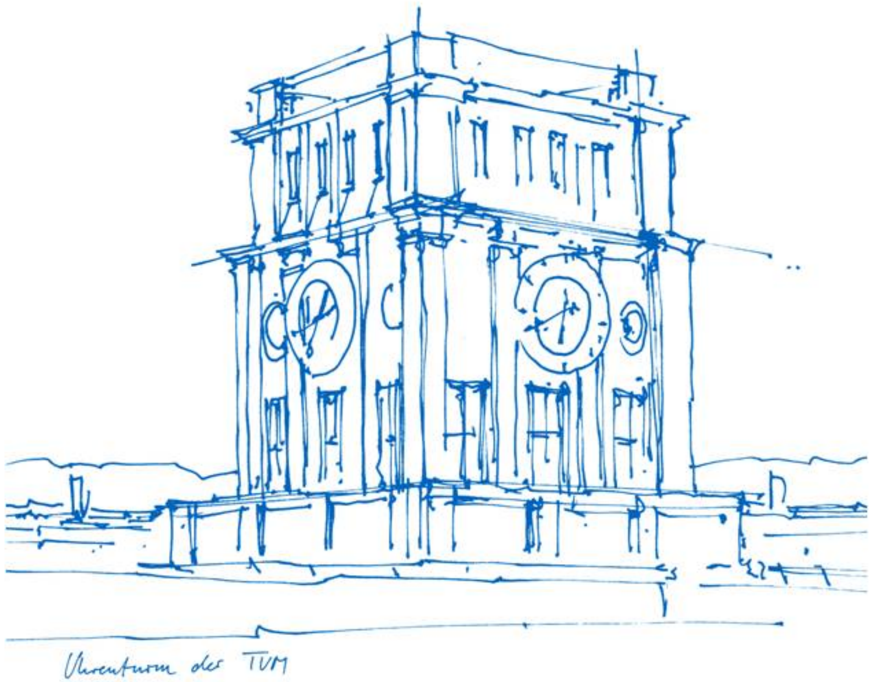
Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich

Reproducible Research for Networked Systems

Georg Carle
Sebastian Gallenmüller

{carle|gallenmu}@net.in.tum.de
http://www.net.in.tum.de/{~carle|~gallenmu}

Acknowledgements:
All members of the Chair of
Network Architectures and Services



Outline

Needs

- Scalable, Resilient and Trustworthy Programmable Networked Systems with Predictable Performance
- Research Infrastructure for Reproducible Experiments

Challenges

Approach

- Framework, Methods and Tools for Reproducible Experiments
- Scientific Large-scale Infrastructure for Computing/Communication Experimental Studies

Conclusions

Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich

Scalable, Resilient and Trustworthy Programmable Networked Systems

Need for Resilient Low-Latency Predictable Network Services

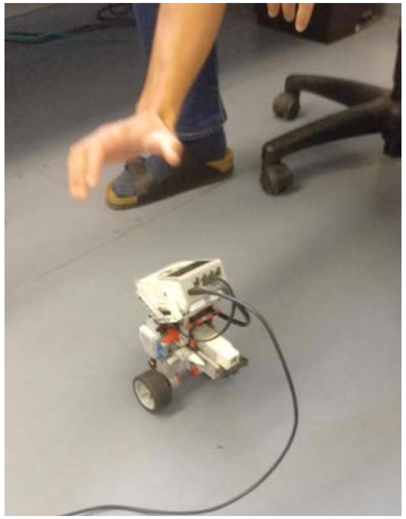
Challenges

- complex architectures
- performance, safety and security requirements

⇨ Need for

- Secure communication, trustworthy implementation
- Network stack + applications: *worst case performance guarantees*
- Scalability, flexibility, affordability, time-to-market

Low-Latency Systems:



Network-Controlled Robot

Power Grid Control

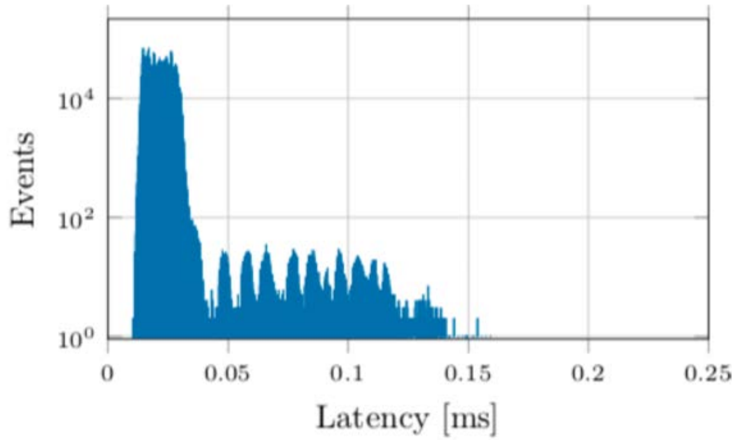
Need: End-to-End Worst-Case Latency Guarantees

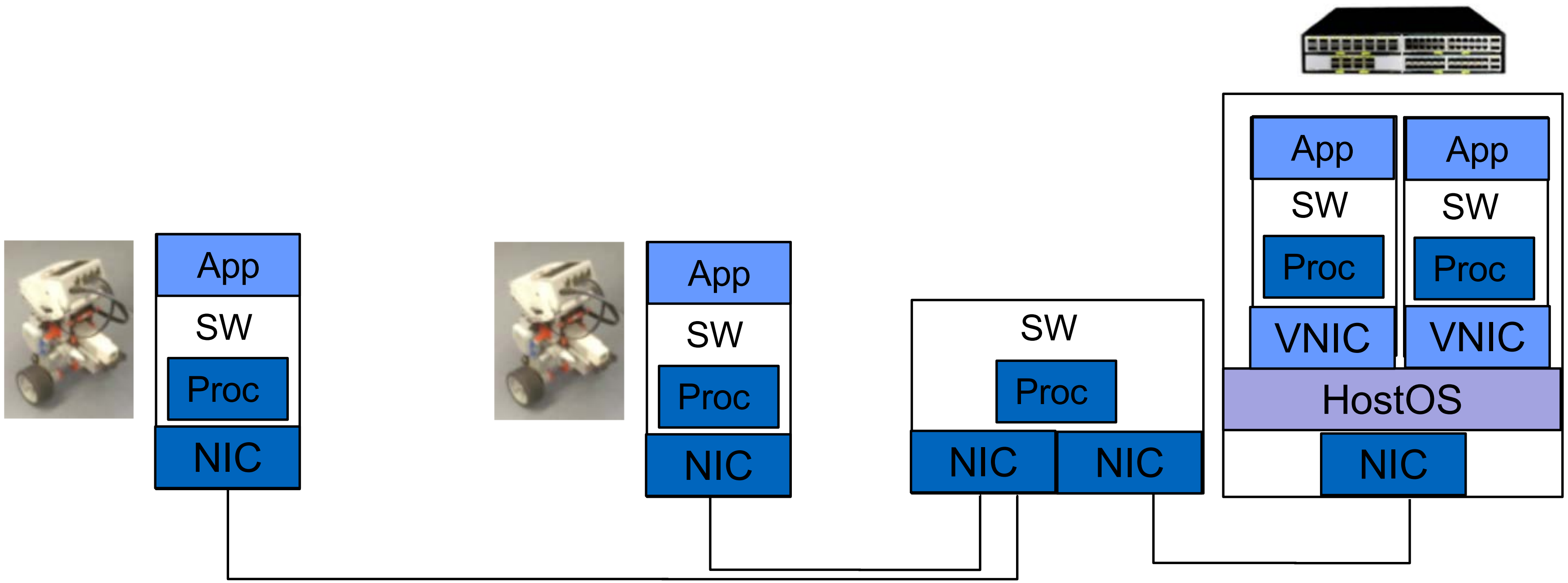
Goal:

- Predictable performance of networked systems

Challenges:

- Complex Hardware + Software
- Programmability
- Issue: latency distribution (long tail)



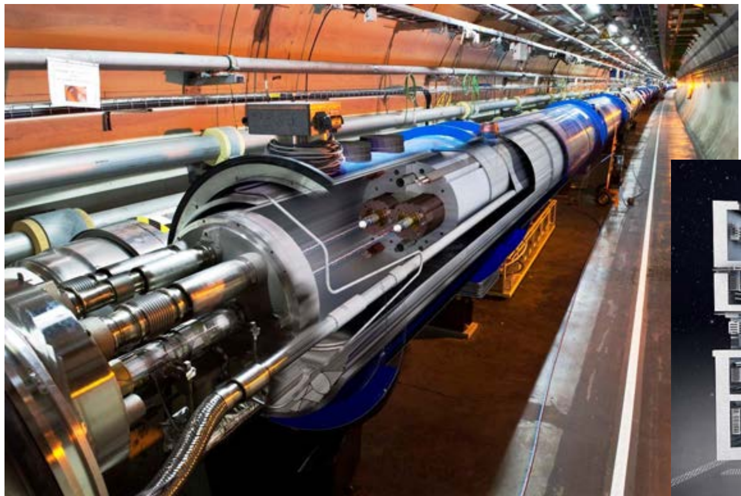


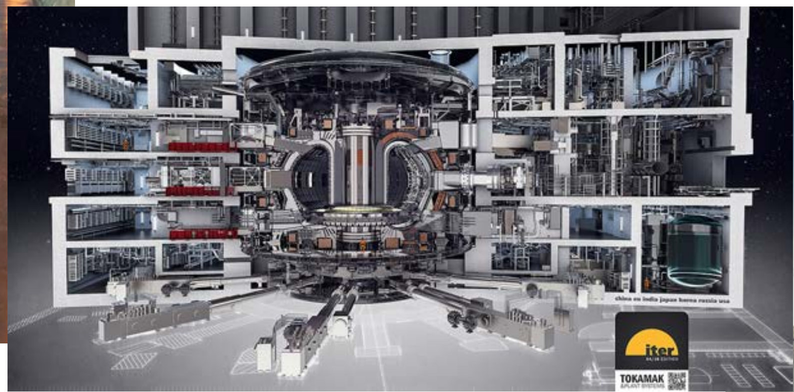
Goal:

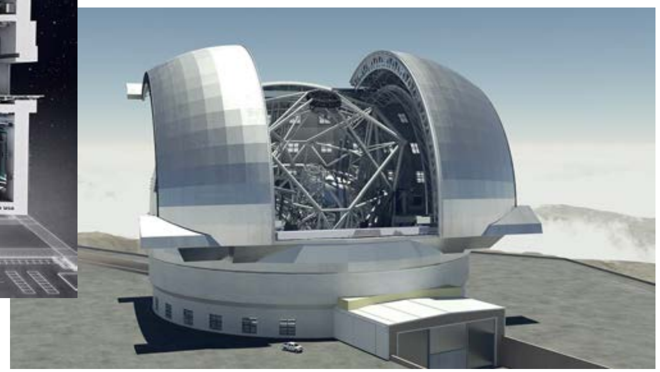
Research Infrastructure for Networked Systems

Natural Sciences Research infrastructures

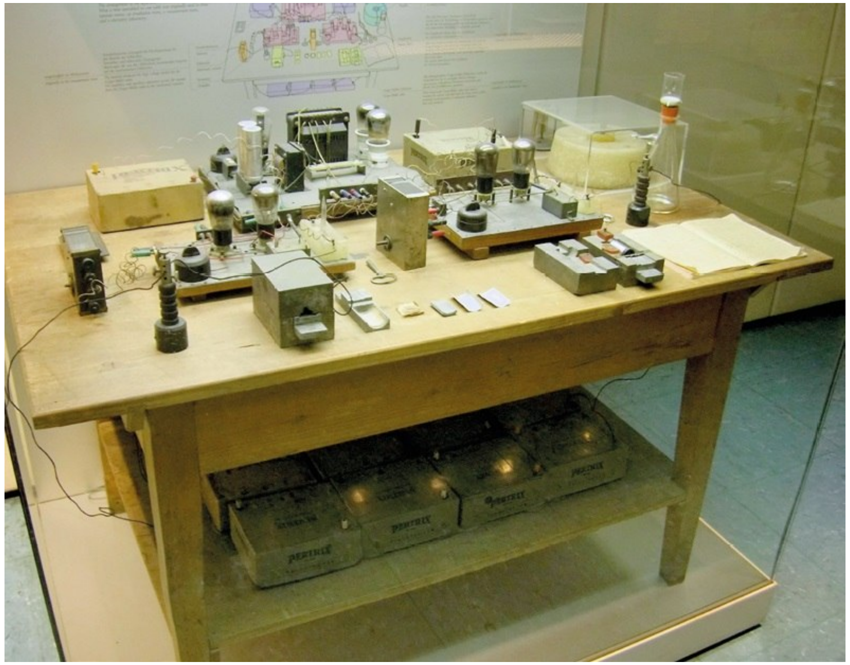
- Large-scale research infrastructures have become a necessity to answer current research questions
- Long-term funding programs allow the creation of infrastructures
  - Large Hadron Collider
  - Fusion Reactor ITER
  - Extremely Large Telescope
- For Computer Science research no such infrastructures exists







Research Infrastructure for Networked Systems



First nuclear fission experiment
(Otto Hahn, Germany 1938)

Networked systems
Reproducible experiments?

Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich

Challenge: Complexity

Complexity of Protocol Stack

Complexity by Programmability

Complexity by Processing Architecture

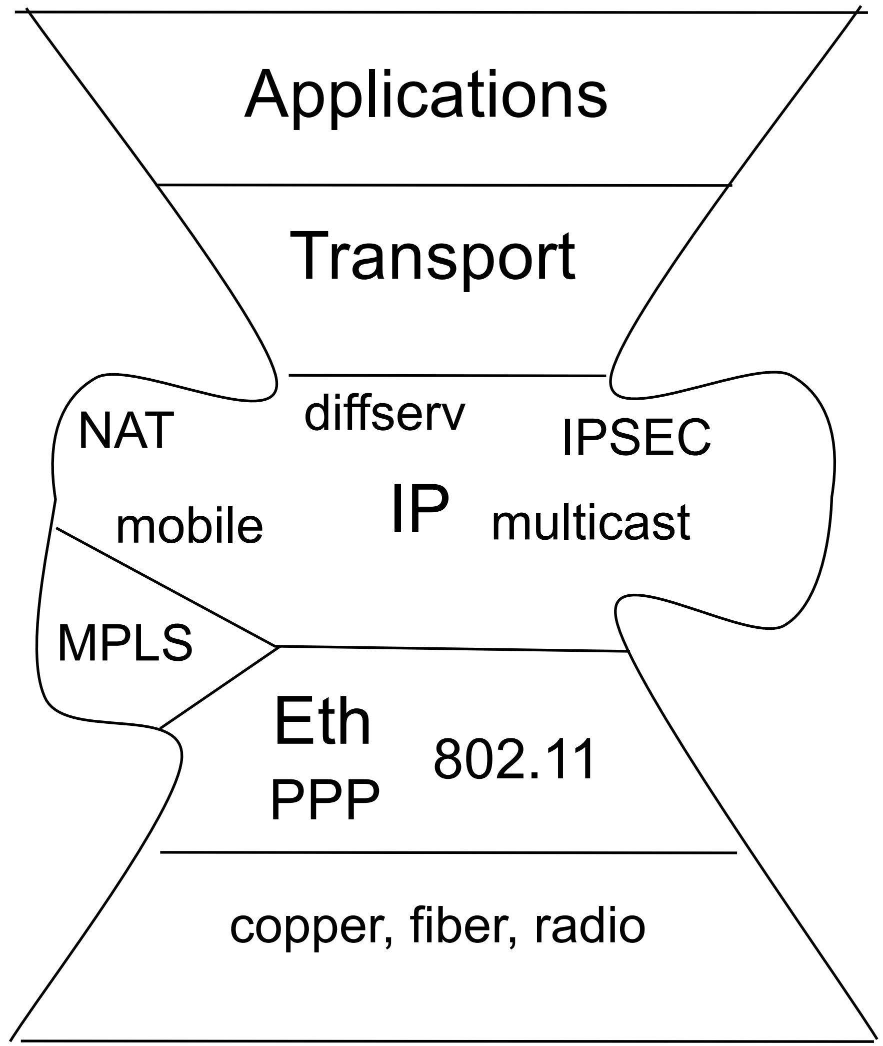
Complexity by Software Architecture

Latency Guarantees

Reproducible Experiments

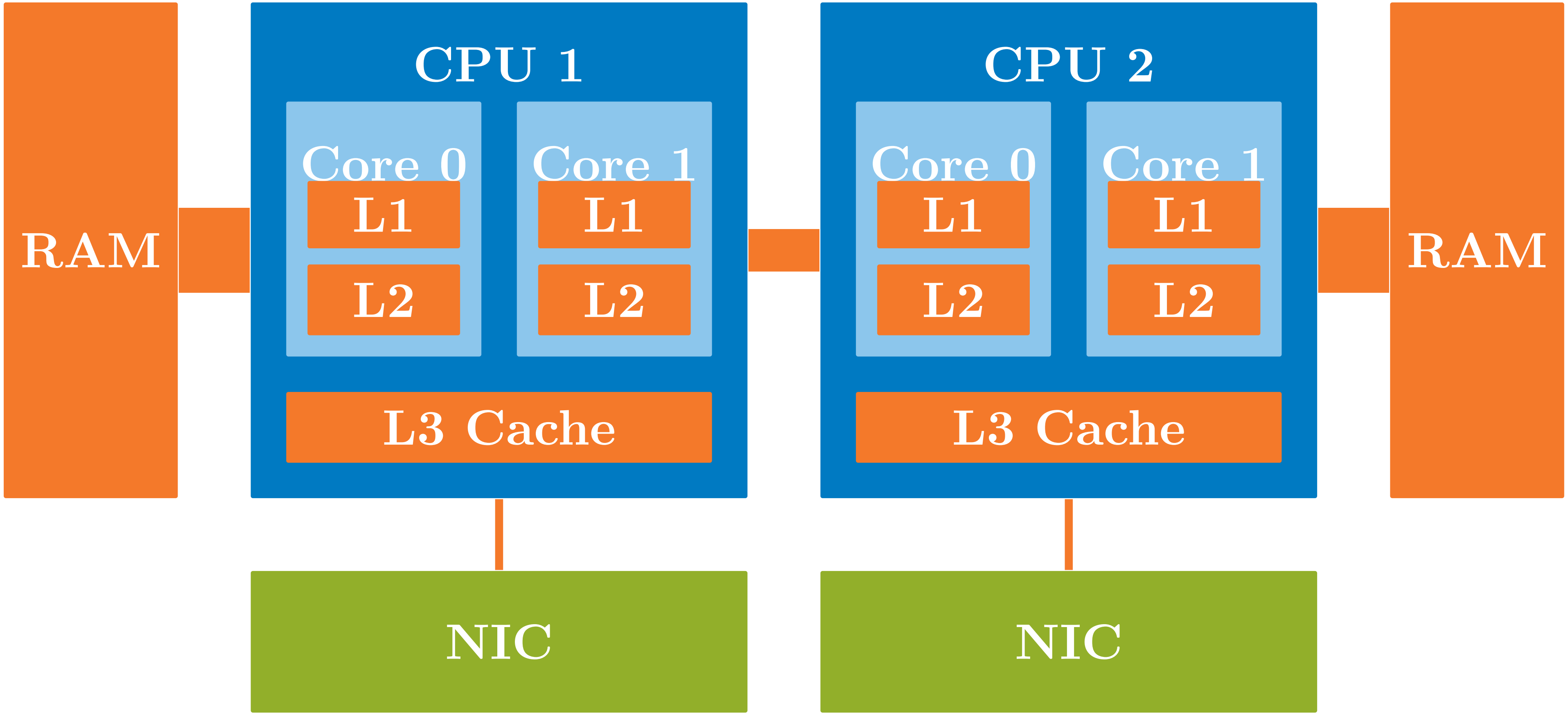
Protocol Stacks are Complex

- TLS, QUIC, MASQUE
- TCP, UDP
- BGP, OSPF,
  VRRP, PIM
- IPsec, IKE, EAP
- IPv4, IPv6, Segment Routing
- VLAN, GTP, IP in IP, GRE, MPLS

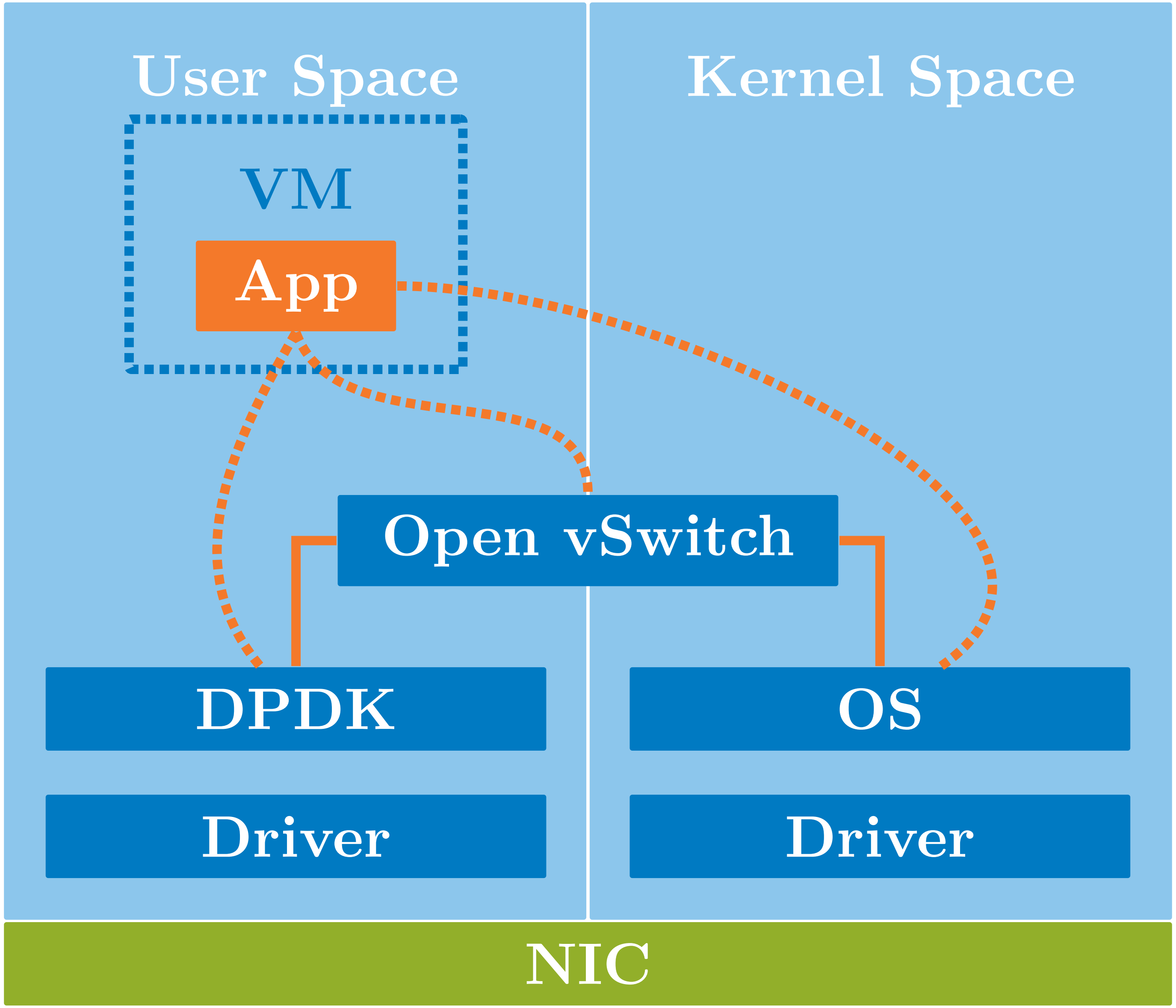


Modern Hardware Architectures are Complex

Non-Uniform Memory Architecture (NUMA)



Modern Software Architectures are Complex

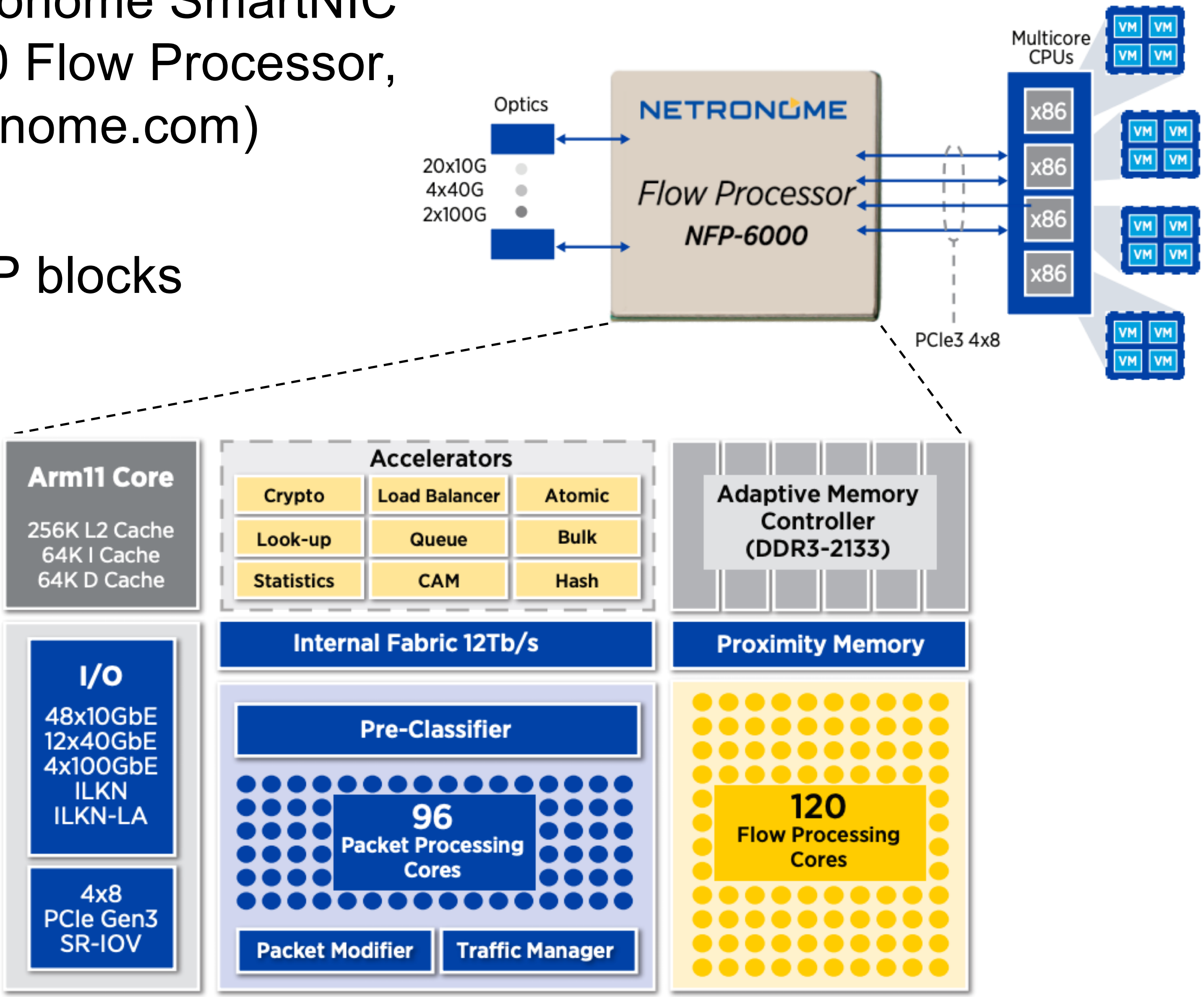


Programmable NICs add Complexity

Programmable packet processing architectures
Example: Netronome SmartNIC
with NFP-6000 Flow Processor,
(cf. www.netronome.com)

NICs

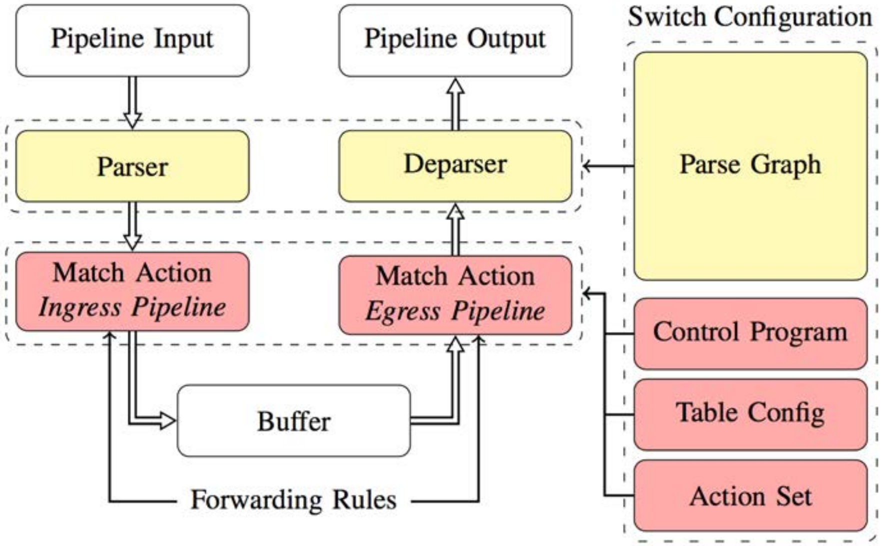
Composable IP blocks



P4 Programmable Packet Processing adds Complexity

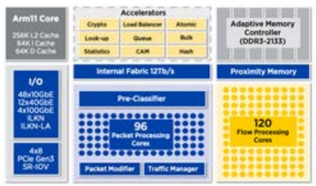
P4 Architecture

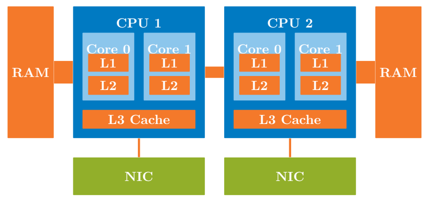
Programmable High-Performance Packet Processing



P4 on different processing targets

- Tofino ASIC-based switch
- P4NetFPGA
- P4 Programming of SmartNIC
- P4 Programming of CPUs (t4p4s DPDK)





P4 Programmable Network Devices

Comparison of P4 Programmable Target Types

	CPU	NPU	FPGA	ASIC
Throughput	+	++	+++	++++
Latency	> 10 µs	5 µs to 10 µs	< 2 µs	< 2 µs
Jitter	−−−−	−−−	−−	−
Resources	++++	+++	++	+
Flexibility	++++	+++	++	+
Example	t4p4s DPDK	NFP-4000 SmartNIC	NetFPGA SUME	Intel Tofino




[ITC2020] Dominik Scholz, Henning Stubbe, Sebastian Gallenmüller, Georg Carle, "Key Properties of Programmable Data Plane Targets," in 32nd International Teletraffic Congress (ITC 32), Osaka, Japan, Sep. 2020

Digital Sovereignty Contribution: High-performance low-latency systems
Programmable with P4, realized using multiple target types, from different vendors

Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich

Reproducible Experiments

Viewpoints on Reproducible Research

ACM SIGCOMM MoMeTools - Workshop on Models, Methods and Tools for Reproducible Network Research
Georg Carle, Hartmut Ritter, Klaus Wehrle,
Karlsruhe, Germany, August 2003

ACM SIGCOMM Reproducibility Workshop
Olivier Bonaventure, Luigi Iannone, Damien Saucez
Los Angeles, USA, August 2017

[Rep17] Q. Scheitle, M. Wählisch, O. Gasser, T. Schmidt, G. Carle,
Towards an ecosystem for reproducible research in computer networking
Proceedings of the ACM SIGCOMM Reproducibility Workshop, 2017

Dagstuhl seminar 18412 “Encouraging Reproducibility in Scientific Research of the Internet“, October 2018

Despite 20 years since first workshop have passed, issues remain

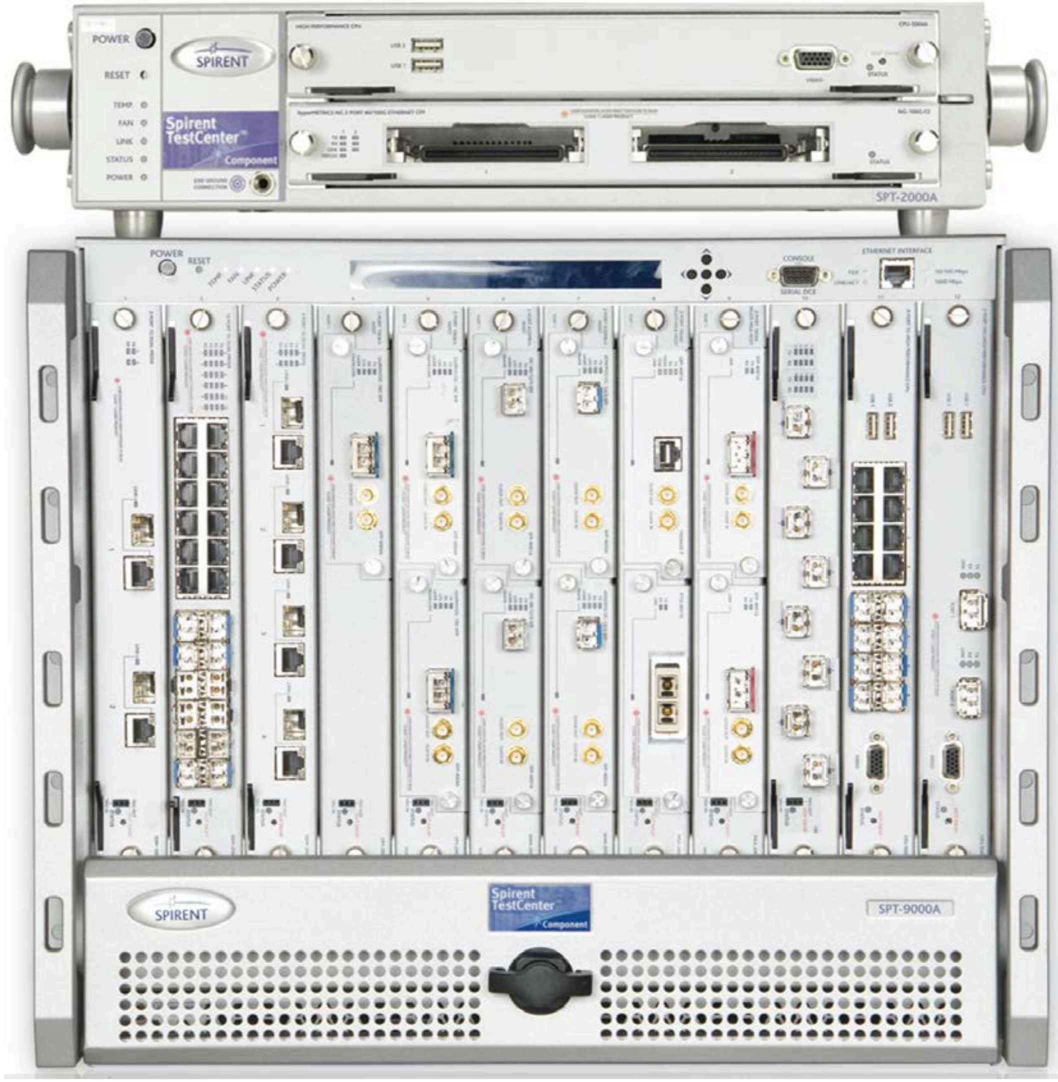
- Which KPIs are relevant?
- How to measure these KPIs?
- How to build **testbeds** to measure these KPIs?
- How to measure in a ***reproducible*** manner?

Hardware Traffic Generators

- Fast
- Precise

but

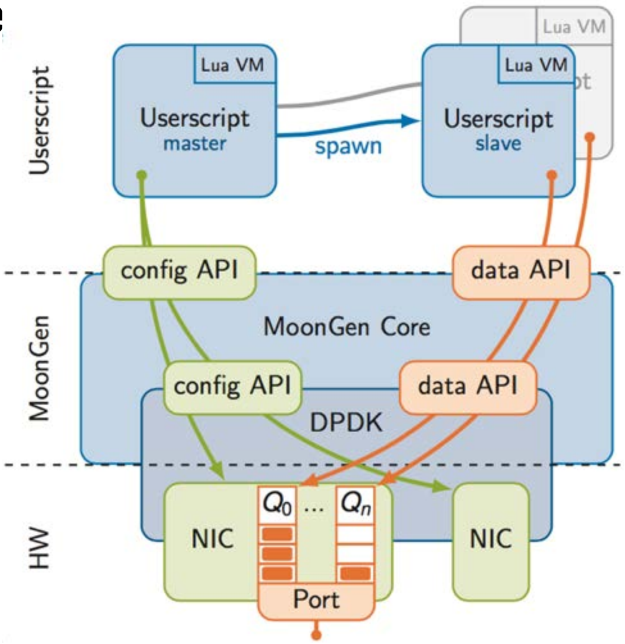
- Expensive
- Difficult to deploy
- Inflexible



Spirent traffic generator

MoonGen

- Inexpensive: Commercial Off-The-Shelf hardware
- Fast: DPDK for packet I/O, multi-core support
- Easy to deploy: simple software setup
- Flexible: user-controlled Lua scripts
- Precise
    - Timestamping: Utilize hardware features found on modern commodity NICs
    - Rate control: Hardware features and novel software approach



[ANRP17] Internet Research Task Force (IRTF) Applied Networking Research Prize, IETF-100, Nov. 2017, https://irtf.org/anrp

[ANCS17] Paul Emmerich, Sebastian Gallenmüller, Gianni Antichi, Andrew Moore, Georg Carle: Mind the Gap – A Comparison of Software Packet Generators, ACM/IEEE Symposium on Architectures for Networking and Communications Systems 2017

Usage of MoonGen/libmoon

Name	Usage scenario	Publication
High-performance applications:		
FlowScope	Tool for high-performance flow capture and analysis	[11], [12]
MoonRoute	Extensible high-performance router	[4], [13]
Benchmarking tools:		
RFC 2544	Modular benchmarking tool	[14], [15]
OPNFV VSPERF	Automated NFV testing framework	[16], [17]
FLOWer	High-performance switch benchmarking	[18], [19]
Traffic & packet generation:		
NFVnice	Throughput and latency measurements	[20]
Verified NAT	Throughput and latency measurements	[21]
PISCES	Throughput measurements	[22], [23]
Sonata	Replaying CAIDA traces	[24]
DoS flood generator	DNS and TCP SYN flooding attack tools	[25]–[27]
MoonGen / libmoon under test:		
MoonGen investigation	Precise and accurate rate control and timestamping	[3], [28], [29]
MoonGen timestamping	Investigation of timestamping for packet generators	[30]
Additions to MoonGen / libmoon:		
MoonStack	Easy-to-use and efficient packet creation	[31]

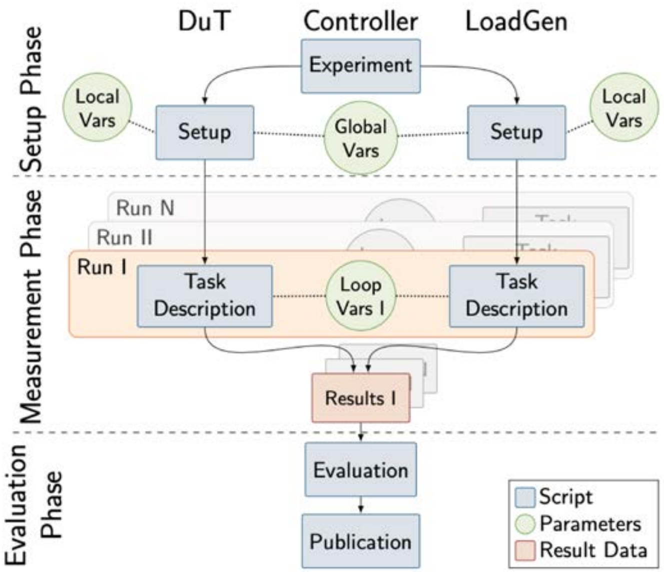
[Comsnets18] Gallenmüller, Scholz, Wohlfart, Scheitle, Emmerich, Carle, “High-Performance Packet Processing and Measurements,” COMSNETS 2018, Bangalore, India, Jan. 2018

TUM Testbed for Reproducible Experiments

- Automated workflow using **pos p**lain **o**rchestrating **s**ervice [pos] workflow for reproducible experiments
- Throughput - packets per second, bytes per second, frame loss rate
- Latency - Median, average, worst case, percentiles, ...
- White-box - Hardware and software events; interrupts, cache misses

[pos] Sebastian Gallenmüller, Dominik Scholz, Henning Stubbe, Georg Carle, "The pos Framework: A Methodology and Toolchain for Reproducible Network Experiments," in The 17th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '21), Munich, Germany, Dec. 2021

[SLICES] ESFRI - European Strategy Forum on Research Infrastructures; pos with TUM Baltikum Testbed: part of SLICES Research Infrastructure https://slices-ri.eu/



pos workflow

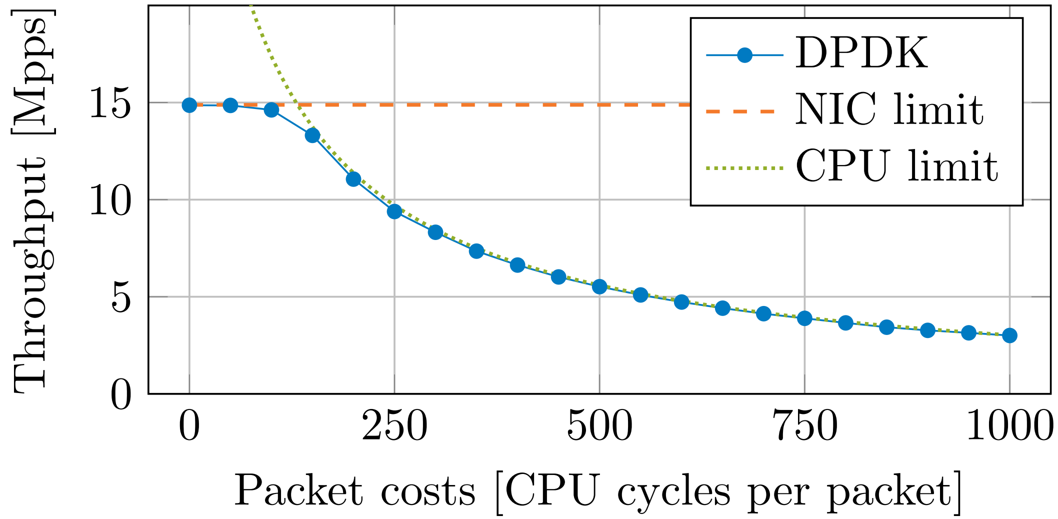
Performance Evaluation: Node Bottlenecks

Hardware

- Network Bandwidth
- NIC Processing Capacity
- PCIe Bandwidth
- Memory Bandwidth
- CPU Cache Size
- CPU Cache Line Length

Software

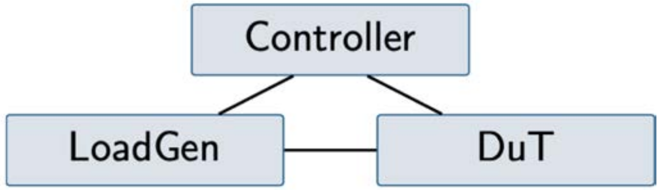
- CPU utilization per packet
- Kernel / network stack overhead



Throughput limit = min(NIC limit, CPU limit)
NIC limit = 14.88 Mpps (10 Gb Ethernet)
CPU limit = available CPU cycles

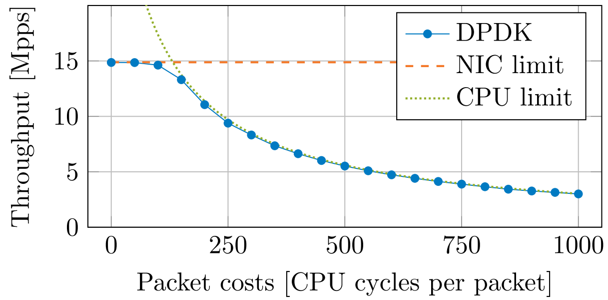
System Analysis

Measurement setup



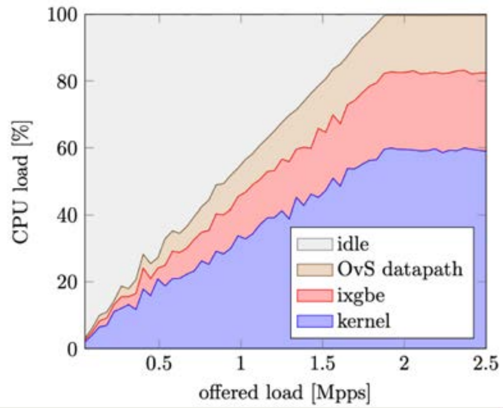
Black-box

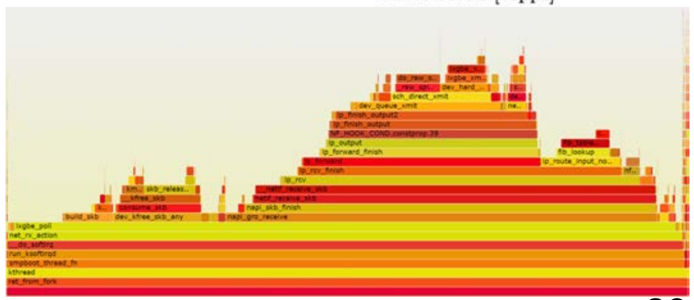
- Throughput
  - Packets / bytes per second
  - Frame loss rate
- Latency
  - Median, average, worst case, percentiles, ...



White-box

- Hardware and software events
  - Cycles, Interrupts, L1/L2/L3 cache misses
  - Per second, per packet, per function





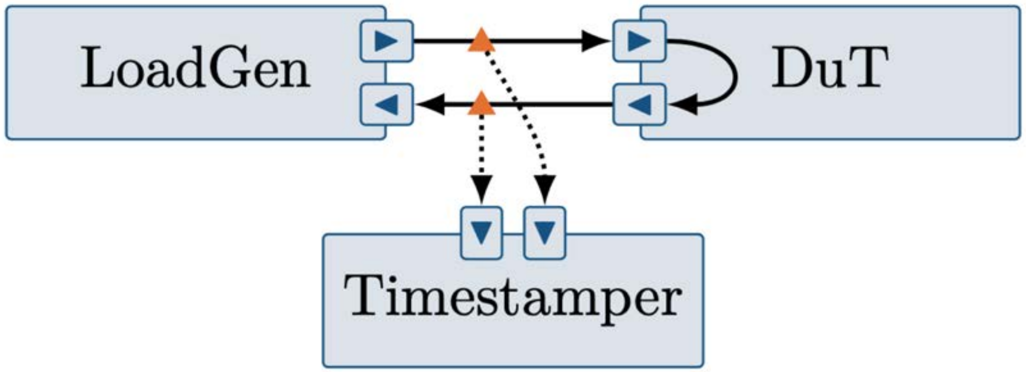
5G Low-Latency Services

5G Ultra-Reliable Low-Latency Communication (URLLC)

- Ultra reliable: 99.999% packet delivery probability
- Low latency: 1ms one-way latency in Radio Access Network (RAN)

5G Service provisioning with Virtual Network Functions (VNF)

- Virtualized environment: Linux, kvm
- Network function: Snort3



percentiles	50th	99th	99.9th	99.99th	99.999th
Snort 3 forwarder	69 µs	88 µs	107 µs	1.7 ms	2.5 ms

⇨ 99.99th percentile already violates URLLC

Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich

SLICES Research Infrastructure

European Scientific Large-Scale Infrastructure for Computing/Communication Experimental Studies

Third generation Mid-Scale Test Platform

USA NSF PAWR (Platforms for Advanced Wireless Research): NSF + Industry, 100M€, 2017-2022

NSF Fabric: NSF, 20 M€, 2019-2023

Colosseum: NSF-DARPA, 20+7,5M$, 2017-2025.

BRIDGES: NSF, 2.5M€, 2020-2023

EU Horizon Europe
ICT 17-19-52, 2018-2022, 205 M€
SNS Stream C, first call, 2022-2025, 25M€

Japan NICT R&D Shared Open Platform
200 M$

China CENI
Chinese Experimental National Infrastructure
2018-2022
190 M€

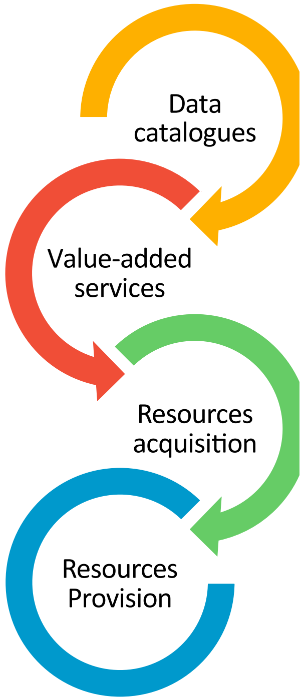
SLICES, first in digital sciences to entered the ESFRI Roadmap 2021

- ESFRI:
  European Strategy Forum on Research Infrastructures

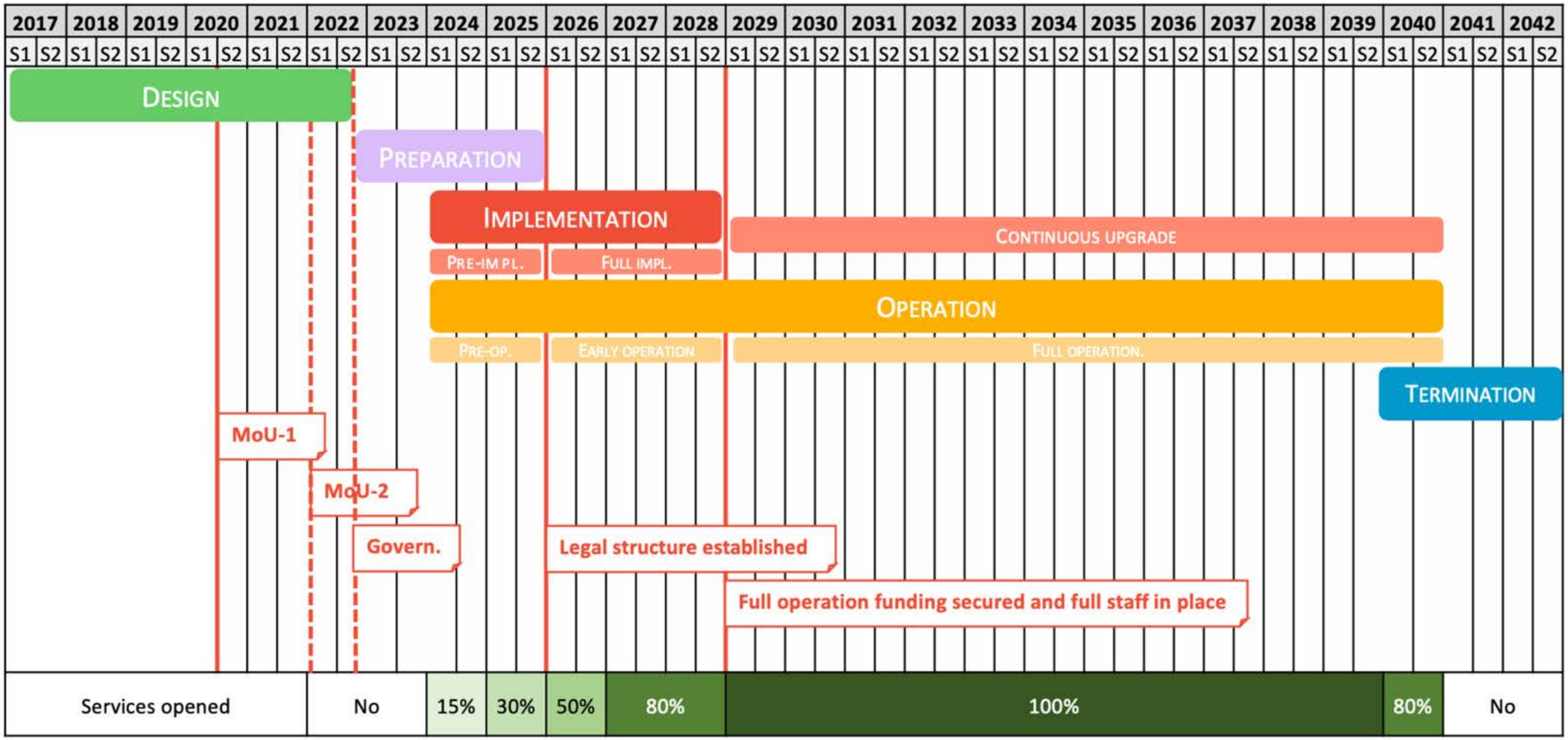
- **SLICES** is an **RI** to support the **academic and industrial research community** that will design, develop and deploy the **Next Generation** of **Digital Infrastructures**

- **SLICES-RI** is a **distributed RI** providing several **specialized instruments** on challenging research areas of Digital Infrastructures, by **aggregating** networking, computing and storage **resources** across countries, nodes and sites

- **Scientific domains:** networking protocols, services, radio technologies, data collection, parallel and distributed computing, cloud and edge-based computing architectures and services



SLICES Timeline



Questions?