The Future of Network Management (of the Future)

An Overview of Low-Latency, High-Throughput AI with Programmable Devices

Luciano Paschoal Gaspary - UFRGS, Brazil

paschoal@inf.ufrgs.br • http://www.inf.ufrgs.br/~paschoal • @lpgaspary

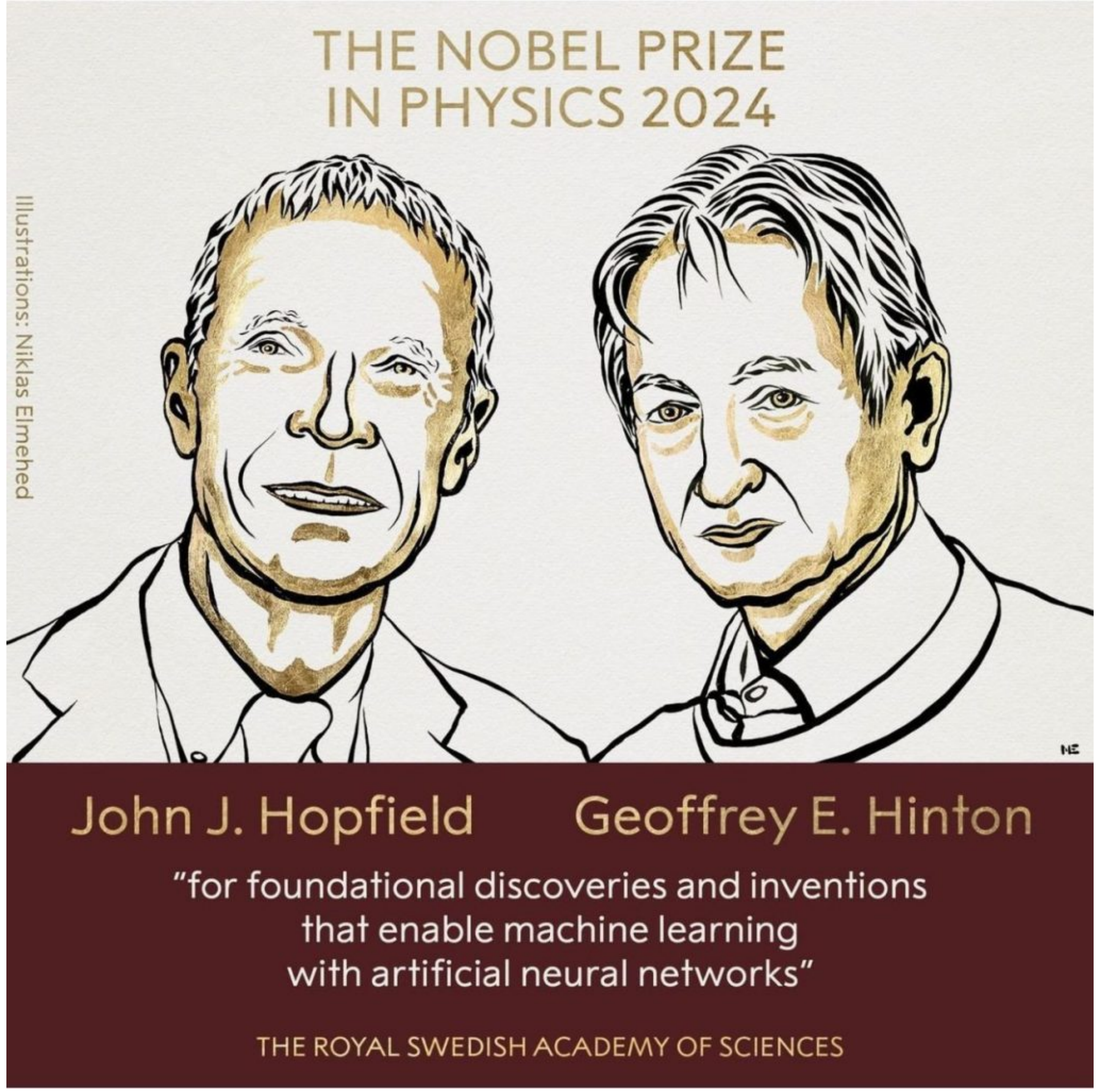
WTR 2025 • 22 October 2025 • Porto Alegre

Initial disclaimer & scope

- The scope is:
  - AI/ML for networks, **not networks for AI/ML**
  - aspects that can be addressed mainly in the data plane, **not in the control or management planes** (multi-dimensional problem)

Initial disclaimer & scope

- The scope is:
  - AI/ML for networks, **not networks for AI/ML**
  - aspects that can be addressed mainly in the data plane, **not in the control or management planes** (multi-dimensional problem)
- In this talk, I sometimes take a **provocative look** at AI/ML for networking :-)



Source: Nobel Foundation

The big “race” towards AI/ML for networks

- The race towards integrating Artificial Intelligence and Machine Learning into Computer Networks is very intense

- If you Google these keywords, you will uncover millions of studies (>2.0M)

The big "race" towards AI/ML for networks

- The race towards integrating Artificial Intelligence and Machine Learning into Computer Networks is very intense

- If you Google these keywords, you will uncover millions of studies (>2.0M)

- Are there **clear "killer" networking apps** that are in need of more AI/ML? Are we lagging behind?

Source: https://www.modernanalyst.com/Resources/BusinessAnalystHumor/tabid/218/ID/5392/Solutions_in_Search_of_Problems.aspx

Why AI/ML for networks?

- What are the “killer” apps for AI/ML? Potentially the ones “*where we are already implicitly employing machine learning, maybe badly*” (Schapira 2023)

Detection of <you name it> from data	Identifying existing **patterns** or **anomalies** in data, often in the present or recent past
Prediction of <you name it> from data	Forecasting **future outcomes** or **events** based on historical data

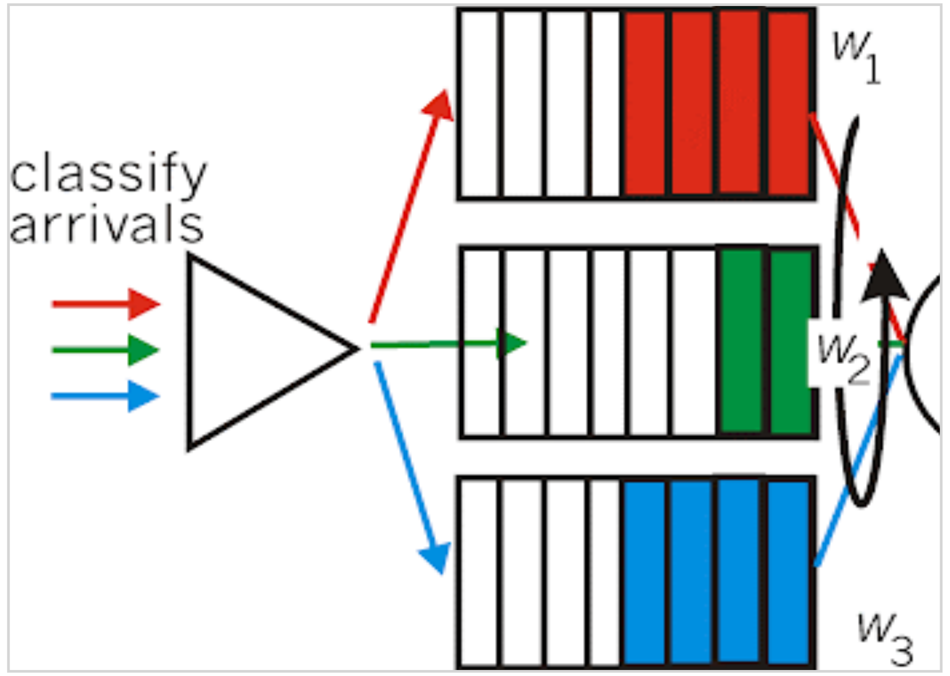
Raouf Boutaba, M.A. Salahuddin, Noura Limam, *et al.*: **A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities**. J Internet Serv Appl 9, 16 (2018).
Michael Schapira: **AI for networking, and networking for AI**. The Networking Channel, 2023. Available at https://www.youtube.com/watch?v=i6DvbfIUPSg.

Why AI/ML for networks?

Carry out **intrusion detection** by looking at unexpected events

Source: https://towardsdatascience.com

Source: Kurose and Ross, 2020.

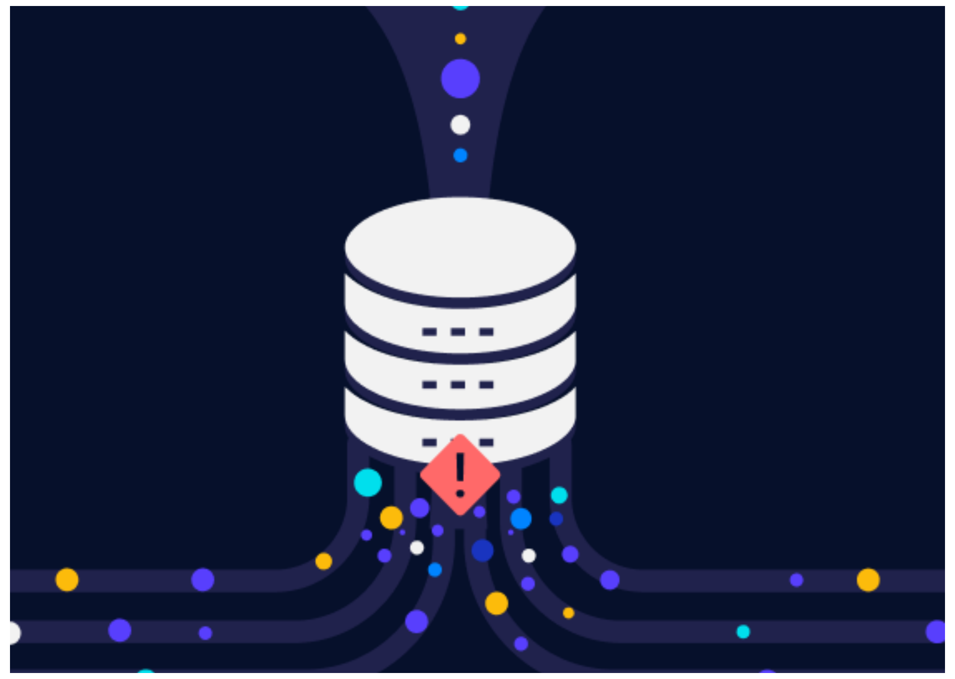


Perform **traffic classification** based on patterns existing on network packets

Execute **traffic engineering** based on predicted future traffic demands

Source: https://engineering.nyu.edu

Source: https://granulate.io



Run **congestion control** based on the prediction of the bottleneck bandwidth

Make **video streaming** decisions based on predicted download times of video chunks

Source: https://www.forbes.com

+ **resource management** + **fault management** + **channel modeling** + ...

Raouf Boutaba, M.A. Salahuddin, Noura Limam, *et al.*: **A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities.** J Internet Serv Appl 9, 16 (2018).
Michael Schapira: **AI for networking, and networking for AI.** The Networking Channel, 2023. Available at https://www.youtube.com/watch?v=i6DvbfIUPSg.

What are we striving for?

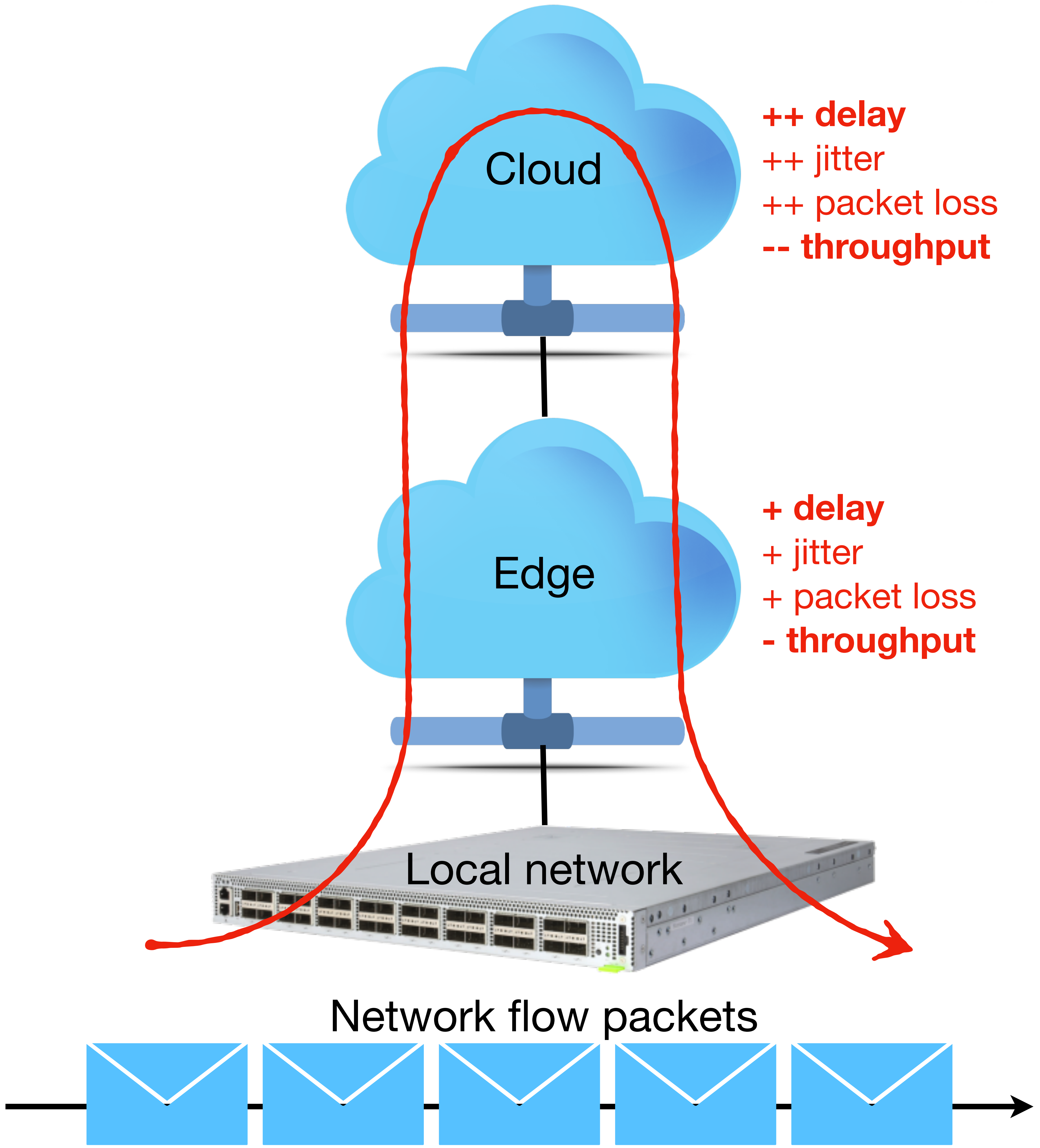
- The ultimate goal of AI/ML for networks is **higher accuracy** and **quick responses**
- Our community has excelled in achieving high-quality models, but we are **falling short** of making it **fast** and **promptly**

Source: RunCzech

Filippo Poltronieri, Cesare Stefanelli, Mauro Tortonesi, and Mattia Zaccarini: **Reinforcement Learning vs. Computational Intelligence: Comparing Service Management Approaches for the Cloud Continuum**. Future Internet 15, no. 11: 359 (2023).
Jonatas Marques, Kirill Levchenko, Luciano Paschoal Gaspary: **IntSight: Diagnosing SLO Violations with In-band Network Telemetry**. CoNEXT 2020: 421-434.

What are we striving for?

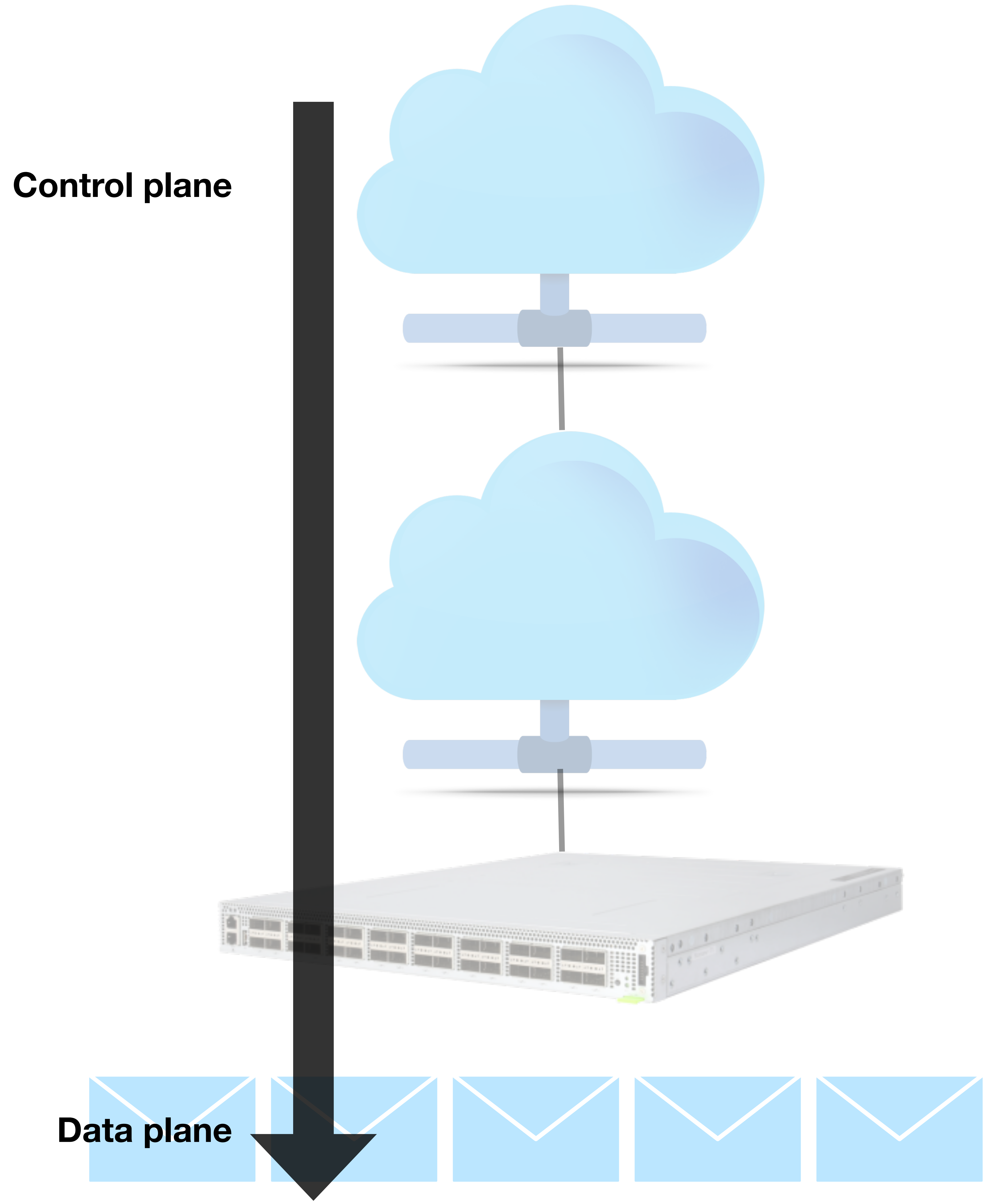
- The ultimate goal of AI/ML for networks is **higher accuracy** and **quick responses**
- Our community has excelled in achieving high-quality models, but we are **falling short** of making it **fast** and **promptly**
- Long control loops, requirements of AI/ML techniques (*e.g.*, training)



Filippo Poltronieri, Cesare Stefanelli, Mauro Tortonesi, and Mattia Zaccarini: **Reinforcement Learning vs. Computational Intelligence: Comparing Service Management Approaches for the Cloud Continuum**. Future Internet 15, no. 11: 359 (2023).
Jonatas Marques, Kirill Levchenko, Luciano Paschoal Gaspary: **IntSight: Diagnosing SLO Violations with In-band Network Telemetry**. CoNEXT 2020: 421-434.

What are we striving for?

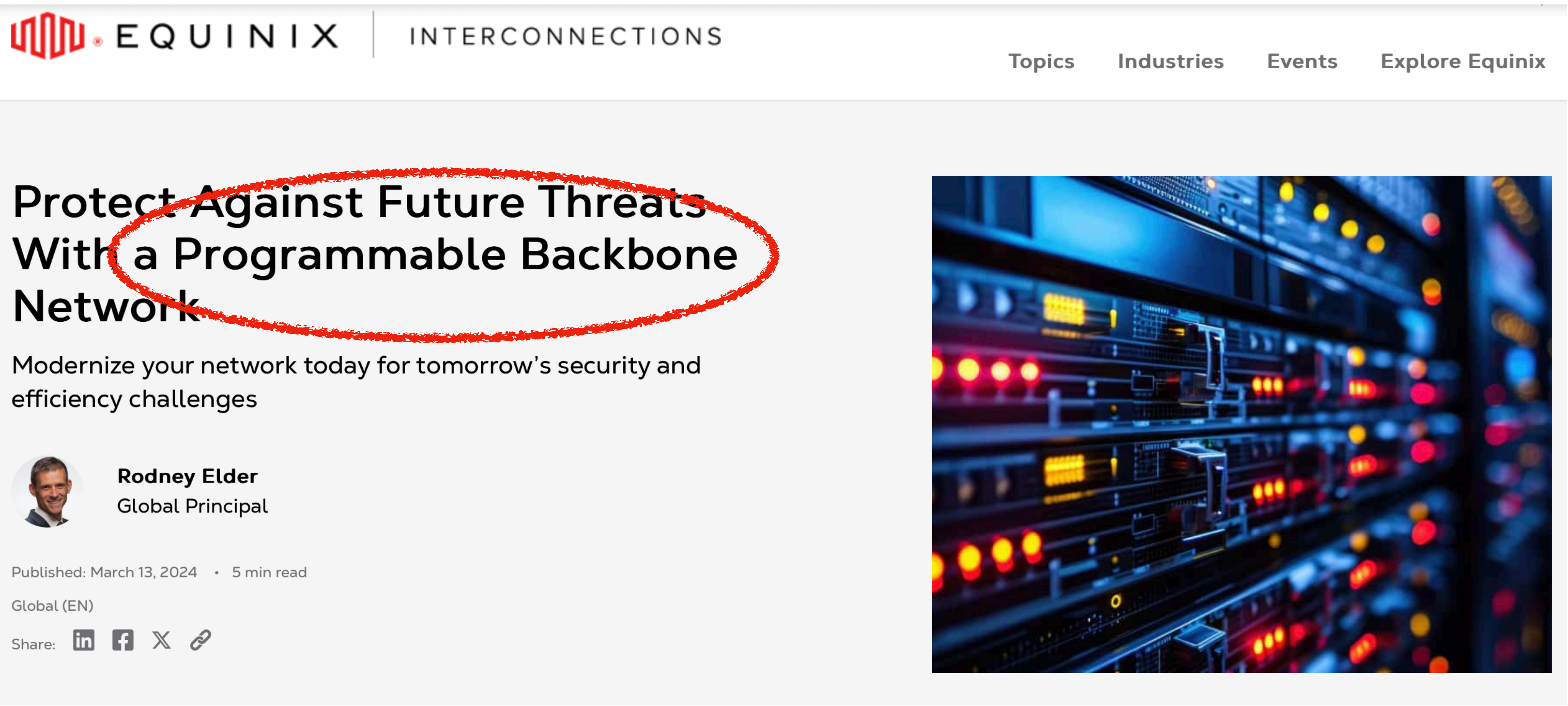
- The ultimate goal of AI/ML for networks is **higher accuracy** and **quick responses**
- Our community has excelled in achieving high-quality models, but we are **falling short** of making it **fast** and **promptly**
- Long control loops, requirements of AI/ML techniques (*e.g.*, training)
- Push needed from control plane-based ML to in-network ML

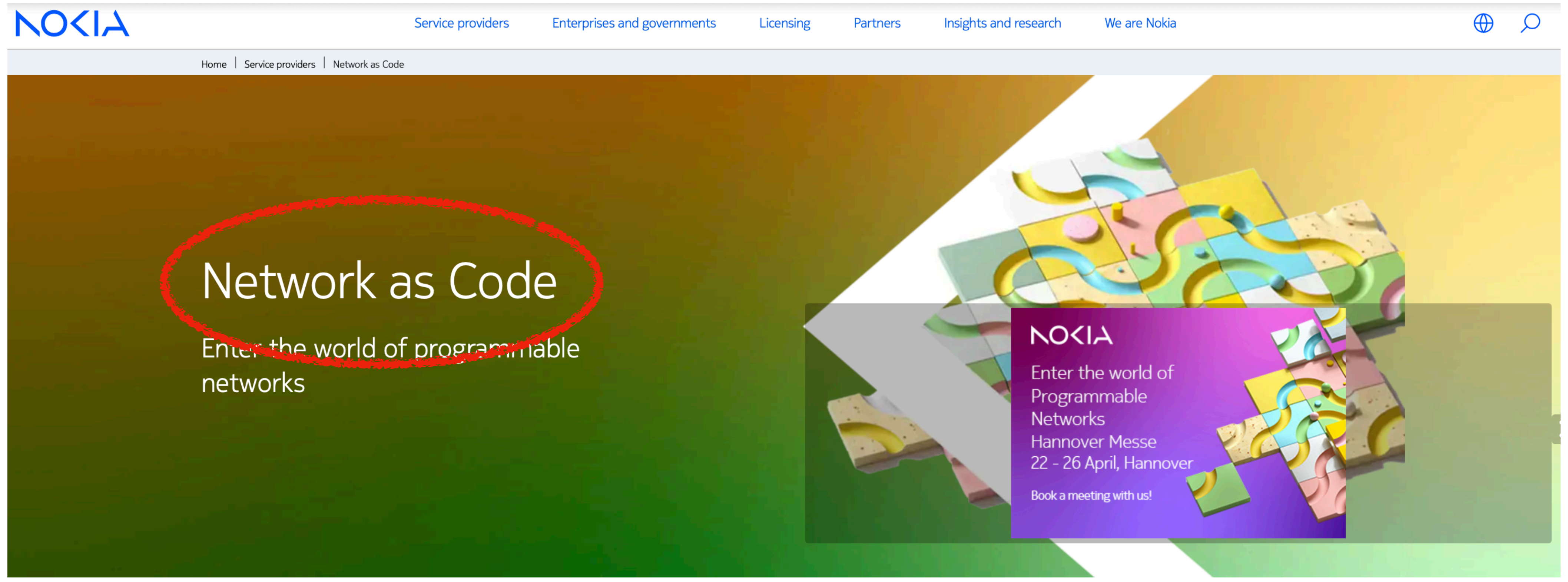


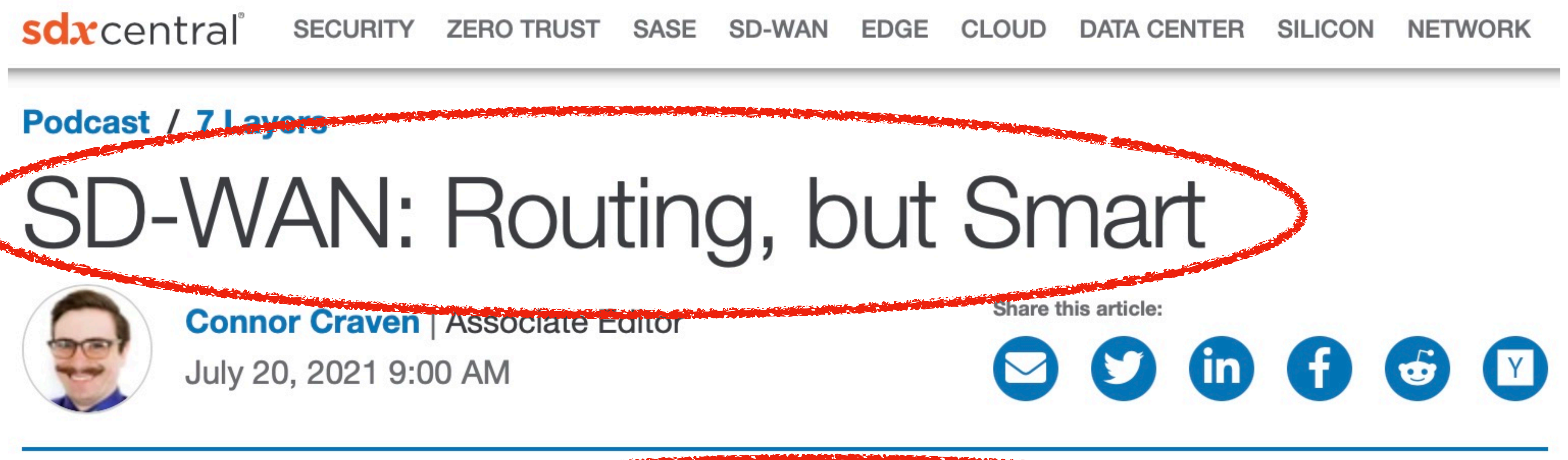
Source: C. Zheng, X. Hong, D. Ding, S. Vargaftik *et al*. (2023)

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

The emergence of programmable networks







Announcing Next-Generation P4-Programmable Datacenter Switching

Oxide Computer Company
8.423 seguidores

14 de outubro de 2024

Oxide is excited to announce a partnership with Xsight Labs to build the next generation of P4-programmable networks on the Oxide Cloud Computer. We'll be building around the just-announced X2 chip from Xsight. At Oxide, partnership is foundational to what we do. Working with the Xsight team has been amazing and was a key component to our decision. We're really looking forward to the advances we'll be making in programmable data-plane networking together. For an inside look at how we've leveraged programmable networks in the Oxide platform, check out our Rack-Scale Networking Oxide and Friends episode. More details to come on an open-source P4 compiler implementation for the X2.

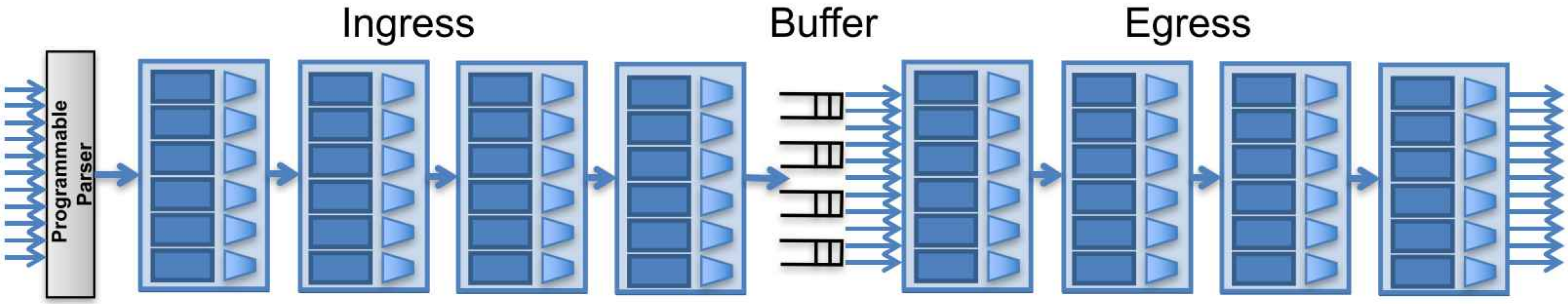
Software-Defined Networks
Programmable Networks
Programmable Data Planes (PDPs)

Weverton Cordeiro, Jonatas Marques, Luciano Gaspary: **Data Plane Programmability Beyond OpenFlow: Opportunities and Challenges for Network and Service Operations and Management**. J. Netw. Syst. Manag. 25(4): 784-818 (2017).

Programmable network logic in hardware

- **Hardware support** to accommodate new software
- **Per-packet** traffic **visibility**
- **Line-rate processing** (sub-microsecond latency)

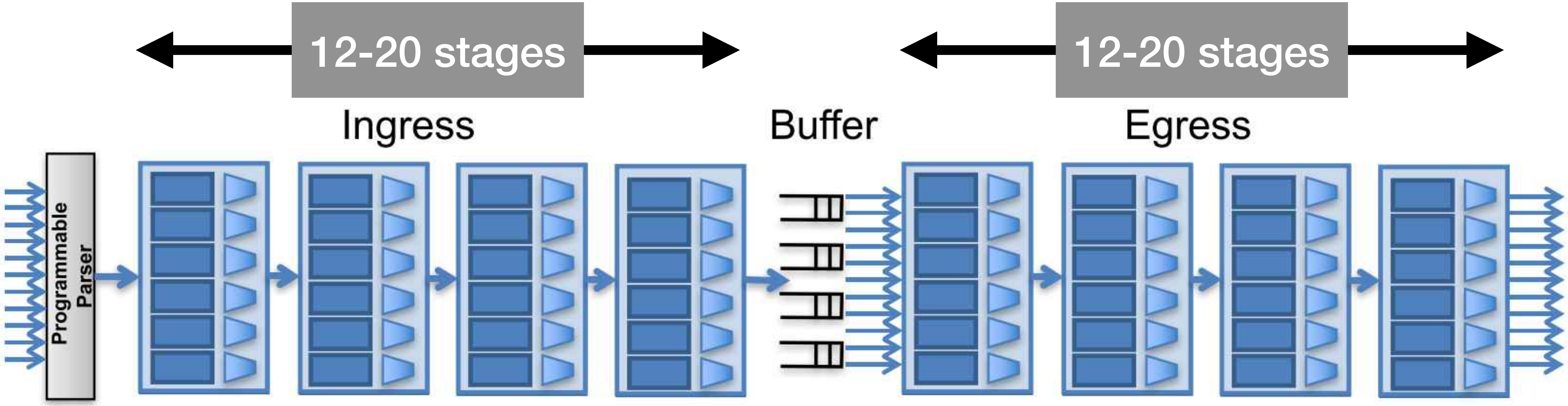
10’s Tbps
aggregated throughput
10’s Billions
packets per second



Source: https://www.infoq.com/presentations/pisa-asic-p4/ *apud* https://p4.org/

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).
Pat Bosshart, Glen Gibb, Hun-Seok Kim, George Varghese, Nick McKeown, Martin Izzard, Fernando A. Mujica, Mark Horowitz: **Forwarding Metamorphosis: Fast Programmable Match-action Processing in Hardware for SDN**. SIGCOMM 2013: 99-110.

Programmable network logic in hardware

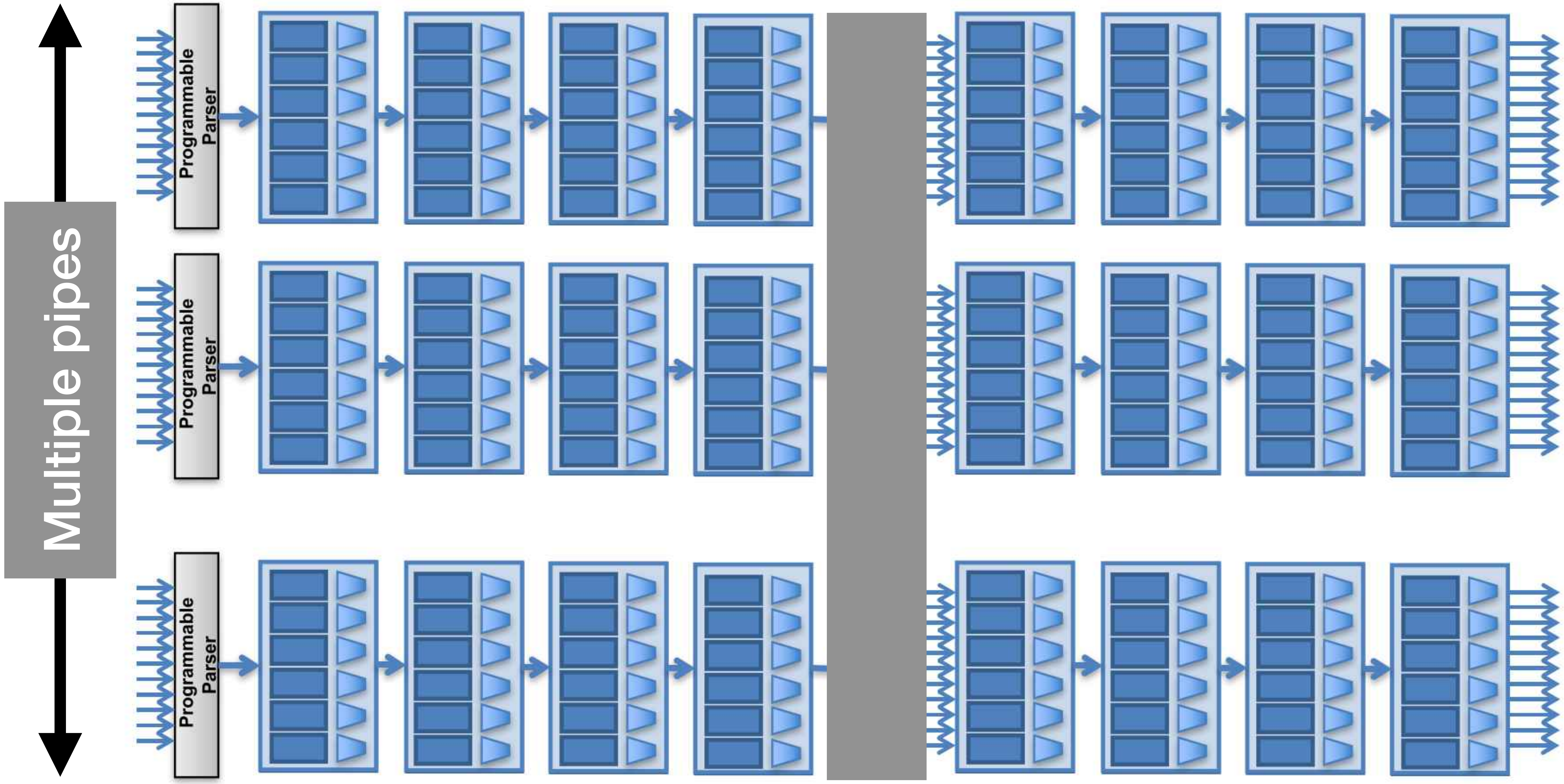


Source: https://www.infoq.com/presentations/pisa-asic-p4/ *apud* https://p4.org/

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

Pat Bosshart, Glen Gibb, Hun-Seok Kim, George Varghese, Nick McKeown, Martin Izzard, Fernando A. Mujica, Mark Horowitz: **Forwarding Metamorphosis: Fast Programmable Match-action Processing in Hardware for SDN**. SIGCOMM 2013: 99-110.

Programmable network logic in hardware



Source: https://www.infoq.com/presentations/pisa-asic-p4/ *apud* https://p4.org/

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).
Pat Bosshart, Glen Gibb, Hun-Seok Kim, George Varghese, Nick McKeown, Martin Izzard, Fernando A. Mujica, Mark Horowitz: **Forwarding Metamorphosis: Fast Programmable Match-action Processing in Hardware for SDN**. SIGCOMM 2013: 99-110.

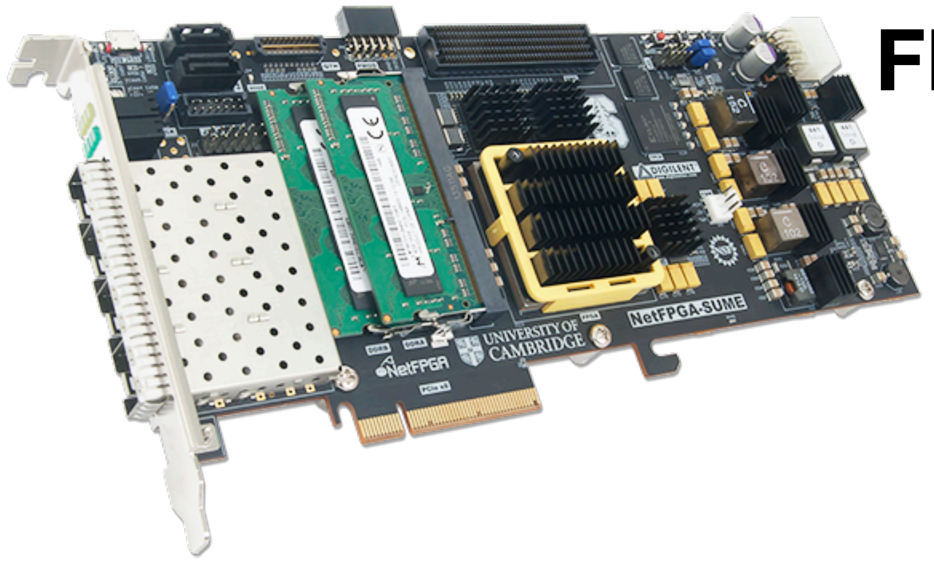
Programmable network devices

A couple of examples

- **Switches**:
  - Edgecore Wedge 100BF-32 Tofino, 32 ports of 100 GbE (6.4 Tbps)
  - **X-Switch**, **128 ports** of **100 GbE** (**12.8 Tbps**)
- **Smart NICs**:
  - NetFPGA-SUME Virtex-7 FPGA Board
  - Xilinx Alveo SN1022 100 GbE
  - NVIDIA BlueField-2 E-Series DPU 100 GbE

ASIC

FPGA

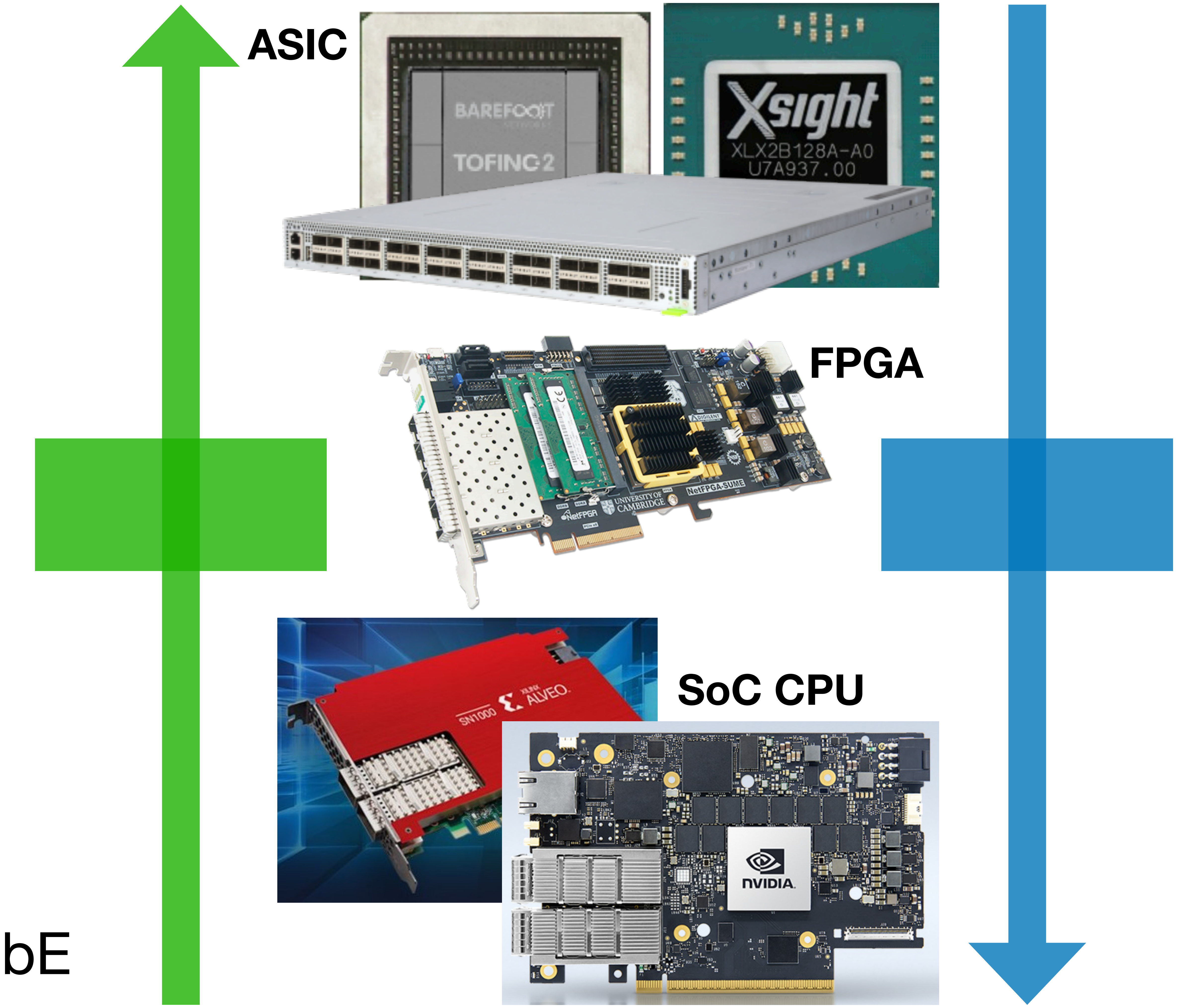


SoC CPU

Programmable network devices

A couple of examples

- **Switches**:
  - Edgecore Wedge 100BF-32 Tofino, 32 ports of 100 GbE (6.4 Tbps)
  - **X-Switch**, **128 ports** of **100 GbE** **(12.8 Tbps)**
- **Smart NICs**:
  - NetFPGA-SUME Virtex-7 FPGA Board
  - Xilinx Alveo SN1022 100 GbE
  - NVIDIA BlueField-2 E-Series DPU 100 GbE



Domain-specific languages for networks

- Domain-specific languages/instructions
- Limited procedural code, *i.e.*, abstraction of a sequence of actions
- Commands with scope isolated to handling packets
- We are **trading expressiveness** and **flexibility** for **safety**, **security**, and **efficiency**
- Examples include: P4, NPL, POF (+ micro-C, C++)

```
parser my_parser() {}

control my_deparser() {}

control my_verify_checksum() {}

control my_compute_checksum() {}

control my_ingress() {}

control my_egress() {}

V1Switch(my_parser(),
         my_verify_checksum(),
         my_ingress(),
         my_egress(),
         my_compute_checksum(),
         my_deparser()) main;
```

Weverton Cordeiro, Jonatas Marques, Luciano Gaspary: **Data Plane Programmability Beyond OpenFlow: Opportunities and Challenges for Network and Service Operations and Management**. J. Netw. Syst. Manag. 25(4): 784-818 (2017).

A new ecosystem of enabling technologies

Software development suites	Intel P4 Studio, Intel P4 Insight, DOCA, Xilinx Vivado, Xilinx SDNet, X-SDK
Domain-specific languages and APIs	P4, NPL, POF, P4Runtime, C++, gRPC, micro-C, …
Programmable ASICs and forwarding devices	Xsight X2, Intel Tofino 2, Broadcom Trident 4, Cavium XPliant, Innovium Teralynx, NetFPGA SUME, …

ML using programmable network devices

The scope of research in the area: phases

- **Training**: offload operations for efficient distributed training
  - *E.g.*, aggregation functions and parameter optimization
  - The typical case of "**networks for AI/ML**"

Ricardo Parizotto, Bruno Loureiro Coelho, Diego Cardoso Nunes, Israat Haque, and Alberto Schaeffer-Filho: **Offloading Machine Learning to Programmable Data Planes: A Systematic Survey**. ACM Comput. Surv. 56, 1, Article 18 (2024).

The scope of research in the area: phases

- **Training**: offload operations for efficient distributed training
  - *E.g.*, aggregation functions and parameter optimization
  - The typical case of "**networks for AI/ML**"
- **Inference**: deploy in-network ML models to network devices
- The remaining of this talk focuses on inference

Ricardo Parizotto, Bruno Loureiro Coelho, Diego Cardoso Nunes, Israat Haque, and Alberto Schaeffer-Filho: **Offloading Machine Learning to Programmable Data Planes: A Systematic Survey**. ACM Comput. Surv. 56, 1, Article 18 (2024).

What has been developed so far?

Source: C. Zheng, X. Hong, D. Ding, S. Vargaftik *et al*. (2023)

Time frame of five years only

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

What has been developed so far?

Source: C. Zheng, X. Hong, D. Ding, S. Vargaftik *et al*. (2023)

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

What has been developed so far?

Source: C. Zheng, X. Hong, D. Ding, S. Vargaftik *et al*. (2023)

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

What has been developed so far?

SVM, Neural Network, Decision Tree, Ensemble Tree models, Naïve Bayes, K-Nearest Neighbor, K-means, Isolation Forest, PCA, Autoencoder, ~~Self-Organizing Map~~, Q-learning, SARSA

Source: C. Zheng, X. Hong, D. Ding, S. Vargaftik *et al*. (2023)

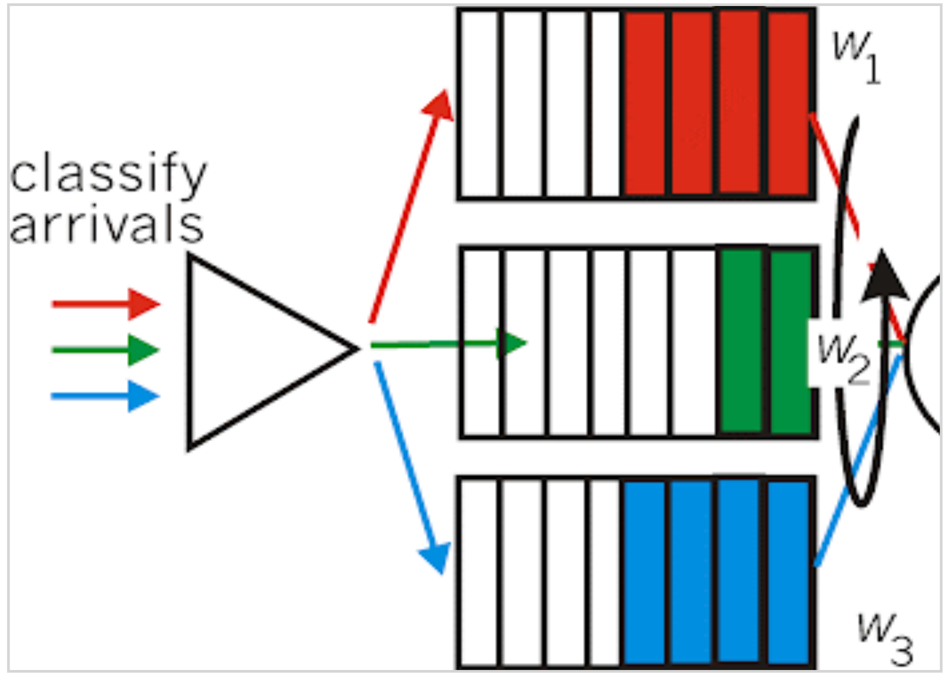
C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

In-network inference: apps and ML techniques

Intrusion detection

Source: https://towardsdatascience.com

Traffic classification

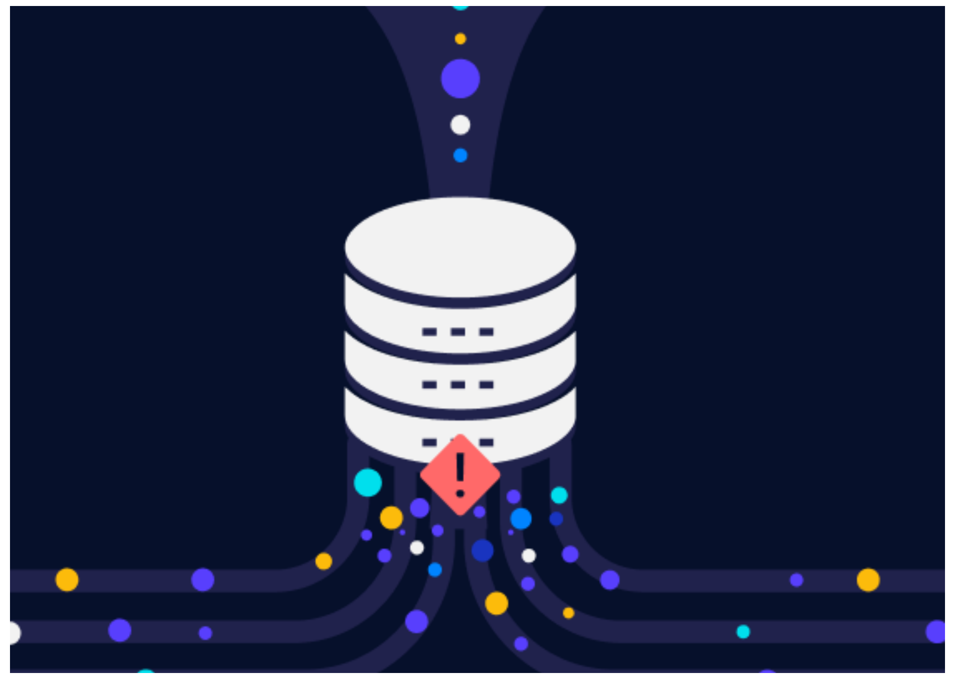


Source: Kurose and Ross, 2020.

Traffic engineering

Source: https://engineering.nyu.edu

Congestion control



Source: https://granulate.io

Video streaming

Source: https://www.forbes.com

Intrusion detection: AE, BNN, Clustering, DNN, DT, ET, KM, NB, KNN, NN, PCA, RL, SVM

Traffic classification: AE, BNN, DNN, DT, ET, KM, KNN, LSTM, NB, PCA, RF, SVM, XGB

+ **resource management** + **fault management** + **channel modeling** + ...

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

Ricardo Parizotto, Bruno Loureiro Coelho, Diego Cardoso Nunes, Israat Haque, and Alberto Schaeffer-Filho: **Offloading Machine Learning to Programmable Data Planes: A Systematic Survey**. ACM Comput. Surv. 56, 1, Article 18 (2024).

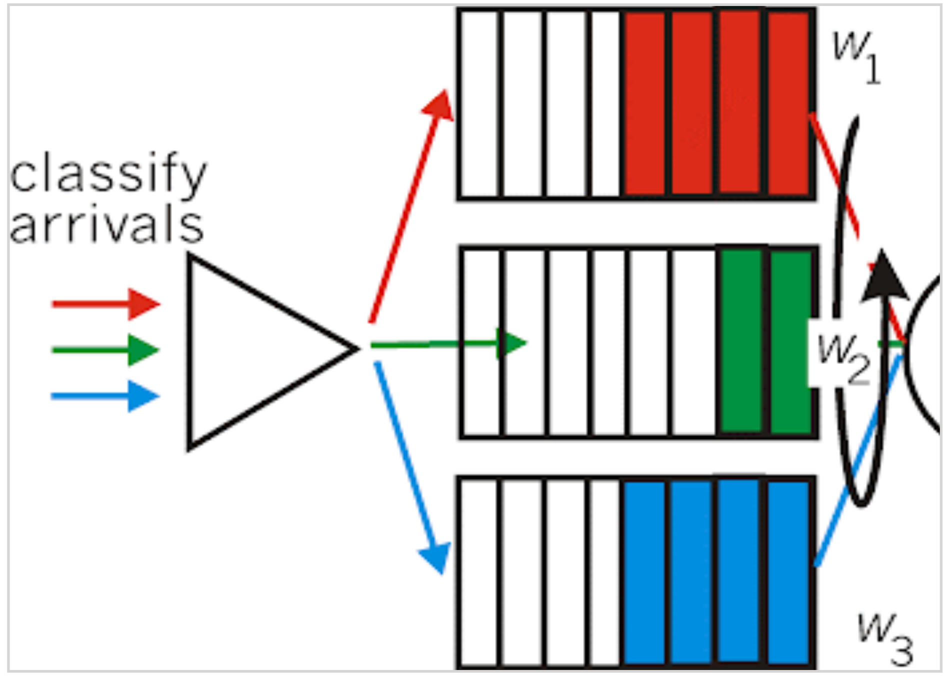
In-network inference: apps and ML techniques

Intrusion detection

Source: https://towardsdatascience.com

AE, BNN, Clustering, DNN, DT, ET, KM, NB, KNN, NN, PCA, RL, SVM

Traffic classification



Source: Kurose and Ross, 2020.

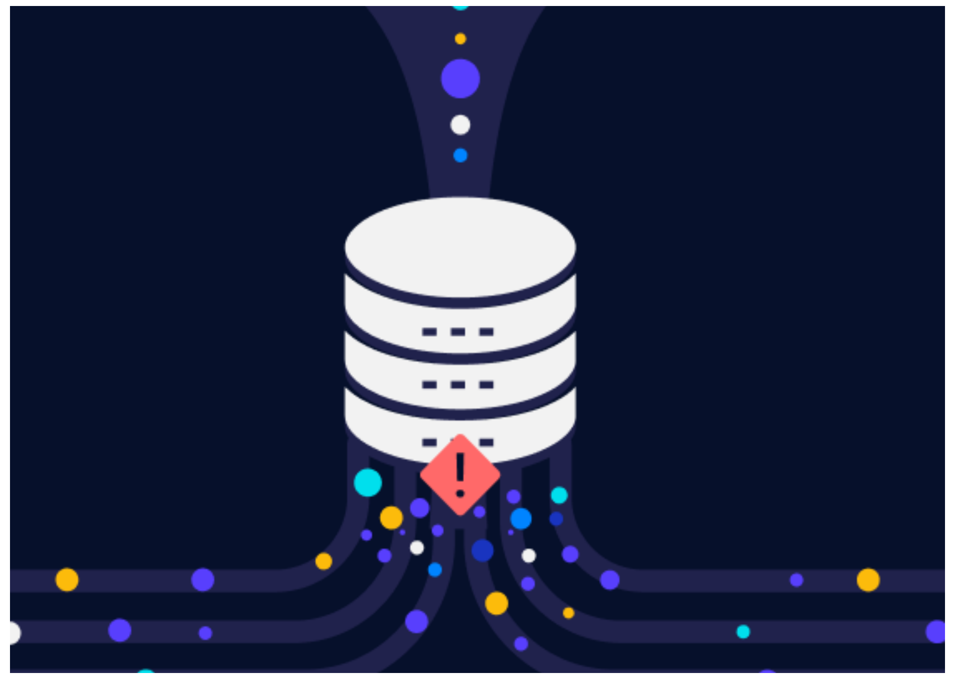
AE, BNN, DNN, DT, ET, KM, KNN, LSTM, NB, PCA, RF, SVM, XGB

Traffic engineering

Source: https://engineering.nyu.edu

RL

Congestion control



Source: https://granulate.io

LSTM

Video streaming

Source: https://www.forbes.com

\+ **resource management** + **fault management** + **channel modeling** + ...

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).
Ricardo Parizotto, Bruno Loureiro Coelho, Diego Cardoso Nunes, Israat Haque, and Alberto Schaeffer-Filho: **Offloading Machine Learning to Programmable Data Planes: A Systematic Survey**. ACM Comput. Surv. 56, 1, Article 18 (2024).

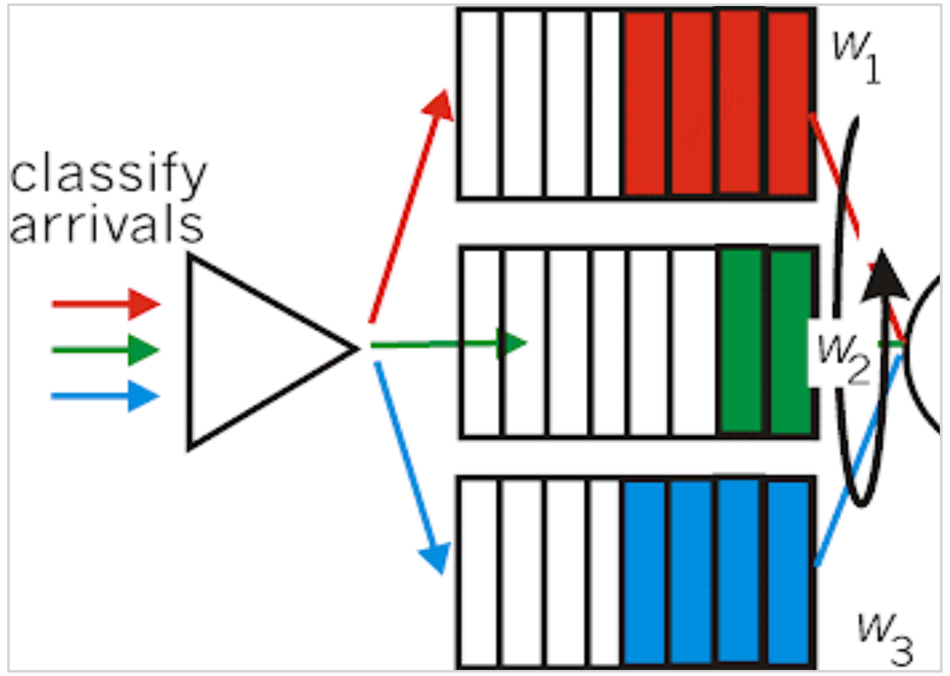
In-network inference: apps and ML techniques

Intrusion detection

Source: https://towardsdatascience.com

AE, BNN, Clustering, DNN, DT, ET, KM, NB, KNN, NN, PCA, RL, SVM

Traffic classification



Source: Kurose and Ross, 2020.

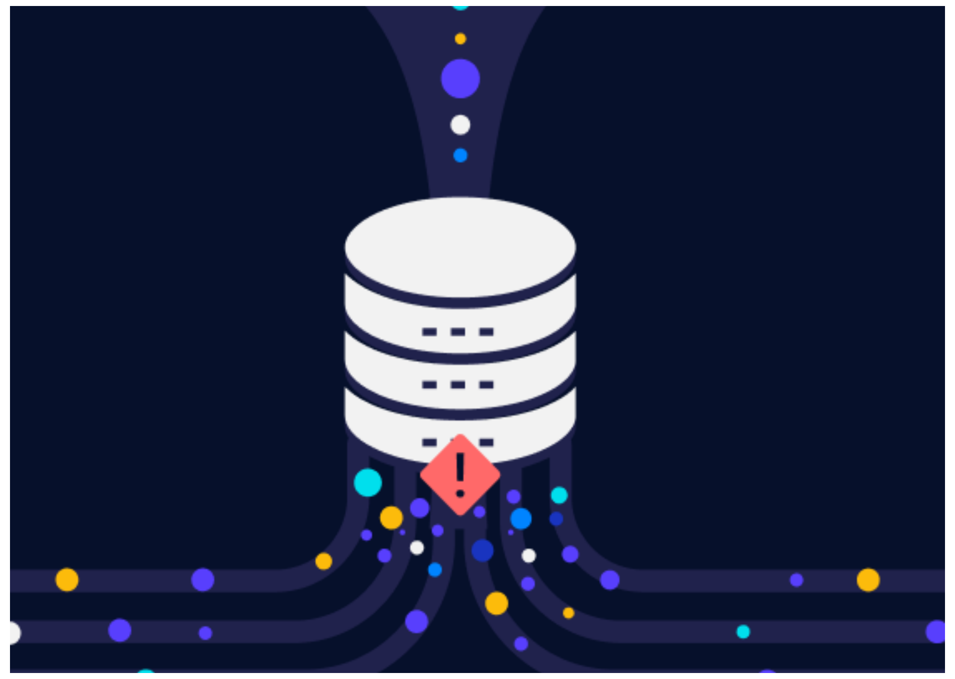
AE, BNN, DNN, DT, ET, KM, KNN, LSTM, NB, PCA, RF, SVM, XGB

Traffic engineering

Source: https://engineering.nyu.edu

RL

Congestion control



Source: https://granulate.io

LSTM

Video streaming

Source: https://www.forbes.com

Opportunity?

\+ **resource management** — Opportunity?

\+ **fault management** — Opportunity?

\+ **channel modeling** — Opportunity?

\+ ...

C. Zheng, X. Hong, D. Ding, S. Vargaftik, Y. Ben-Itzhak, N. Zilberman: **In-Network Machine Learning Using Programmable Network Devices: A Survey**. IEEE Communications Surveys & Tutorials (2023).

Ricardo Parizotto, Bruno Loureiro Coelho, Diego Cardoso Nunes, Israat Haque, and Alberto Schaeffer-Filho: **Offloading Machine Learning to Programmable Data Planes: A Systematic Survey**. ACM Comput. Surv. 56, 1, Article 18 (2024).

Are we there yet? What’s next?

New promising network technologies and low-latency, high-throughput networked applications

Source: https://medium.com/swlh/lets-explore-a-new-reality-the-future-of-vr-ar-9f73ed38364d

Source: https://www.dashtech.org

Source: https://olhardigital.com.br

4th International Workshop on High-Precision, Predictable, and Low-Latency Networking (HiPNet 2022). http://www.cnsm-conf.org/2022/workshop_HiPNet.html.

Alexander Clemm, Maria Torres Vega, Hemanth Kumar Ravuri, Tim Wauters, Filip De Turck. Toward **Truly Immersive Holographic-Type Communication: Challenges and Solutions**. IEEE Commun. Mag. 58(1): 93-99 (2020).

Many, many opportunities

Adequate abstractions for in-network ML

Adequate abstractions for in-network ML

Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML

New compelling applications and use cases

Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML

Support for runtime programmability

New compelling applications and use cases

Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML	Support for runtime programmability	
	Handling of encrypted traffic	
	New compelling applications and use cases	Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML

Support for runtime programmability

Handling of encrypted traffic

Convenient compilation & deployment systems

New compelling applications and use cases

Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML

Support for runtime programmability

Reduction of resource and communication overhead

Handling of encrypted traffic

Convenient compilation & deployment systems

New compelling applications and use cases

Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML

Support for runtime programmability

Reduction of resource and communication overhead

From *ad-hoc* constructs to reusable management libraries

Handling of encrypted traffic

Convenient compilation & deployment systems

New compelling applications and use cases

Implementation and deployment of large-scale ML models

Adequate abstractions for in-network ML

Support for runtime programmability

Reduction of resource and communication overhead

From *ad-hoc* constructs to reusable management libraries

Handling of encrypted traffic

Balancing short (data plane) and long (control plane) control loops (inference *x* training)

Convenient compilation & deployment systems

New compelling applications and use cases

Implementation and deployment of large-scale ML models

Final remarks

- In-network AI/ML can potentially allow for the execution of some networking tasks a lot better (*e.g.*, detection, prediction, etc.)
- And, very notably, with **per-packet visibility** and unprecedented **low latency**
- There are plenty of research opportunities to be explored!
- We can and will deliver breakthrough innovations based on in-network AI/ML to network and service management
- It is not a matter of “if” but “when”

Source: ChatGPT

Thank you!

Picture taken by Prof. Mauro Tortonesi during NOMS 2023 (held in Miami)

Luciano Paschoal Gaspary - UFRGS, Brazil

paschoal@inf.ufrgs.br • http://www.inf.ufrgs.br/~paschoal • @lpgaspary

WTR 2025 • 22 October 2025 • Porto Alegre