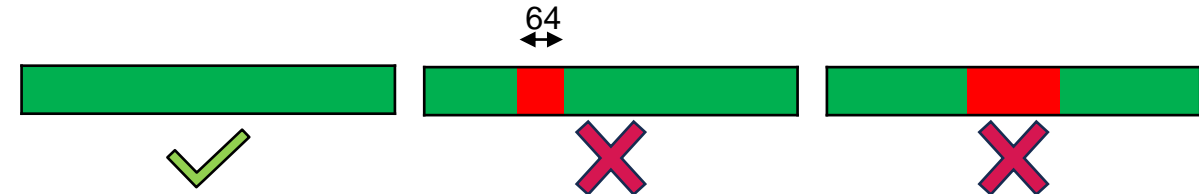
Accuracy: Ability to correctly flag payloads that contain potentially malicious executables

Full 512-bit chunks of executables



	Accuracy	FPR	FNR
SR	95.31 ± 0.33	5.04 ± 0.53	4.33 ± 0.35
Ternary	97.83 ± 0.16	1.74 ± 0.37	2.59 ± 0.38

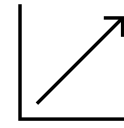
Executable embedded in 512-bit chunks



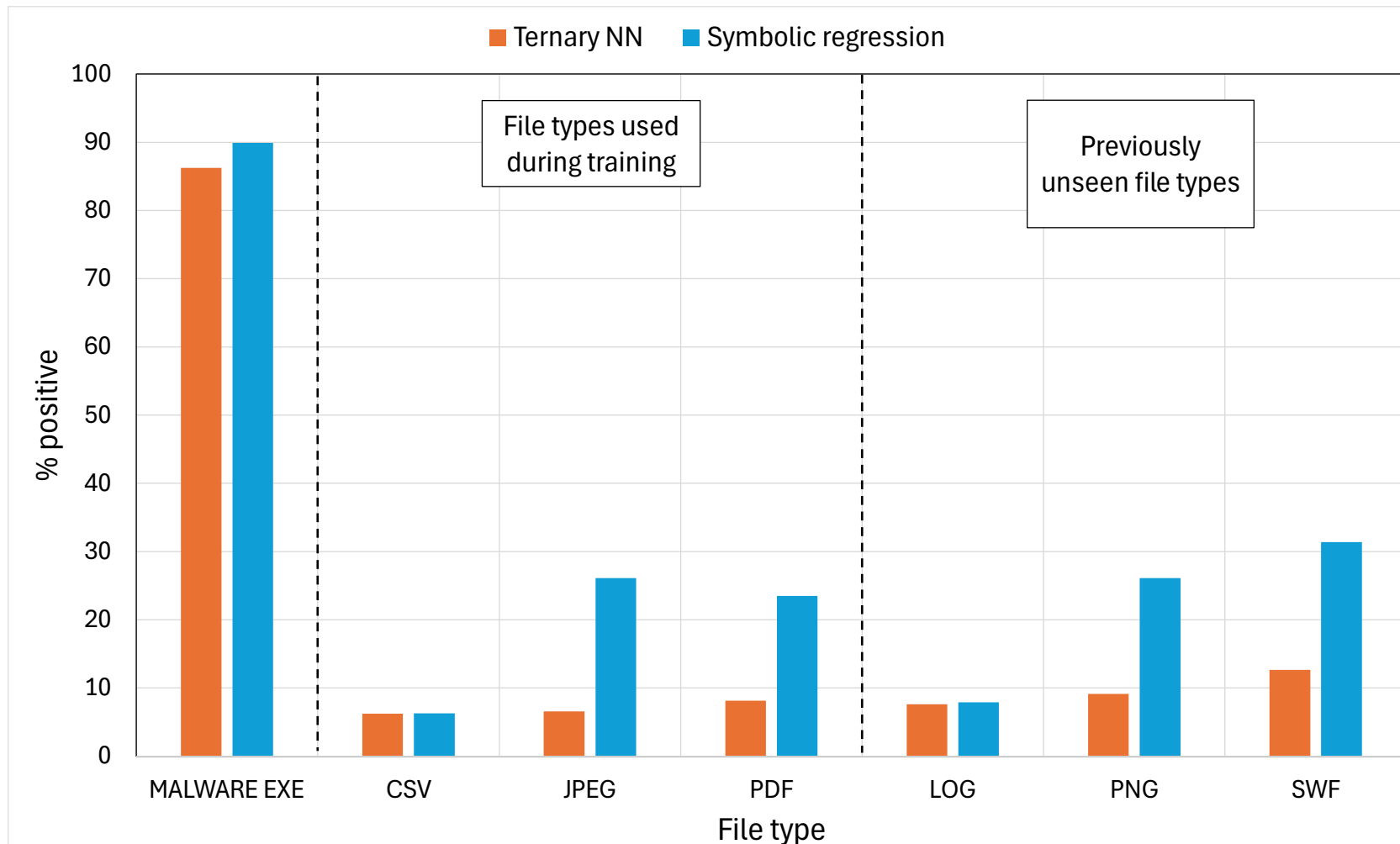
	Accuracy	FPR	FNR
SR	83.47 ± 0.57	13.11 ± 0.72	19.95 ± 0.58
Ternary	89.36 ± 0.39	6.25 ± 1.37	15.04 ± 0.86

For full chunks, both models provide sufficient accuracy. In the more complicated case of embedded threats, the ternary neural network clearly outperforms Symbolic Regressions in terms of accuracy and especially FPR / FNR.

Model generalizability



Evaluation of DPI on full files, including malware, training material and unseen data types. Again, TNN outperforms SR in the edge cases quite clearly.

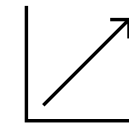


Model	Latency	ii
Ternary	44ns	1cc
SR	24ns	1cc

Model	LUT [%]	FF [%]	DSP [%]	BRAM [%]
Ternary	30,062 [2.3%]	11,363 [0.4%]	0 [0%]	0 [0%]
SR	472 [0.04%]	271 [0.01%]	2 [0.02%]	4 [0.05%]

Resources & latency: Both models achieve a minimal resource footprint (<1%) and latency (50ns). Results are reported post-Place and Route on AMD Alveo U55C.

Network performance evaluation

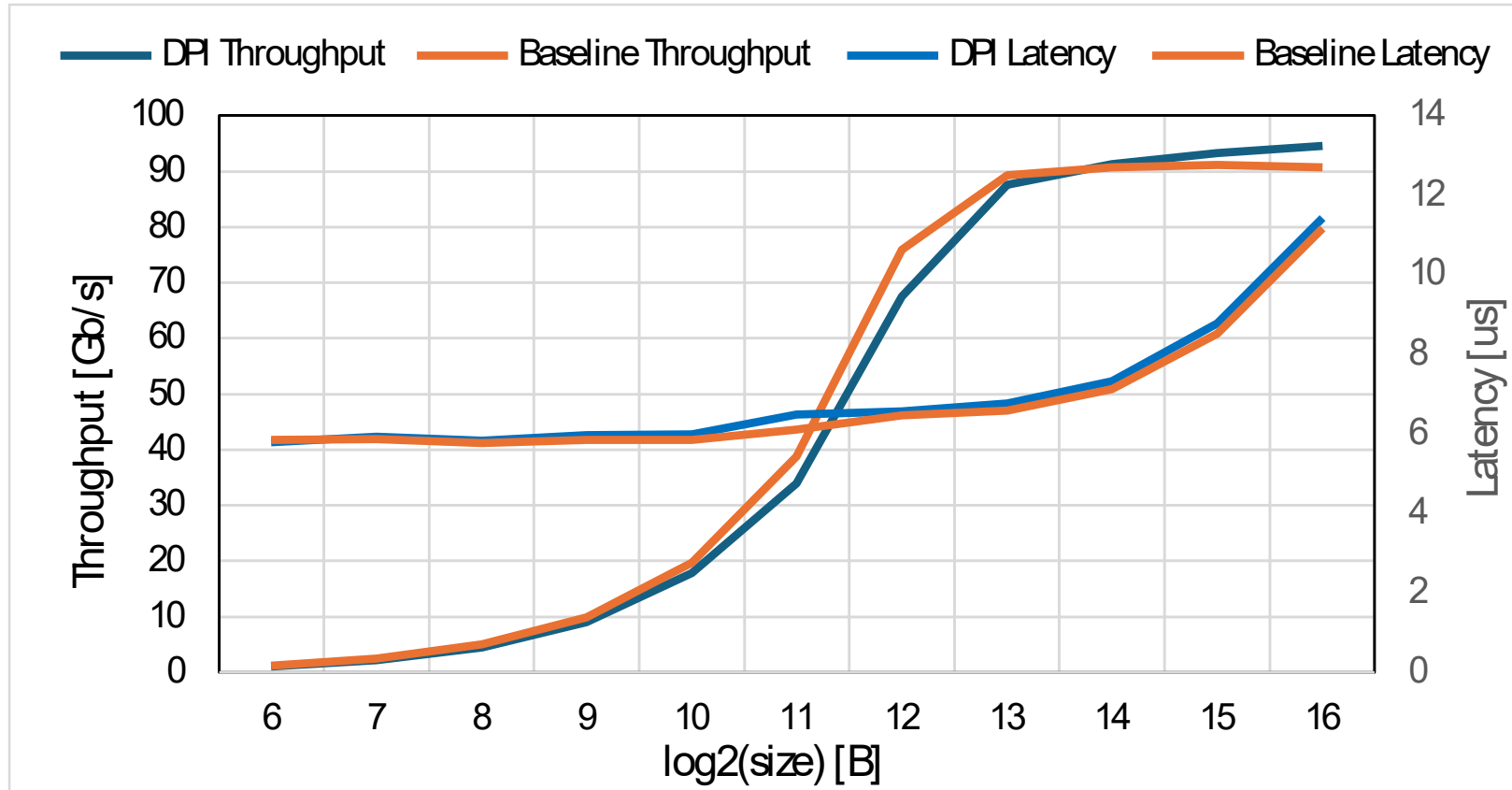


Latency

Exchange of single messages, ping-pong-style

Throughput

Batched message-transmission



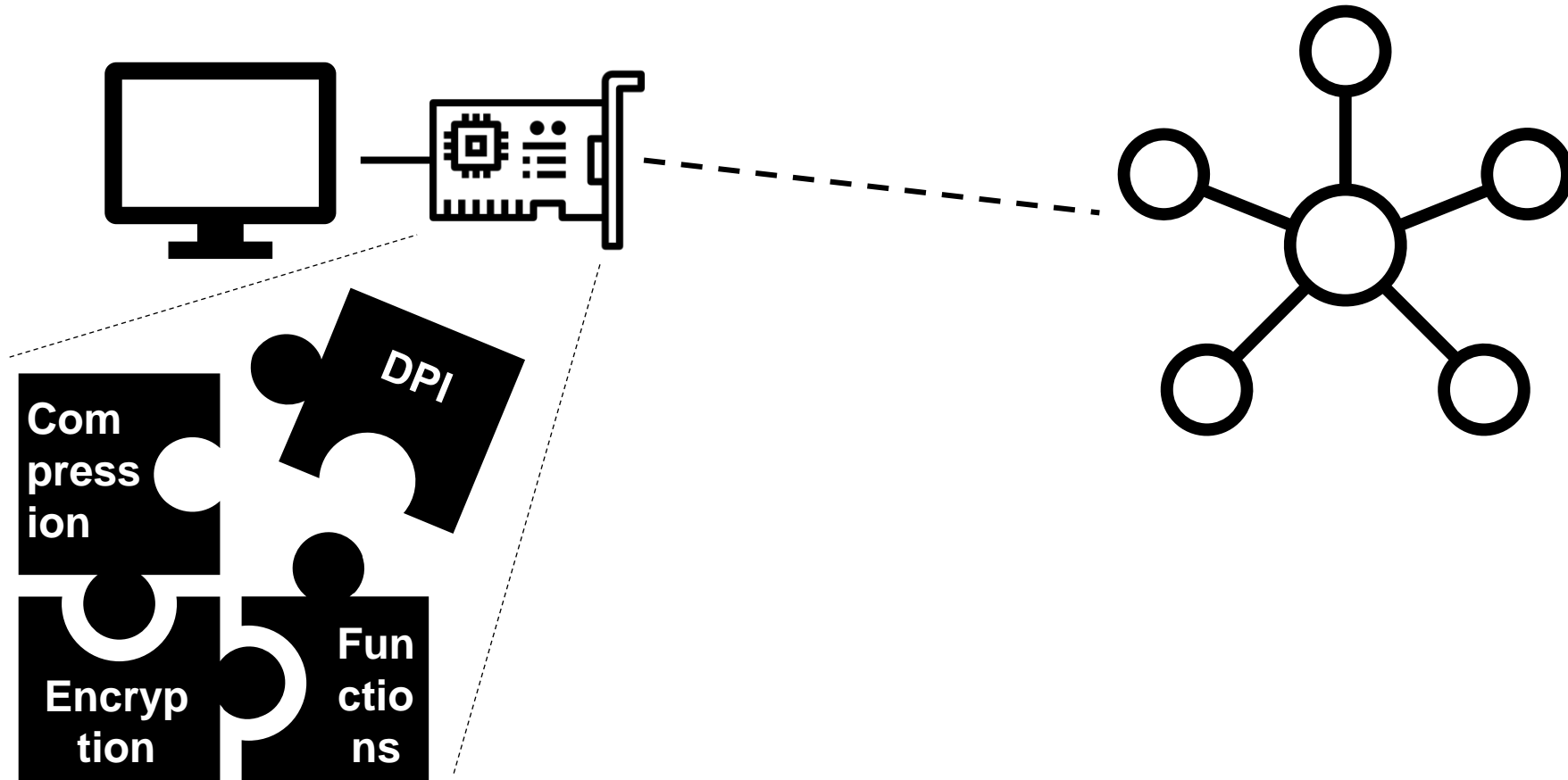
Basically **no performance deterioration** visible through the integration of DPI: Our DPI performs at **line rate** and does not introduce a performance bottleneck for performance benchmarking.



Contributions:

- Integration of ultra-low latency neural networks and symbolic regression directly on the network datapath; **first contribution for RDMA traffic**
- Achieving **100Gbps throughput** while consuming less than **1% of all available resources**
- Models **generalize to previously unseen data types** as well as systems with different configurations
- Clearly identifying **advantages of FPGA-NICs**, since the contributions would not be possible with ASIC-based NICs.

Adding more features to an FPGA-based AI SmartNIC, such as encryption, compression and function offloading (e.g., data pre-processing, gradient compression)



Thank you for your attention!

Questions?

We would like to thank **AMD** for their generous donation of the **Heterogeneous Accelerated Compute Clusters (HACC)** at ETH Zurich (<https://systems.ethz.ch/research/data-processing-on-modern-hardware/hacc.html>), on which the FPGA experiments and GPU model training were conducted.

Link to our open-sourced RDMA-stack:
<https://github.com/fpgasystems/fpga-network-stack>

Data sources

- [1] V. Ljubovic, “Programming homework dataset for plagiarism detection,” 2020. [Online]. Available: <https://dx.doi.org/10.21227/71fw-ss32>
- [2] “TheAlgorithms/C-plus-Plus: Collection of various algorithms in mathematics, Machine Learning, computer science and physics implemented in C++ for educational purposes.” [Online]. Available: <https://github.com/TheAlgorithms/C-Plus-Plus>
- [3] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt, “Bringing science to digital forensics with standardized forensic corpora,” Digital Investigation, vol. 6, 09 2009.
- [4] “Digital Corpora: corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/.” [Online]. Available: <https://corp.digitalcorpora.org/corpora/files/CC-MAIN-2021-31-PDF-UNTRUNCATED/>
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255[6]
- [6] “Huge collection of all algorithms implemented in multiple languages.” [Online]. Available: <https://github.com/AllAlgorithms/cpp>
- [7] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011. [Online]. Available: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf