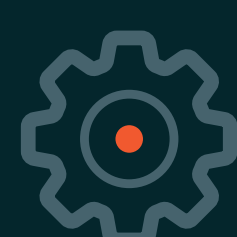




scalacenter

For open source. For education.

JOIN THE SCALA COMMUNITY!



19

Workshops



1M+

Scala Users



72%

Students taking MOOCs for free



74%

Employed Students

Vibrant Job Market

800k+ learners globally

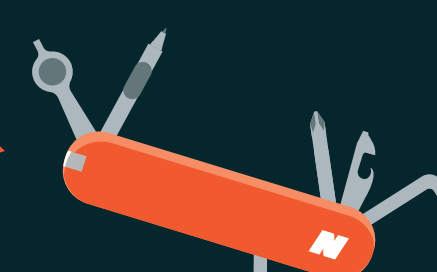
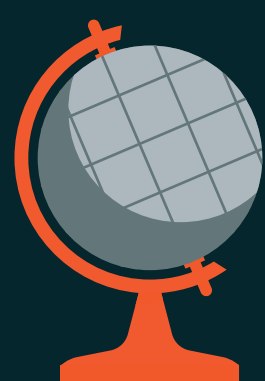
Large Community

1M+ Scala users globally



REALISING SCALA OPEN SOURCE FULL POTENTIAL

CHF 800k Annual budget



Simplified user experience

20+ tools developed & maintained

Responsible stakeholders

15 Advisory board members



14k

Monthly viewing hours



255k+

Monthly site visits



203

Countries



11%

Female students



CODE EDITORS SUPPORT
1.3M downloads



Scaladex

OSS ECOSYSTEM CURATION
28k monthly users



IN-BROWSER CODE EXECUTION AND SNIPPET SHARING
26k monthly users



SCALA 3 MIGRATION HARNESS

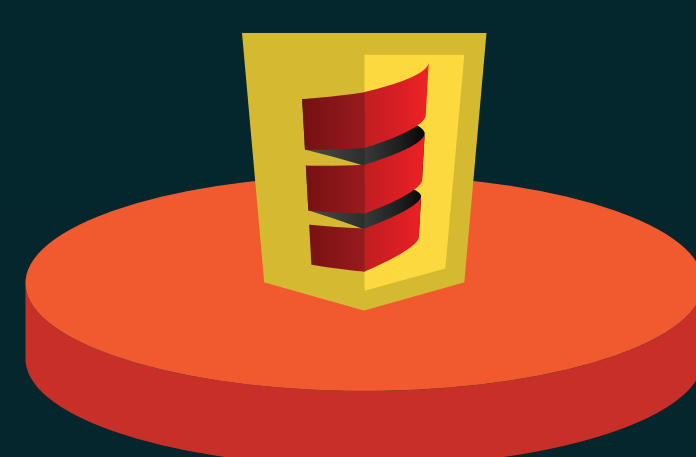


FAST COMPILATION
340k downloads

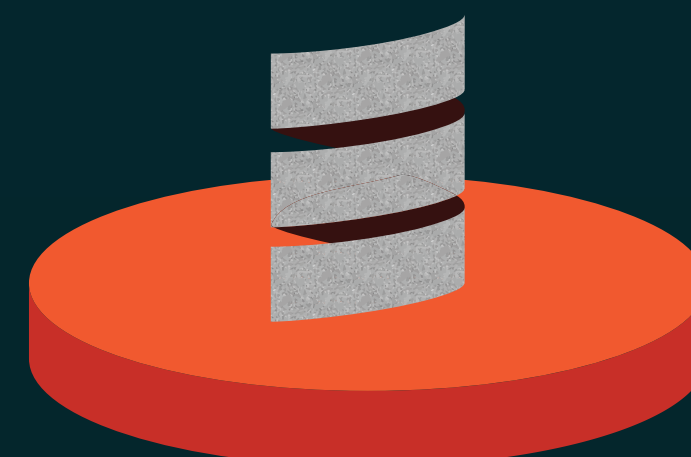


AUTOMATIC CODING STYLE ENFORCEMENT
470k downloads

PLATFORMS



SCALA.JS



SCALA NATIVE



Check out SCALA 5 year report

EPFL

SwissBorg

knóldus
knowledge - commitment - results

Goldman Sachs



Lightbend

47

Spotify

LUNATECH

VIRTUSLAB

databricks

Ali Ansari[†], Shanjing Lin[†], Bugra Eryilmaz[†], Ayan Chakraborty[†], Rafael Pizarro[†],

Babak Falsafi^{†‡}, Michael Ferdman[‡]

[†]PARSA, EPFL

[‡]EcoCloud, EPFL

[‡]Stony Brook University

Cloud Server Efficiency

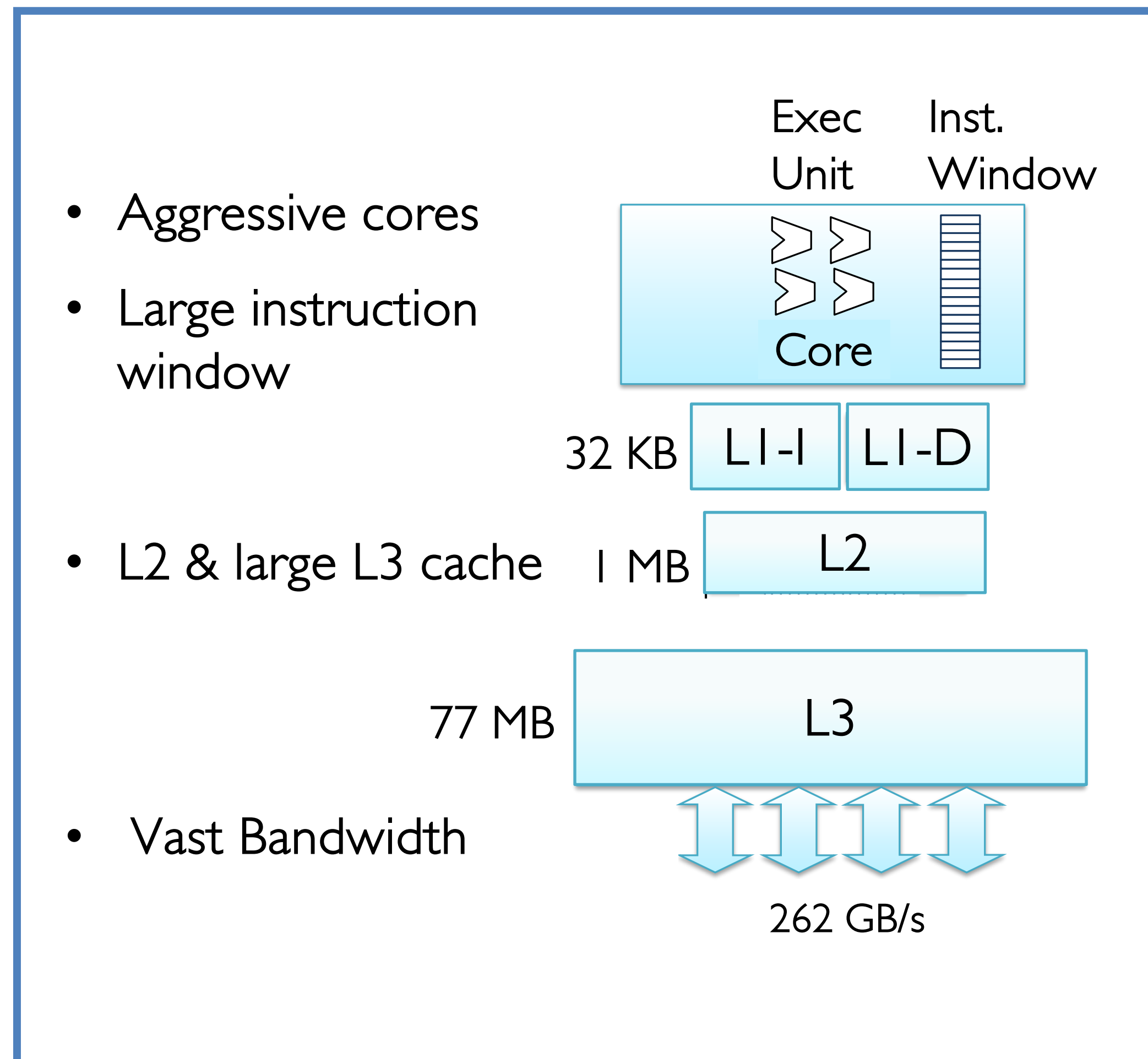


amazon.com, YouTube, Microsoft Bing, Twitter, G

Constant demand for more servers

Increasing costs of HW, space & power

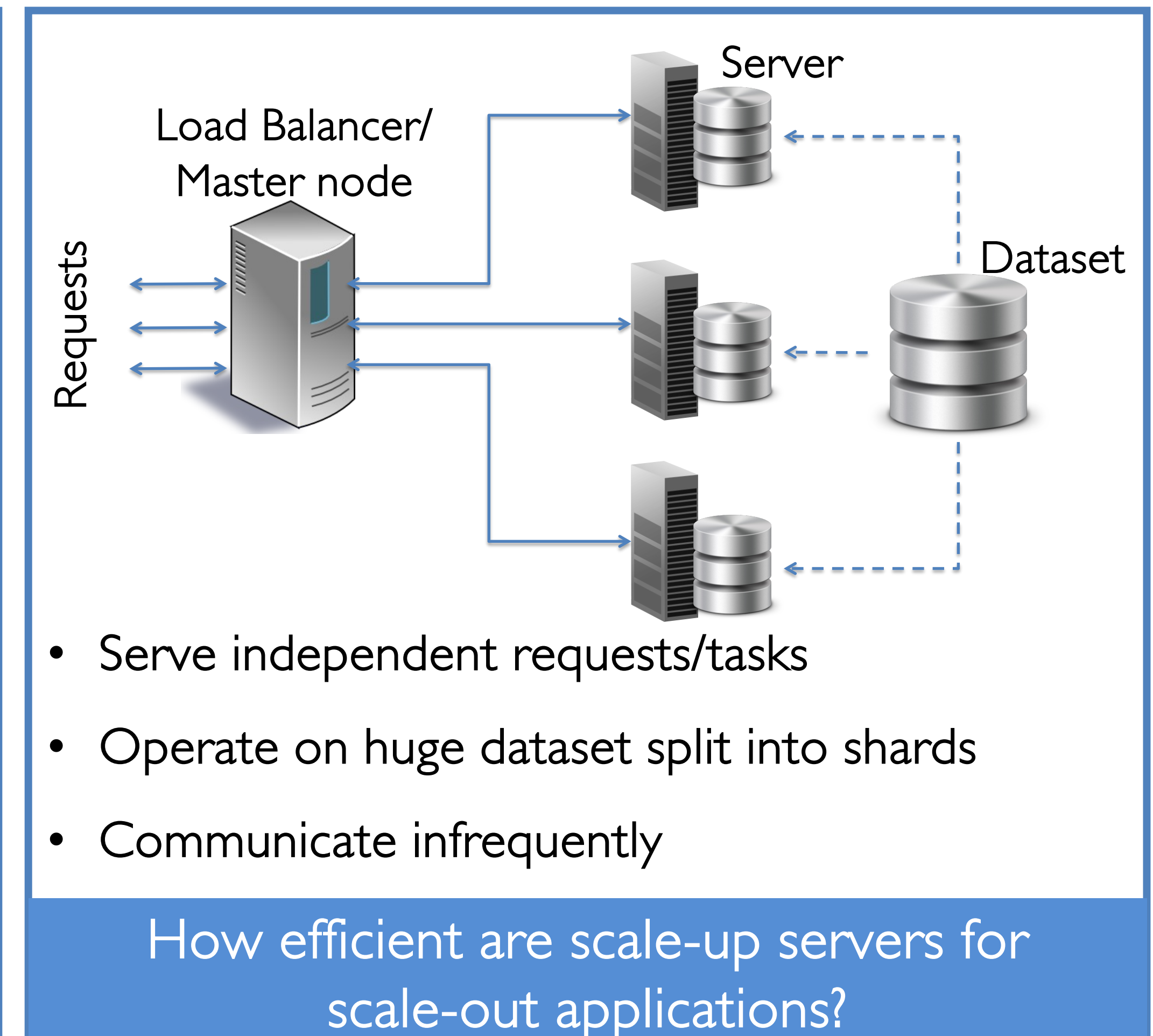
Modern Servers are Scale-Up



- Aggressive cores
- Large instruction window
- L2 & large L3 cache
- Vast Bandwidth

Exec Unit, Inst. Window, Core, 32 KB, LI-I, LI-D, 1 MB, L2, 77 MB, L3, 262 GB/s

Cloud Applications are Scale-out

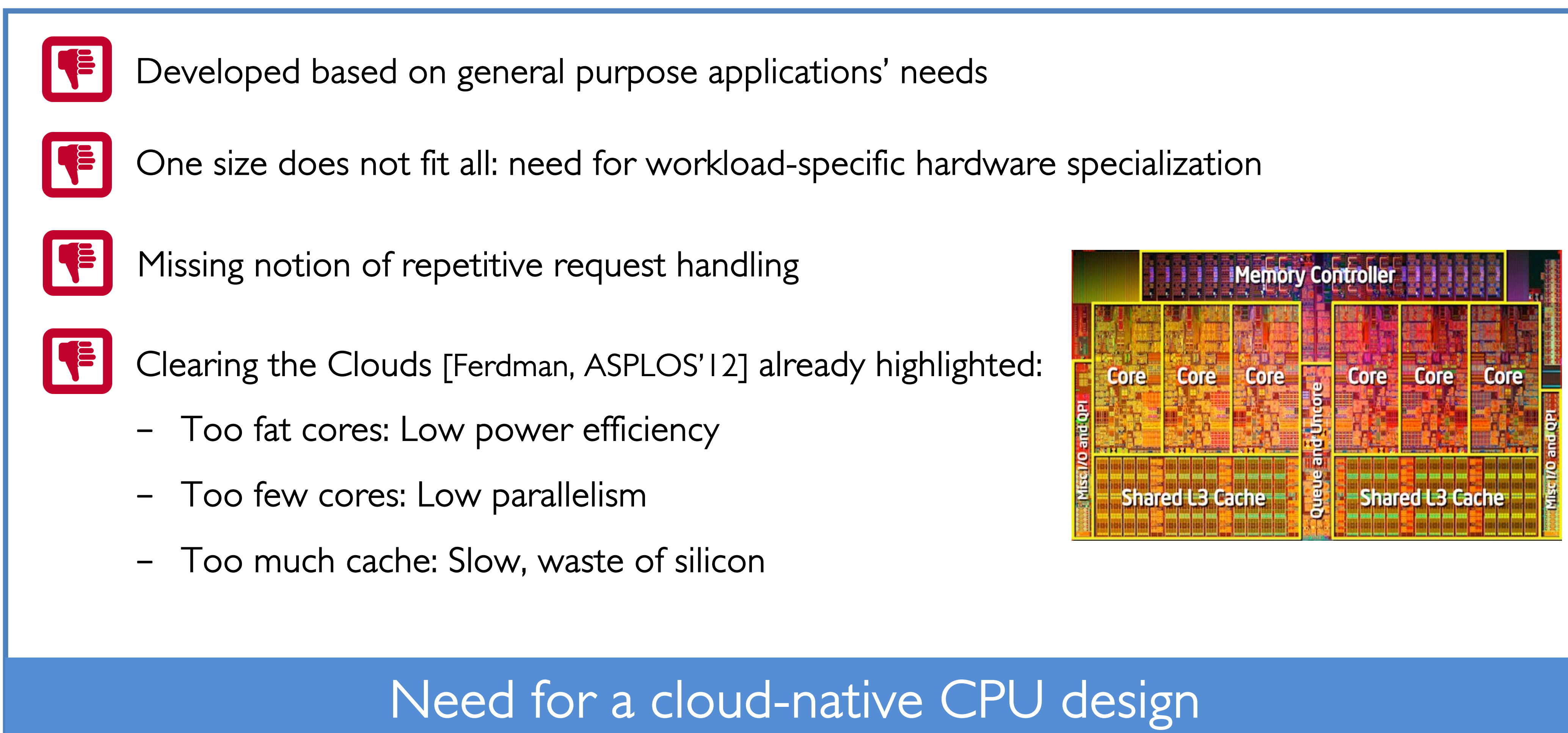


Load Balancer/ Master node, Server, Dataset, Requests

- Serve independent requests/tasks
- Operate on huge dataset split into shards
- Communicate infrequently

How efficient are scale-up servers for scale-out applications?

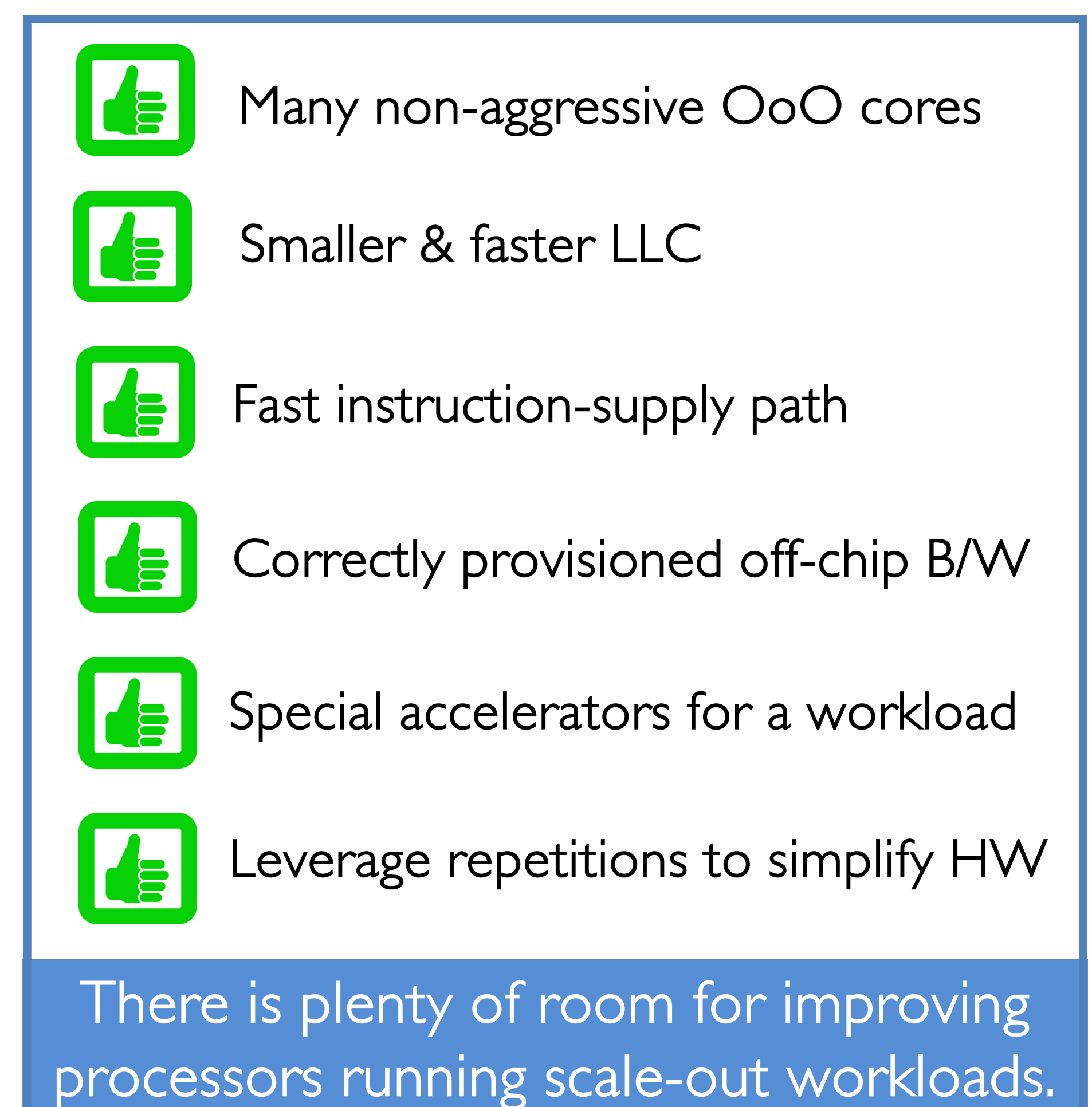
Why not Conventional Scale-Up Processors?



- Developed based on general purpose applications' needs
- One size does not fit all: need for workload-specific hardware specialization
- Missing notion of repetitive request handling
- Clearing the Clouds [Ferdman, ASPLOS'12] already highlighted:
 - Too fat cores: Low power efficiency
 - Too few cores: Low parallelism
 - Too much cache: Slow, waste of silicon

Need for a cloud-native CPU design

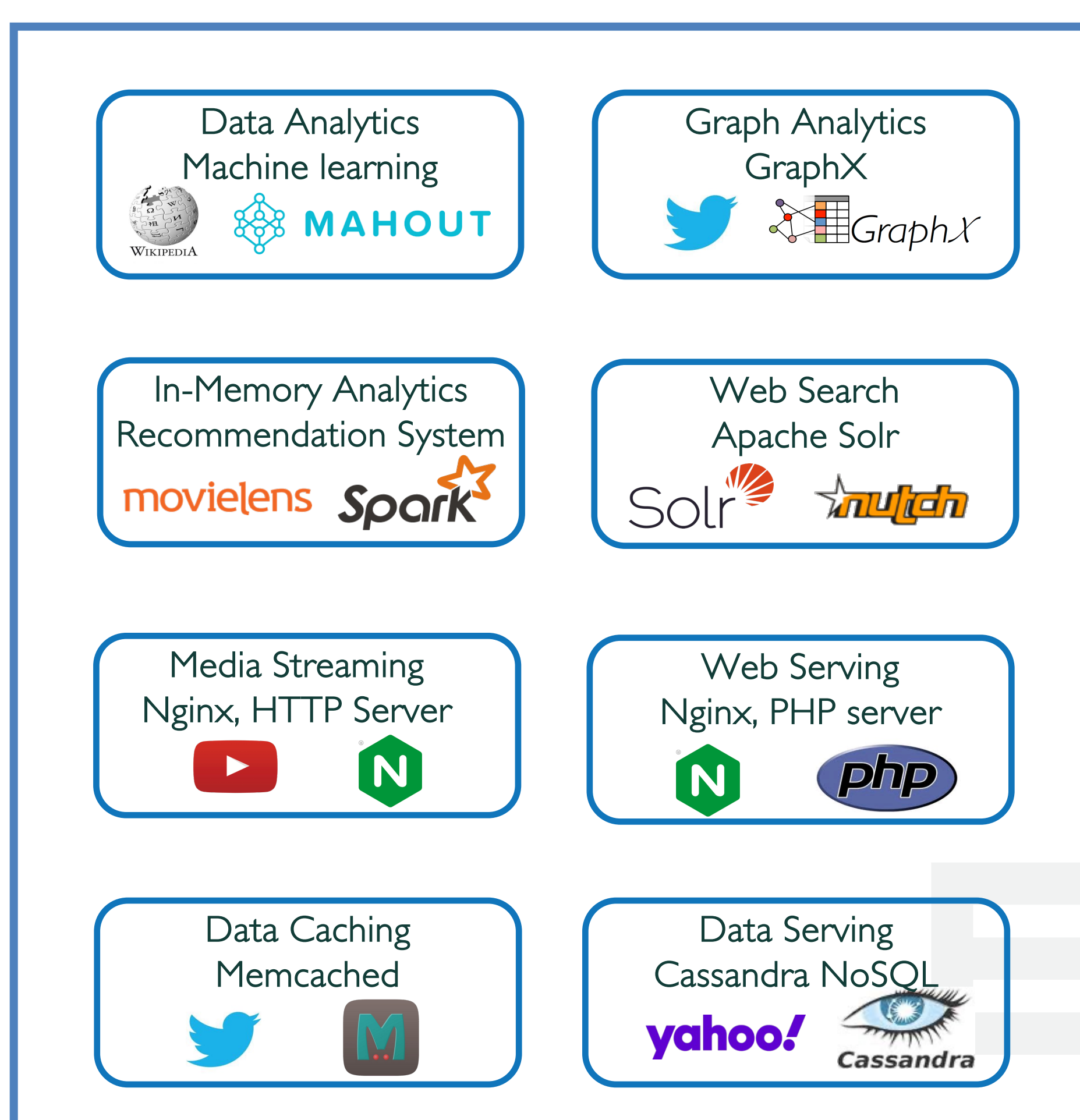
Processors for Scale-Out Workloads



- Many non-aggressive OoO cores
- Smaller & faster LLC
- Fast instruction-supply path
- Correctly provisioned off-chip B/W
- Special accelerators for a workload
- Leverage repetitions to simplify HW

There is plenty of room for improving processors running scale-out workloads.

CloudSuite 4.0



Data Analytics Machine learning MAHOUT

Graph Analytics GraphX

In-Memory Analytics Recommendation System movielens Spark

Web Search Apache Solr

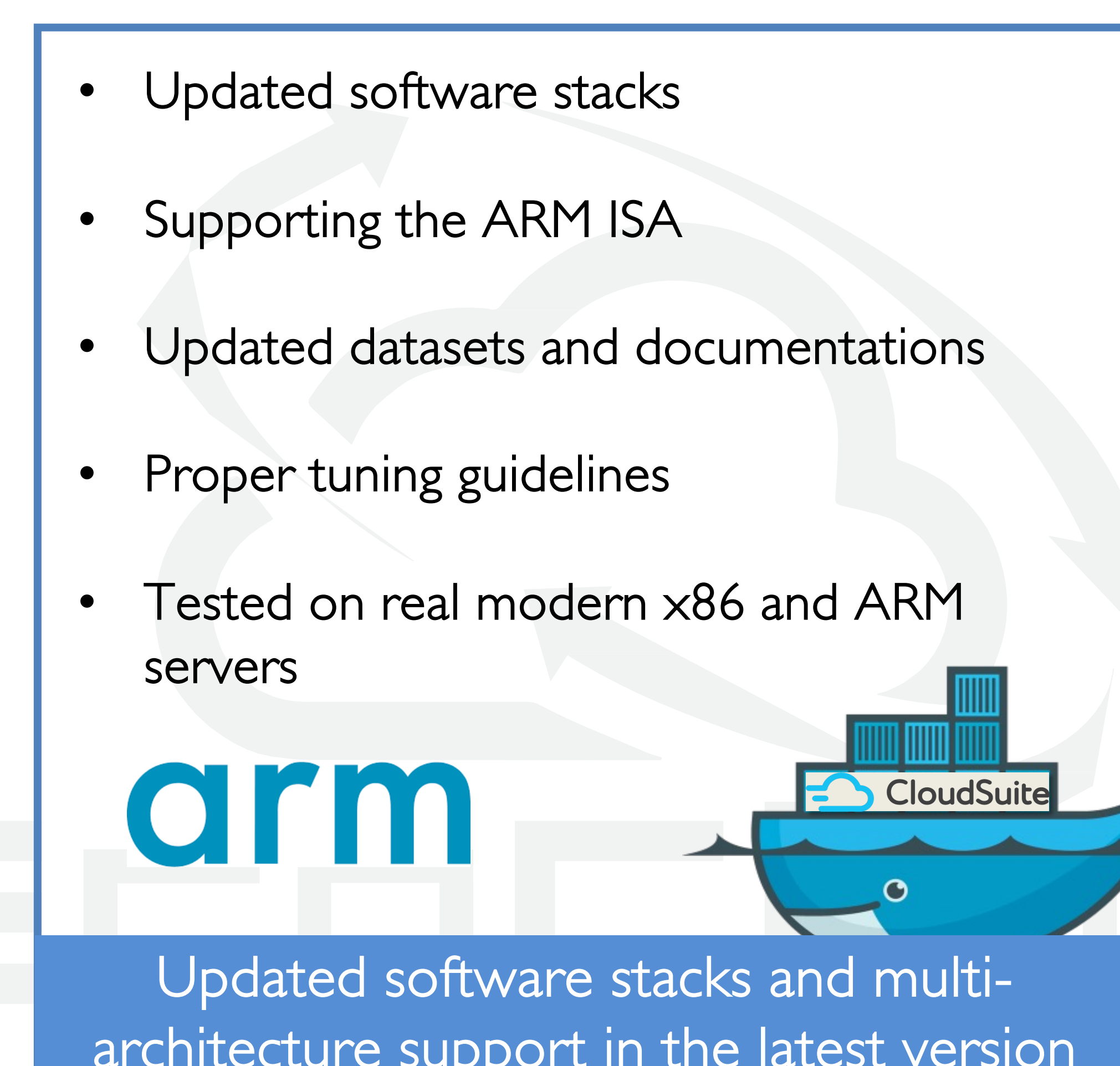
Media Streaming Nginx, HTTP Server

Web Serving Nginx, PHP server

Data Caching Memcached

Data Serving Cassandra NoSQL yahoo! Cassandra

What is New in CloudSuite 4.0?

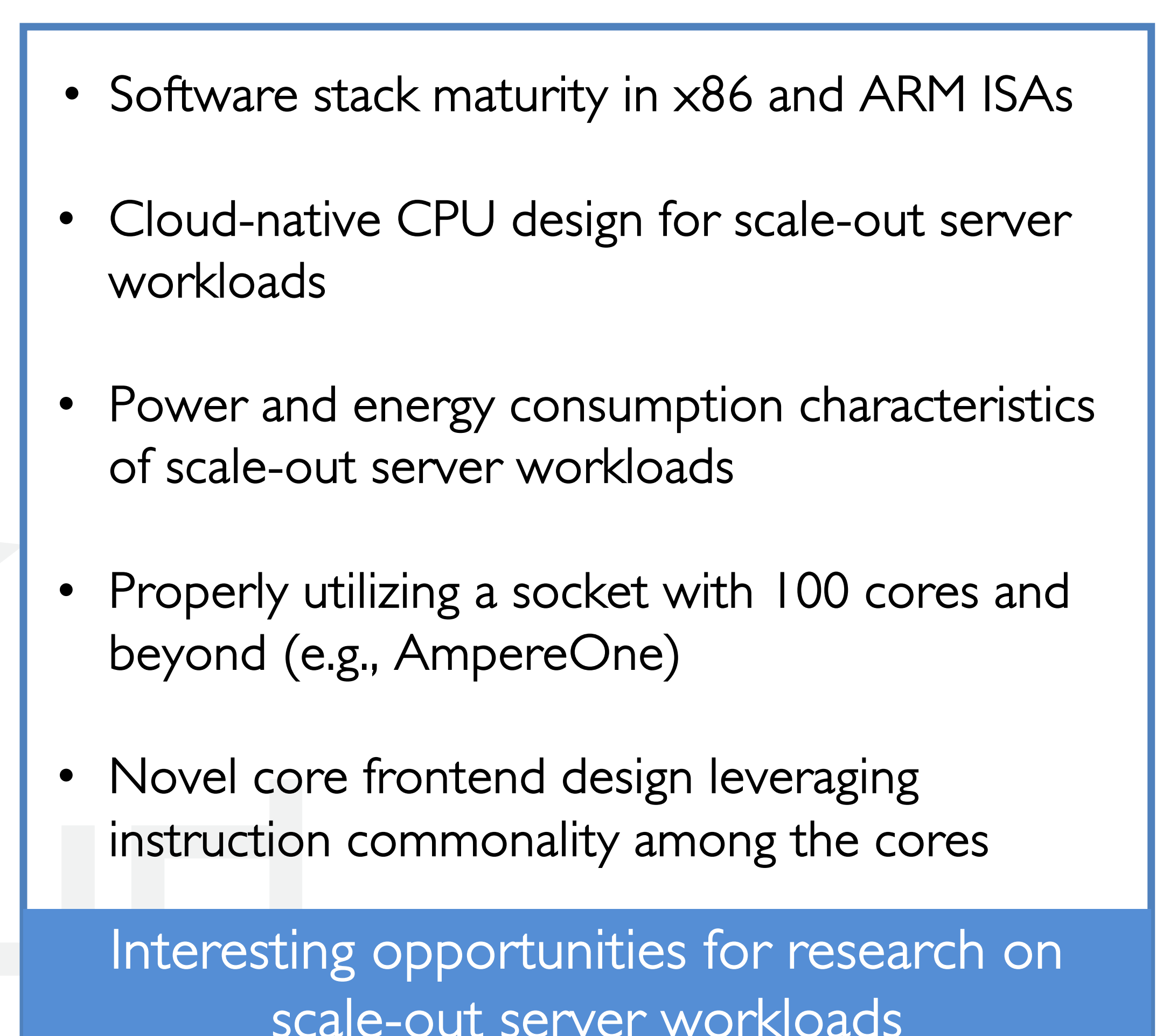


- Updated software stacks
- Supporting the ARM ISA
- Updated datasets and documentations
- Proper tuning guidelines
- Tested on real modern x86 and ARM servers

arm CloudSuite

Updated software stacks and multi-architecture support in the latest version

Research Directions

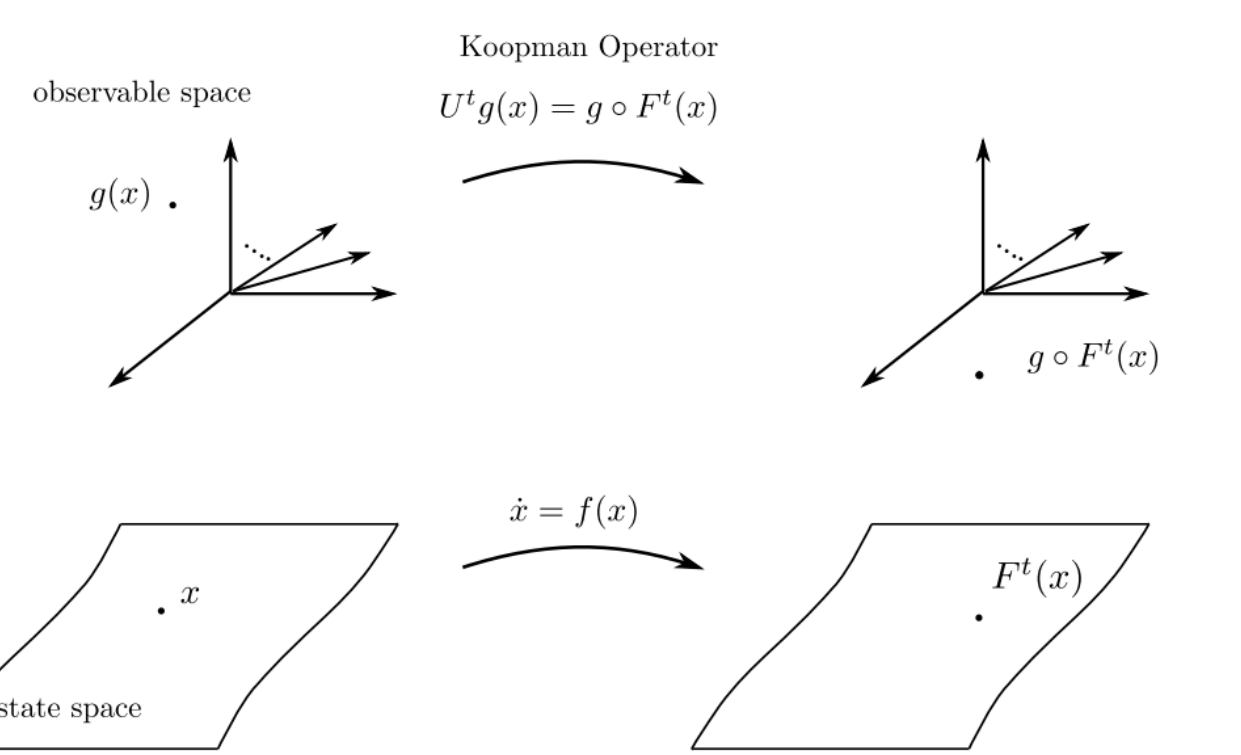


- Software stack maturity in x86 and ARM ISAs
- Cloud-native CPU design for scale-out server workloads
- Power and energy consumption characteristics of scale-out server workloads
- Properly utilizing a socket with 100 cores and beyond (e.g., AmpereOne)
- Novel core frontend design leveraging instruction commonality among the cores

Interesting opportunities for research on scale-out server workloads

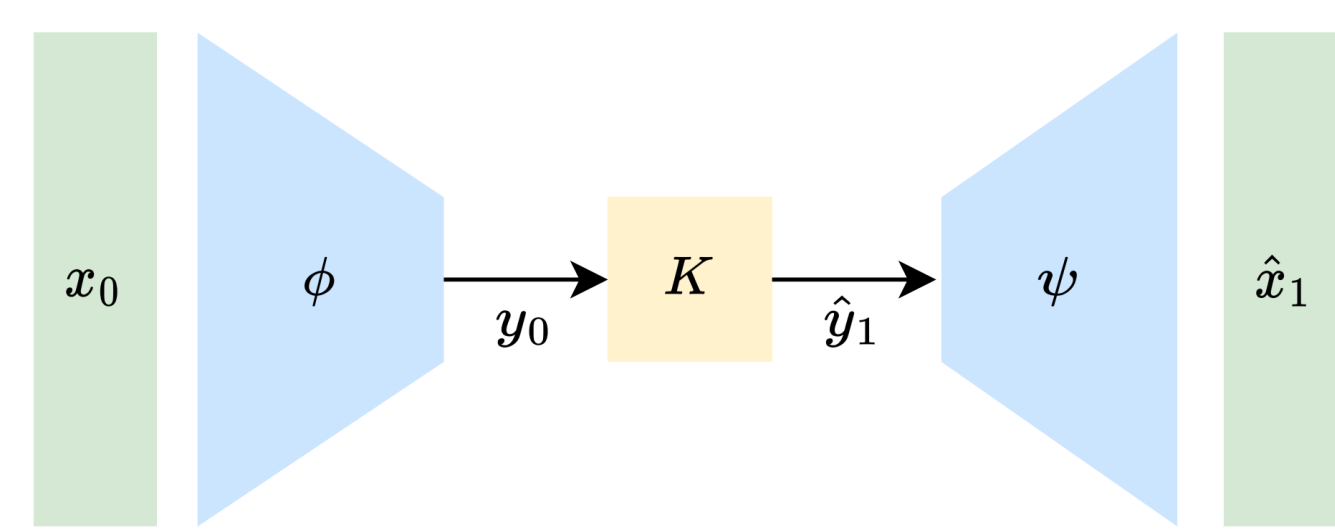
Motivation

- A crucial criterion for reliable health indicators is their capability to discern a degradation trend
- Koopman theory provides a means to find coordinates where dynamics becomes linear
- Based on this theory, there was previously proposed the Deep Koopman Operator (DKO) approach that learns linear dynamics operator in the latent space of an autoencoder
- In this work, **our objective** is to evaluate if DKO and proposed in this work **Koopman-inspired Degradation Model (KIDM)** are able to construct reliable health representation of industrial dynamical systems and develop an approach that is able to capture the hidden degradation dynamics for the trendable degradation predictions
- We learn the dynamics of the process by DKO and KIDM and evaluate the learned representations in the Remaining Useful Life (RUL) prediction

**Koopman Operator**

- ✓ Finite-dimensional
- Infinite-dimensional
- Nonlinear
- ✓ Linear

- There exist infinite-dimensional space where the system dynamics becomes linear
- In applications, one usually looks for finite-dimensional approximation of such space

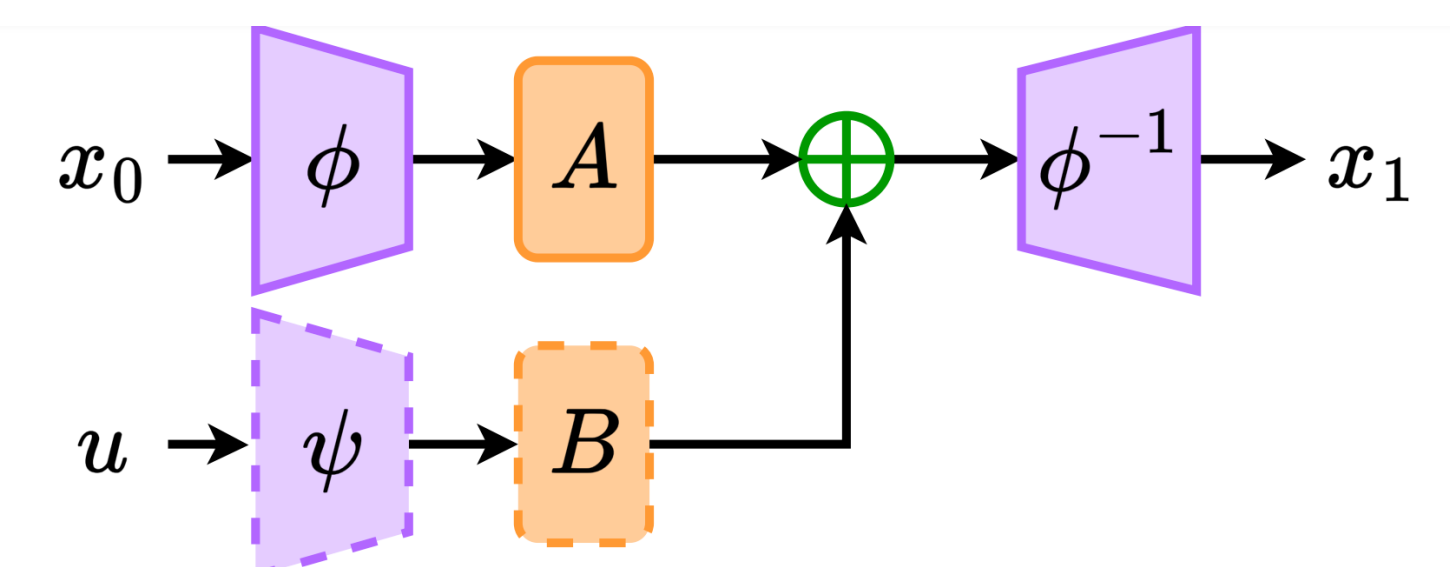
Deep Koopman Operator (DKO)

$$\mathcal{L}_{rec} = \|x_t - \psi(\phi(x_t))\|$$

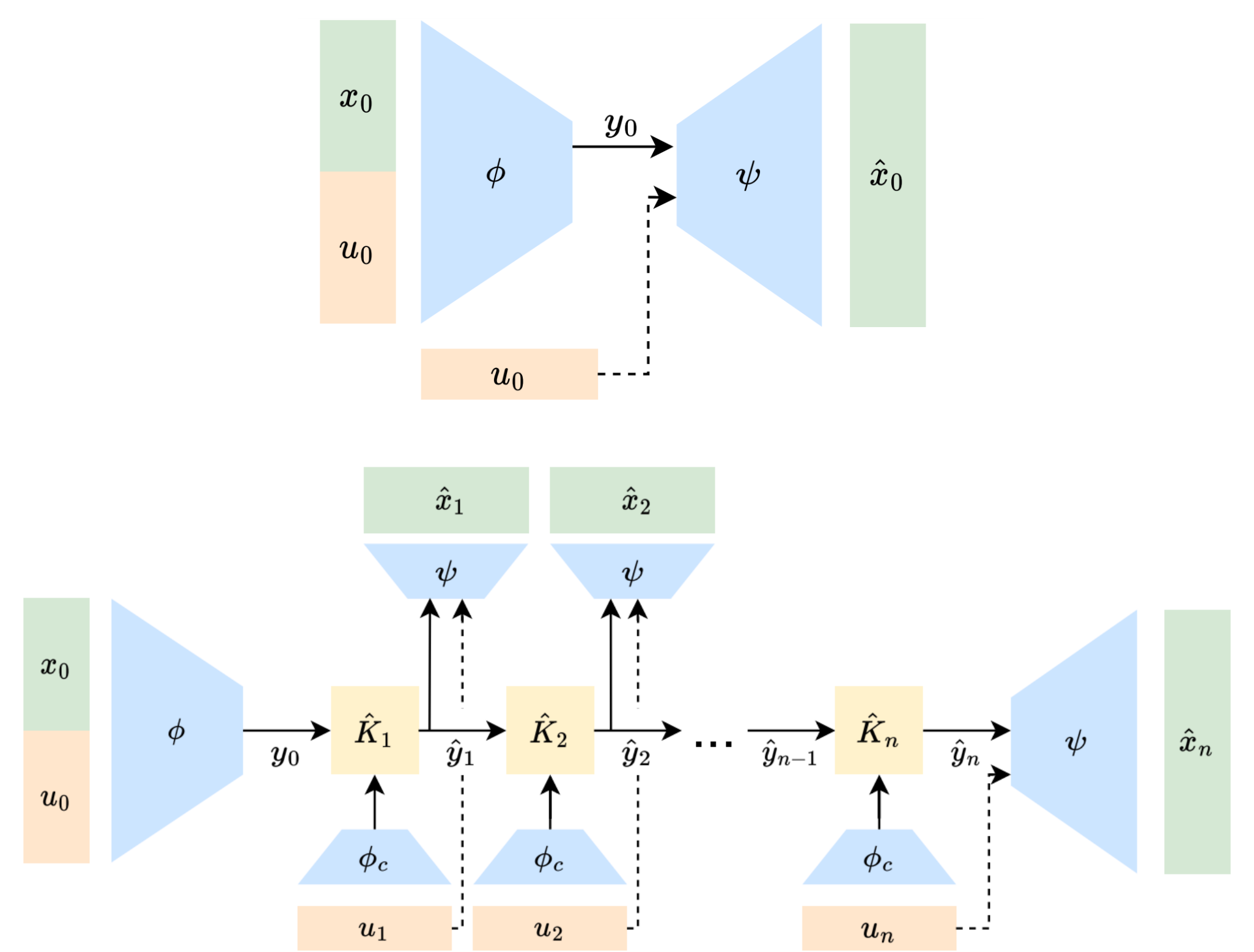
$$\mathcal{L}_{lin} = \|\phi(x_{t+m}) - \mathbf{K}^m \phi(x_t)\|$$

$$\mathcal{L}_{pred} = \|x_{t+m} - \psi(\mathbf{K}^m \phi(x_t))\|$$

- Based on autoencoder architecture
- Learns invertible mapping of system state to latent space
- Further, learns linear dynamics in the latent space and in the initial system state space over multiple timesteps

Previous DKO-based model for control systems

- Previous DKO-based setups for controlled systems mainly have a linear mapping to Koopman observables
- Linear mapping of control may limit the choice of observables
- Control has significant influence on the observables space

Proposed Koopman-Inspired Degradation Model (KIDM)

$$\mathcal{L}_{rec} = \|x_t - \psi(\phi(x_t), u_t)\|$$

$$\mathcal{L}_{lin} = \|\phi(x_{t+m}) - \hat{K}_{t+m} \hat{K}_{t+m-1} \dots \hat{K}_{t+1} \phi(x_t)\|$$

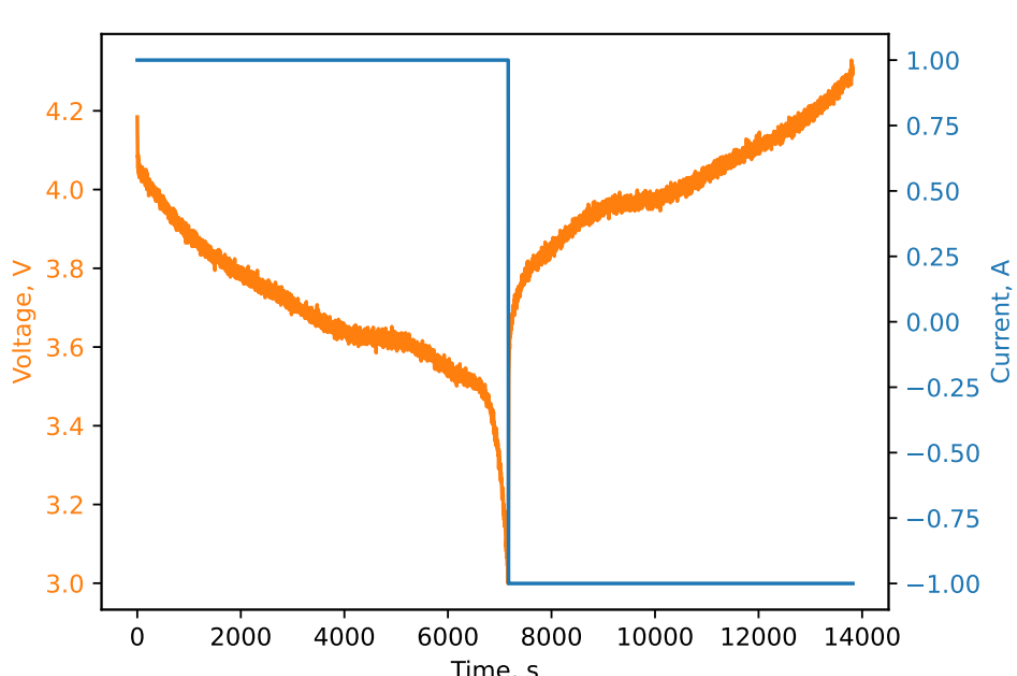
$$\mathcal{L}_{pred} = \|x_{t+m} - \psi(\hat{K}_{t+m} \hat{K}_{t+m-1} \dots \hat{K}_{t+1} \phi(x_t))\|$$

- Koopman operator is produced by encoding the control vector
- Control vector is applied again, together with the latent representation, to predict the future state
- Reconstruction task without applying Koopman operators in the latent space allows Koopman operator to learn the dynamics of degradation

Case studies**1. Li-ion battery under constant current load**

Hidden health parameters:

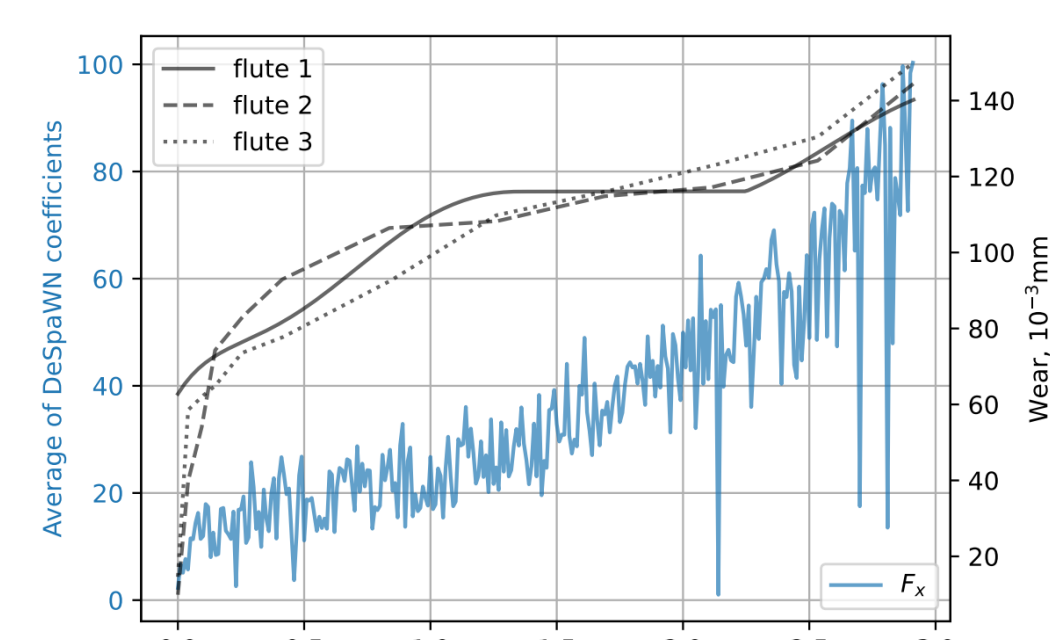
- Number of available Li-ions
- Internal resistance of battery

**2. CNC milling machine cutters**

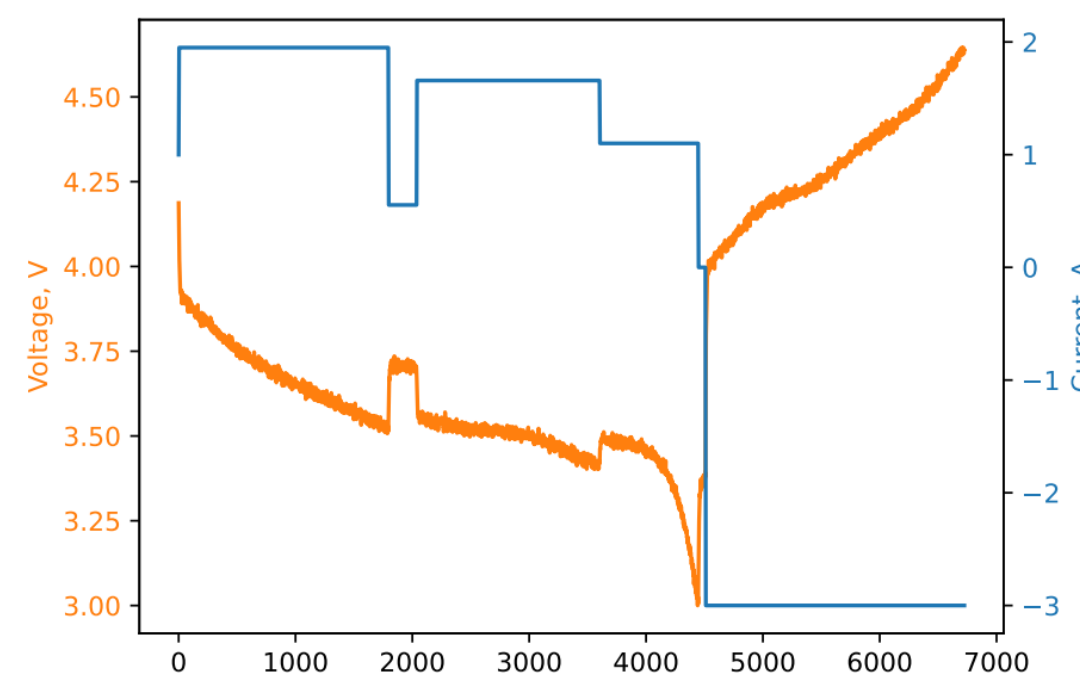
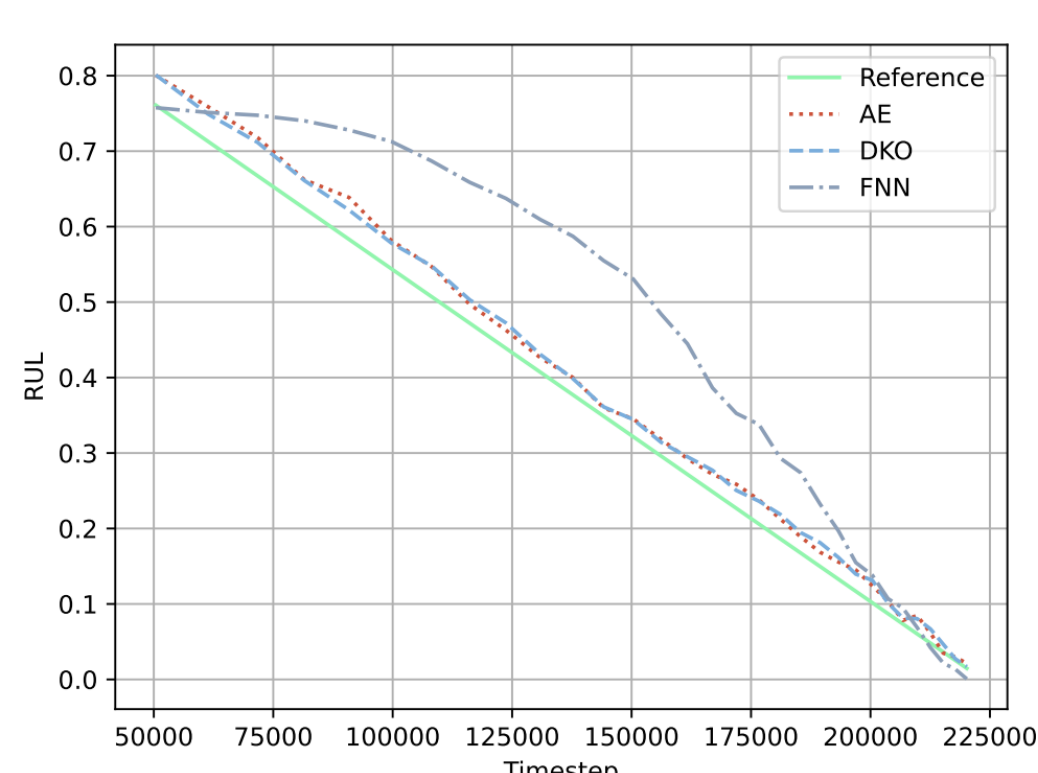
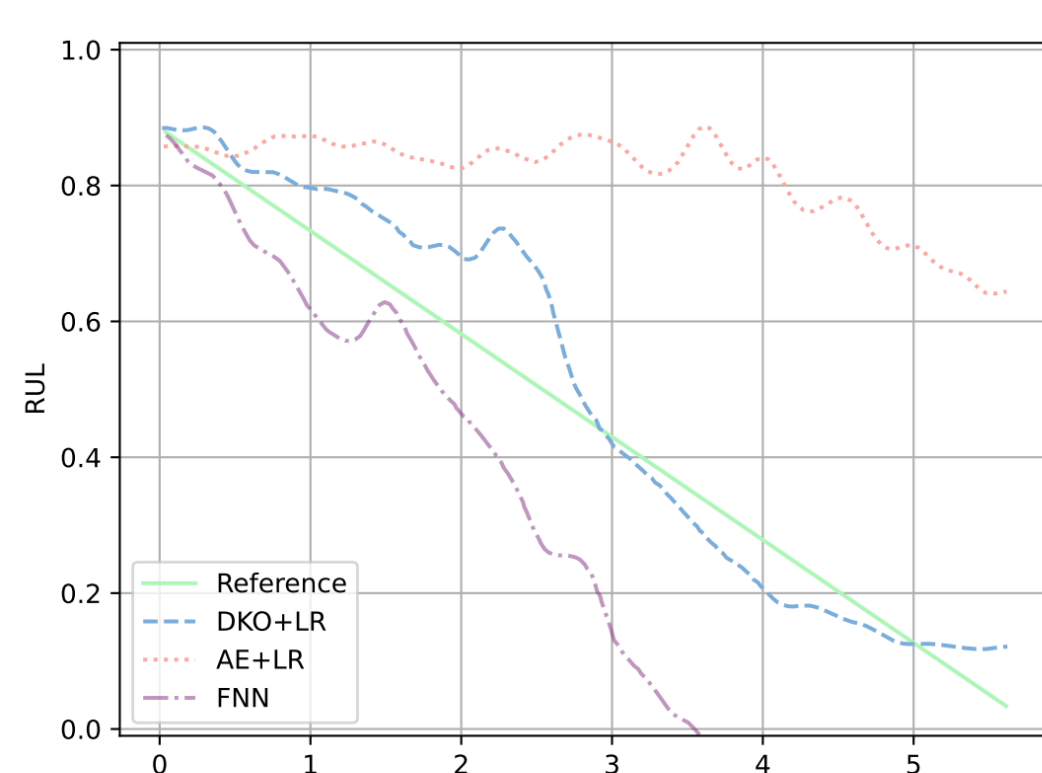
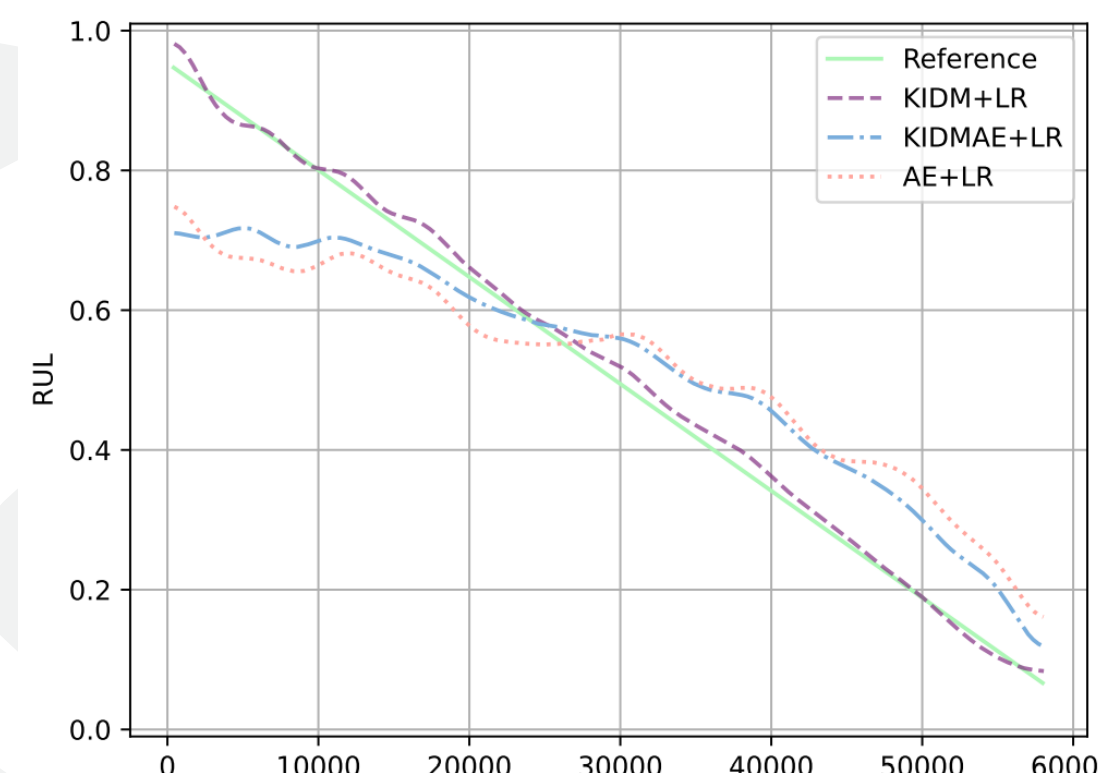
High frequency time-series. Each cutter has 3 flutes that accumulate wear during the lifetime.

Hidden health parameters:

- Number of available Li-ions
- Internal resistance of battery

**3. Li-ion battery under varying current load**

Discharge current is random piece-wise constant function. Hidden health parameters are similar to the case 1.

**RUL predictions****1. Li-ion battery under constant current load****2. CNC milling machine cutters****3. Li-ion battery under varying current load****Conclusions**

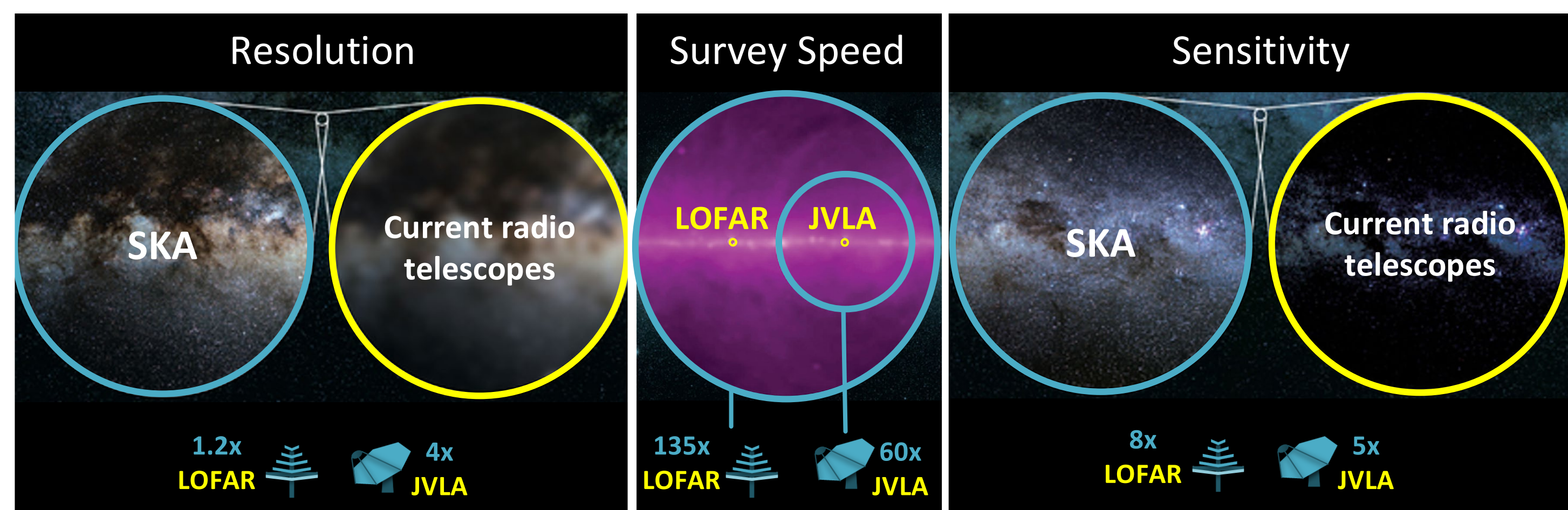
- DKO can be used to represent the health state of the dynamical system in the unsupervised manner
- Incorporating multiple steps prediction loss helps to build a more robust model
- We proposed KIDM algorithm showed superior performance in early lifetime RUL prediction
- DKO and KIDM approaches are robust in RUL prediction even with one ground truth run-to failure trajectory
- Learned Koopman operators can be further analyzed

References

- [1] B. Lusch, J. N. Kutz, S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nature Communications 9 (2018) 4950.
- [2] X. Li, 2010 phm society conference data challenge, IEEE Dataport (2021). doi:10.21227/jdx-dyy51. URL <https://dx.doi.org/10.21227/jdx-dyy51>
- [3] M. J. Daigle, K. Goebel, Model-based prognostics with concurrent damage progression processes, IEEE Transactions on Systems, man, and cybernetics: systems 43 (3) (2012) 535–546.

Denisa Constantinescu¹, Benoît Denkinger¹, Miguel Peon Quiros², Jean-Paul Kneib^{3,4}, David Atienza^{1,2}
¹Embedded Systems Laboratory, ²EcoCloud, ³LASTRO – École Polytechnique Fédérale de Lausanne, ⁴SKACH

Enabling Science Beyond Capabilities of Best Telescopes



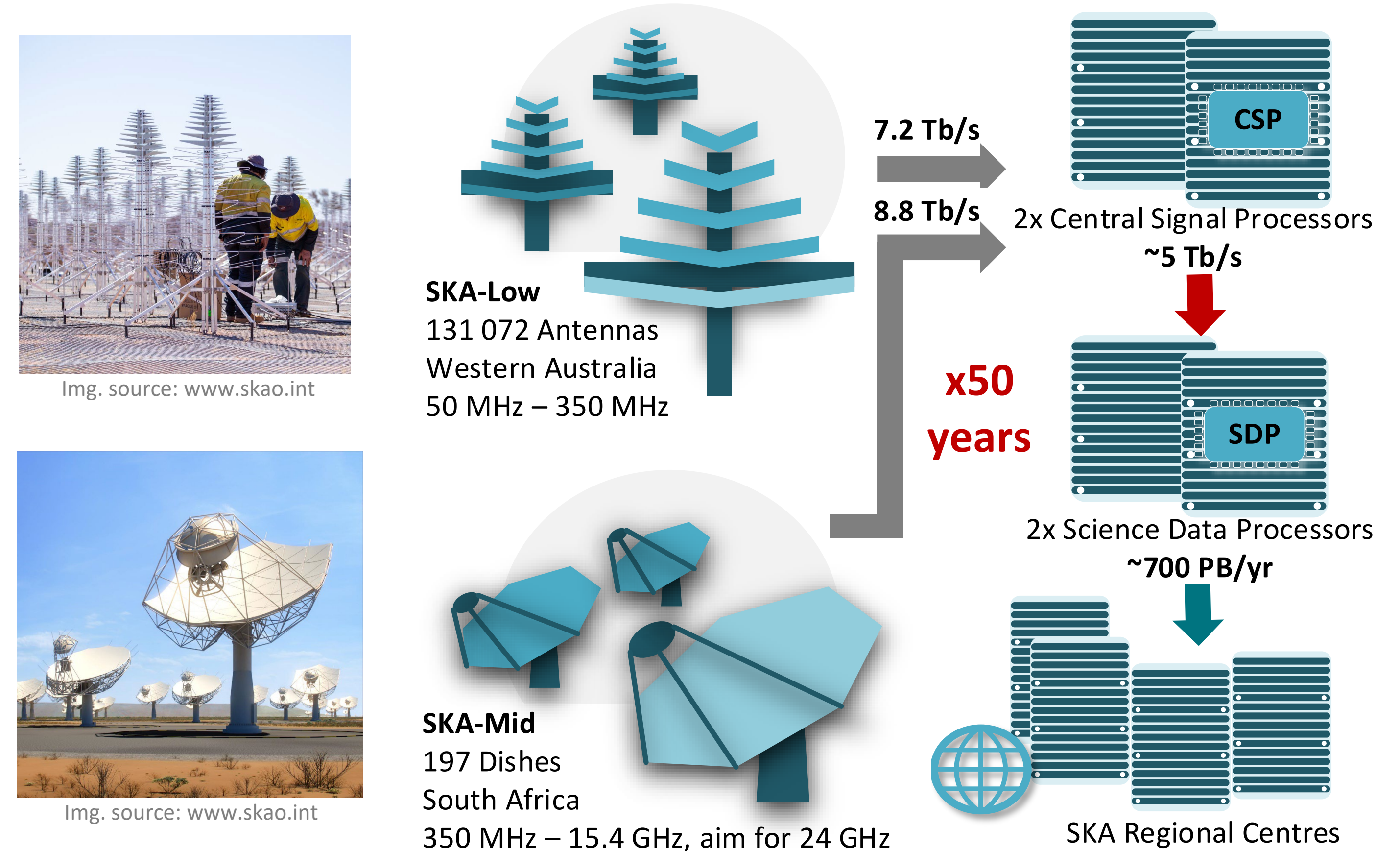
SKA Science Goals Include

Images source: www.skao.int



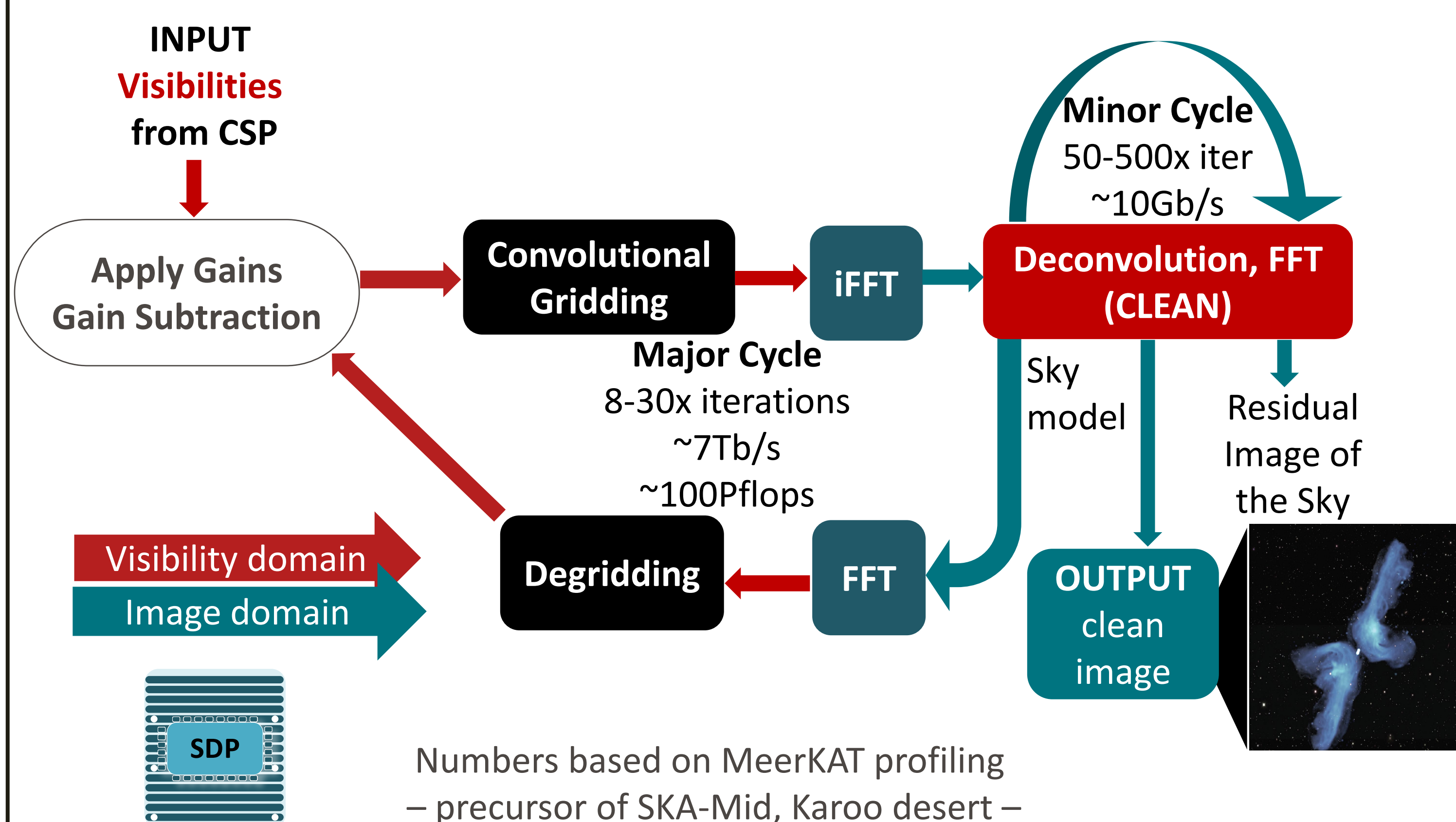
SKAO Consortium

The Square Kilometer Array Observatory in a Nutshell



Energy-Efficient HPC + Big Data Challenge

1 MWatt Budget for 250 PFlops for Science Data Processor



Top 3 Green500 Adastra
 [Power: 0.92 MWatt, 58.021 GFlops/Watts]
 ~Scales to 63 PFlops for 1 MWatt

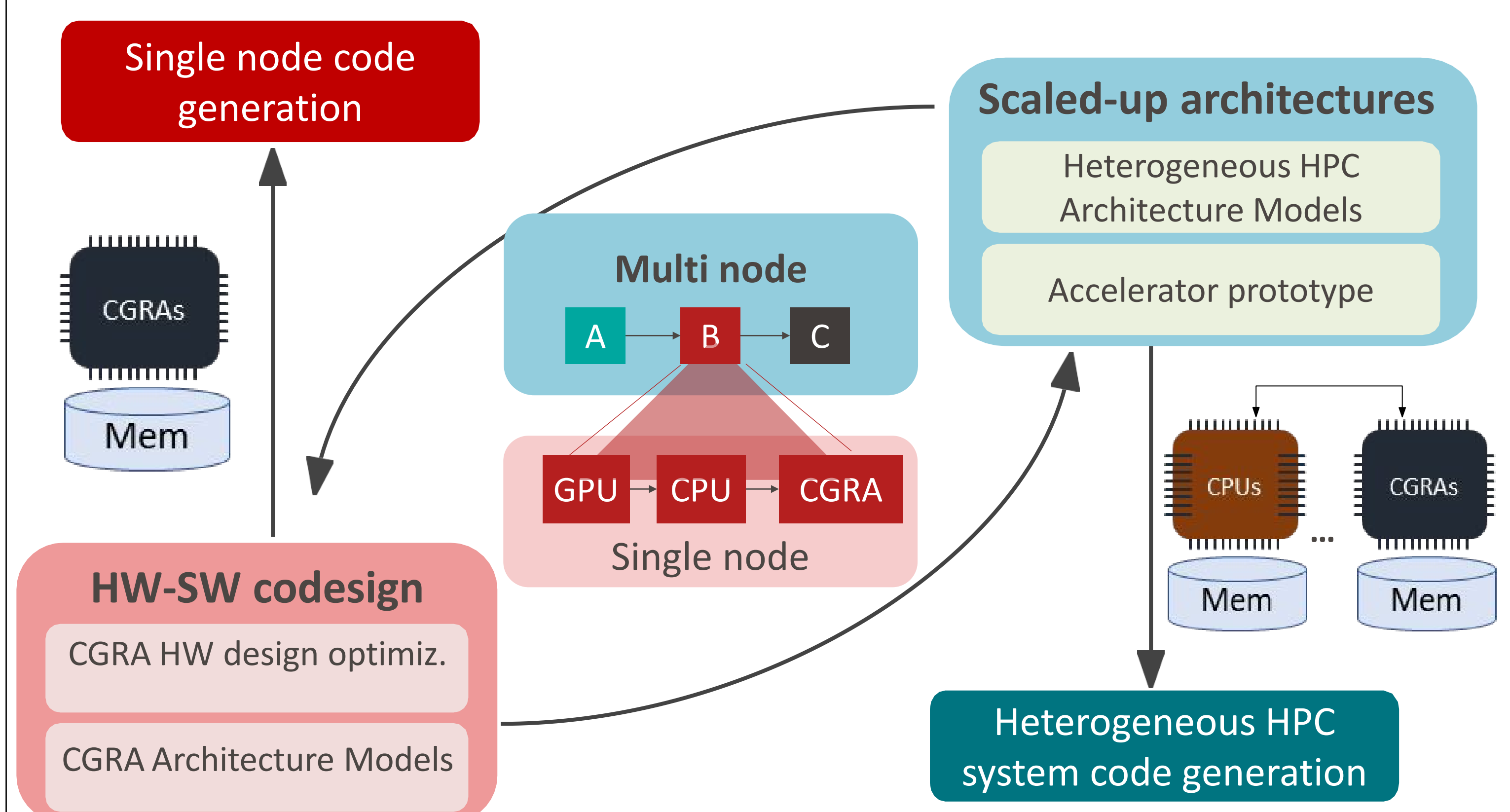
Need Technological Innovations
4x Energy Efficiency!

Sustainable computing + big data + high throughput

Software pipeline not fixed and evolving

Proposal: Multi-Scale Codesign for Sustainable Acceleration

Heterogeneous execution framework for COTS and domain-specific accelerators



Single node - heterogeneous multiprocessing system integration challenge

- New domain-specific accelerator for interferometry computation
- Methods for heterogeneous computing scheduling to integrate custom design accelerator with commercial off-the-shelf accelerators (GPUs, FPGAs, IPUs)
- Parametric architecture: can accommodate different image sizes, accuracy

Multi-node - scalability challenge

- Multi-node scalability of energy efficiency, sustainability, performance & accuracy
- Dynamic workload – many different pipelines
- Highly heterogeneous data center infrastructure
- Extendable multi-node architecture – new types of accelerators

Conclusions

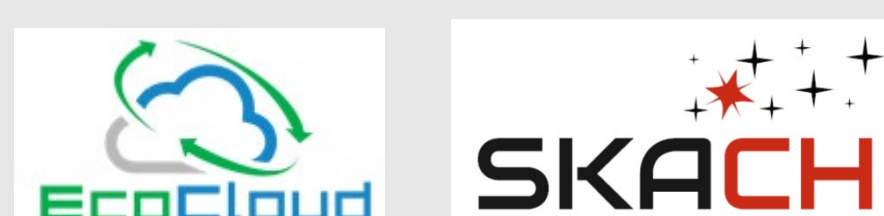
Needs	Business, Research & Innovation Opportunities
Scalable & sustainable HPC for big data	Cloud computing and storage Emissions-aware planning for SKA CSPs, SDPs, and Regional Centres
Real-time requirements	High-throughput technology (I/O, memory, computing)
Old legacy radio-interferometry software	Refactor and scale for modern hardware and HPC
Tradeoffs: programmability, throughput, energy-efficiency	Programmable domain-specific accelerators

Initial seed project between EcoCloud and SKACH

to be extended with new ANR/SNSF project: "SEAMS: Sustainable & Energy Aware Methods for SKA"

Programmable domain-specific accelerators (e.g, CGRA)

Mature, production-level open-source design frameworks
 Hardware IPs: high bandwidth memories, PCIe/CXL
 SwissChip, EuroPractice flagship demonstrators
Heterogeneous execution frameworks

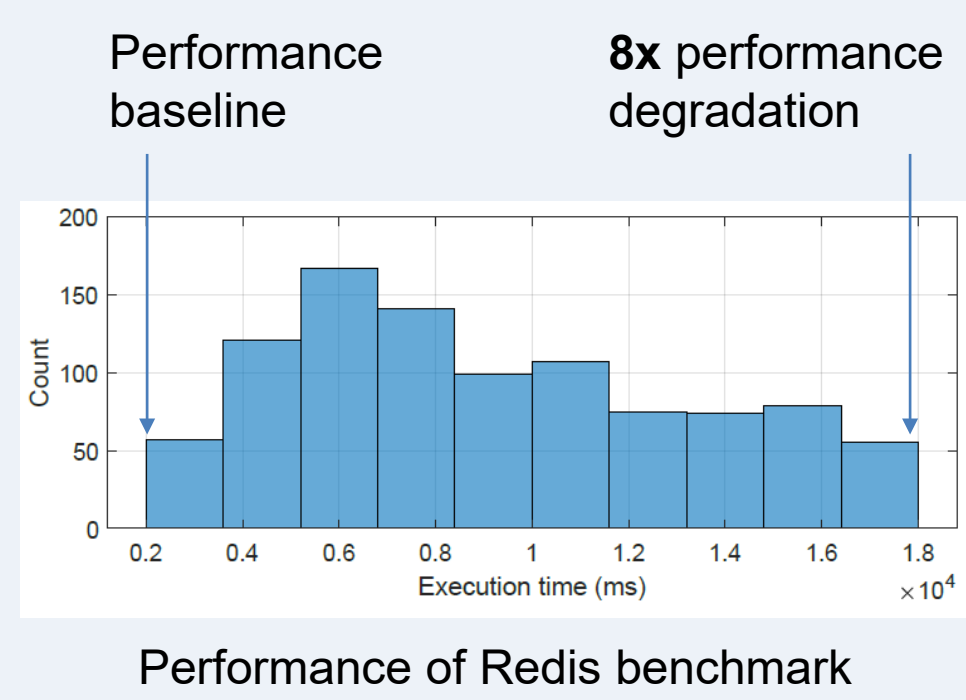
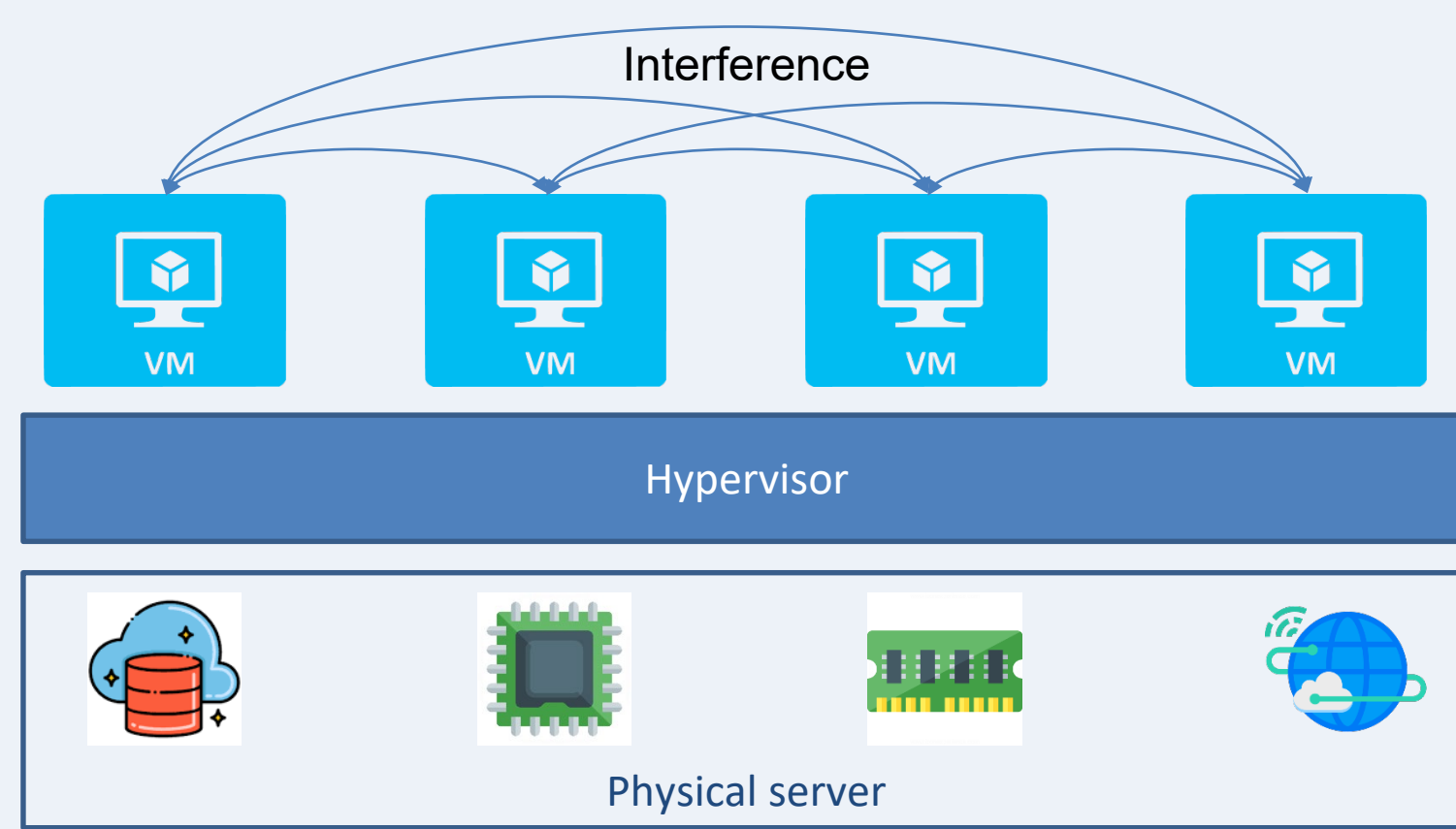


References

1. Swart, G.P., Dewdney, P.E. and Cremonini, A., 2022. **Highlights of the SKA1-Mid telescope architecture**. *Journal of Astronomical Telescopes, Instruments, and Systems*, 8(1), pp.011021-011021.
2. Ernst, N.A., Klein, J., Bartolini, M., Coles, J. and Rees, N., 2023. **Architecting complex, long-lived scientific software**. *Journal of Systems and Software*, p.111732.
3. Gac, N., Nezan, J.F., Ferrari, A., Ferrari, C., Quinson, M. and Dumez-Vioux, C., 2023. **Prototypage rapide d'un supercalculateur dédié à la radioastronomie**.
4. An, T., Wu, X., Lao, B., Guo, S., Xu, Z., Lv, W., Zhang, Y. and Zhang, Z., 2022. Status and progress of China SKA Regional Centre prototype. *Science China Physics, Mechanics & Astronomy*, 65(12), p.129501.
5. Constantinescu, D.A., Denkinger, B.W., Peon Quiros, M. and Atienza Alonso, D., 2023, June. **Tradeoffs in Low-Power Accelerators Design for Large-Scale Interferometers**. In *The Platform for Advanced Scientific Computing (PASC) Conference*.
6. Feng, Wu-chun, and Kirk Cameron. "The green500 list: Encouraging sustainable supercomputing." *Computer* 40.12 (2007): 50-55. www.top500.org/lists/green500/

Introduction

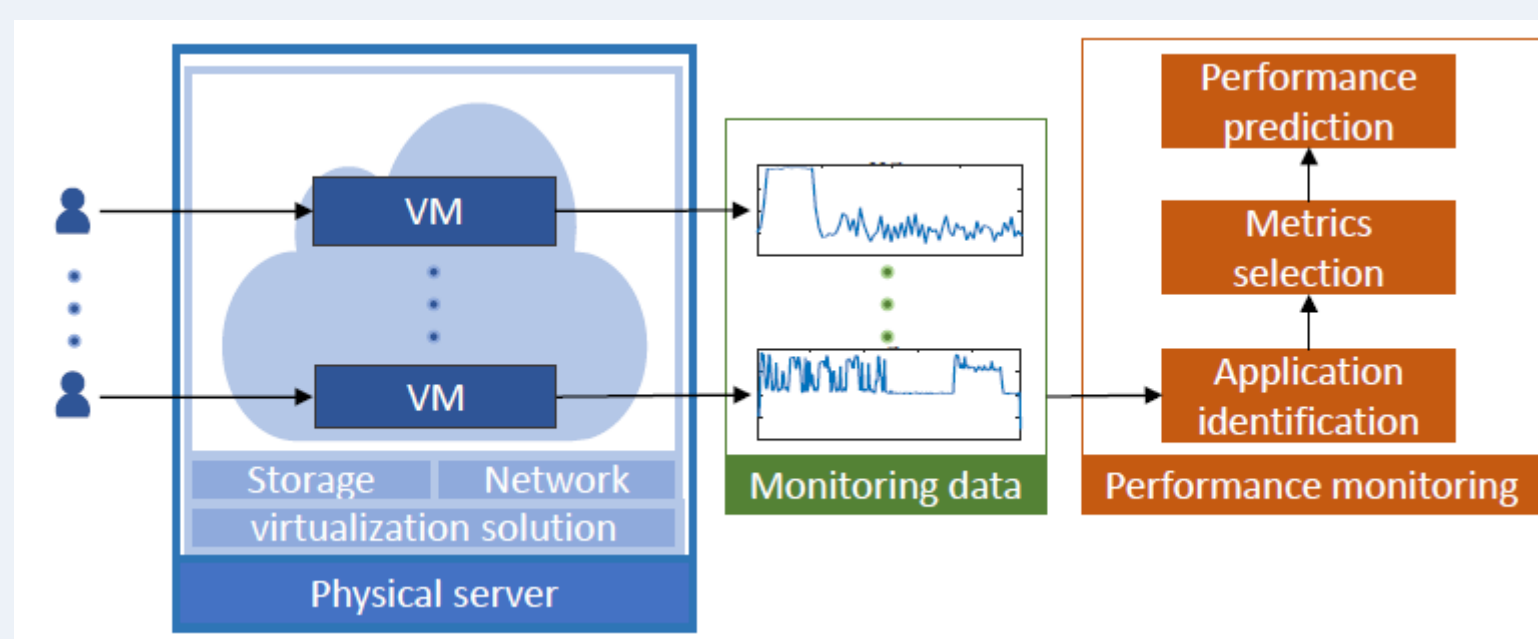
- Infrastructure as a service (IaaS):
 - IaaS contributes to 25% market share in all cloud services
 - IaaS is forecasted to experience the highest end-user spending growth in 2023 at 30.9%
- Performance problem of IaaS:



- Collocated VMs can suffer from severe performance degradation!

Motivation and challenges

- Motivation:
 - Performance prediction of VMs
- Challenges:
 - VM is a black-box
 - Resource contention induced performance variation
 - Workload induced performance variation



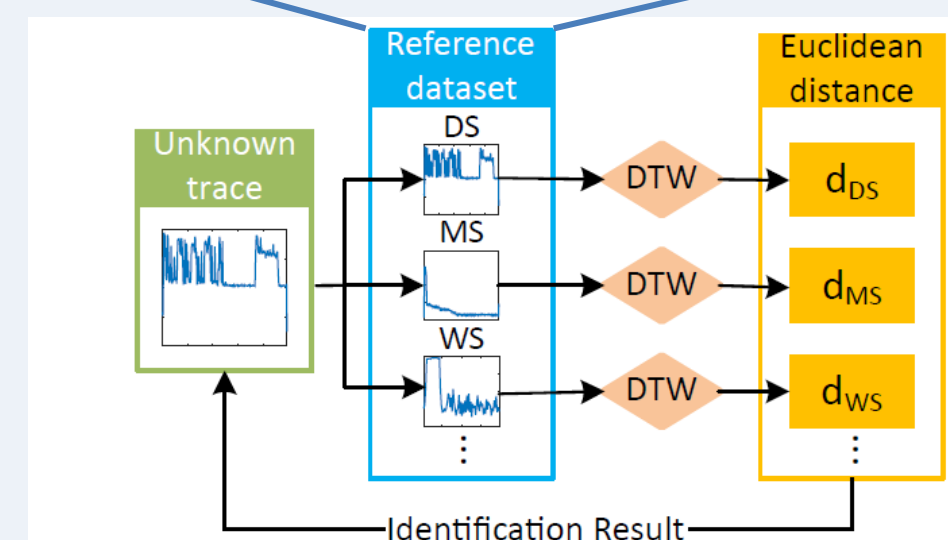
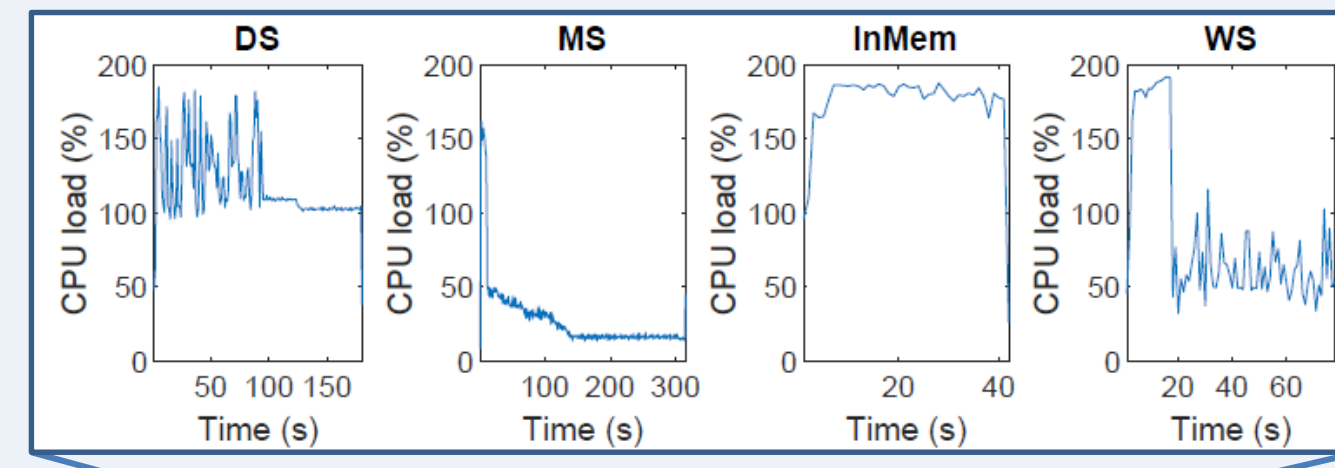
Acknowledgments

This work has been partially supported by the EC H2020 RECIPE FET-HPC project (No. 801137), the ERC Consolidator Grant COMPUSAPIEN (No. 725657), an Industrial Grant from Huawei Cloud to ESL-EPFL, and Grant PID2021-126576NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

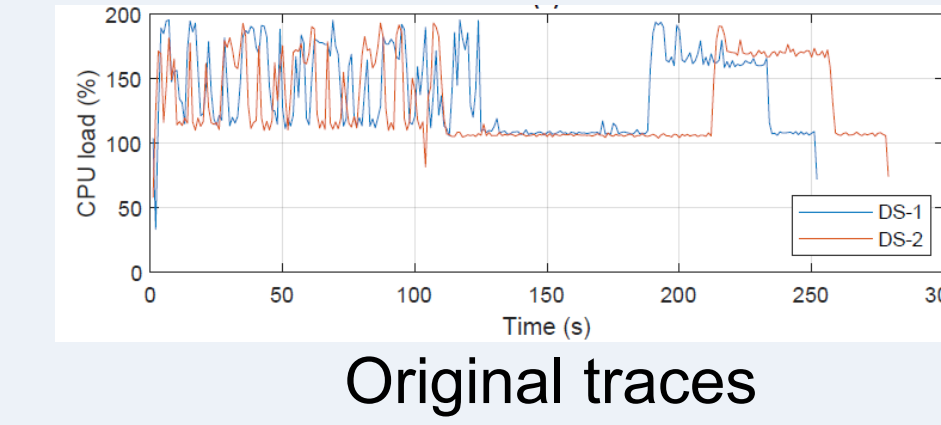


D. Huang, L. Costero, A. Pahlevan, et al. CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds arXiv preprint arXiv: 2309.16333

Application Identification

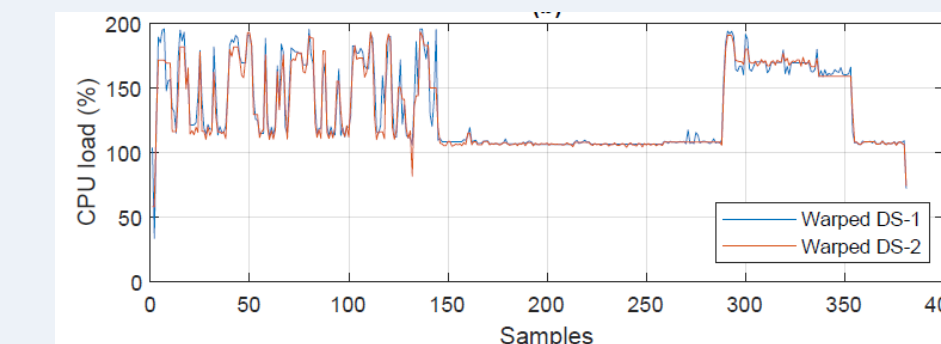


- Offline: Create the reference dataset
- Online: Dynamic time warping (DTW) - based identification



Original traces

- Low prediction accuracy on original traces
- Misprediction due to large distance



Warped traces

- Warped traces are minimized in the distance
- For better identification accuracy

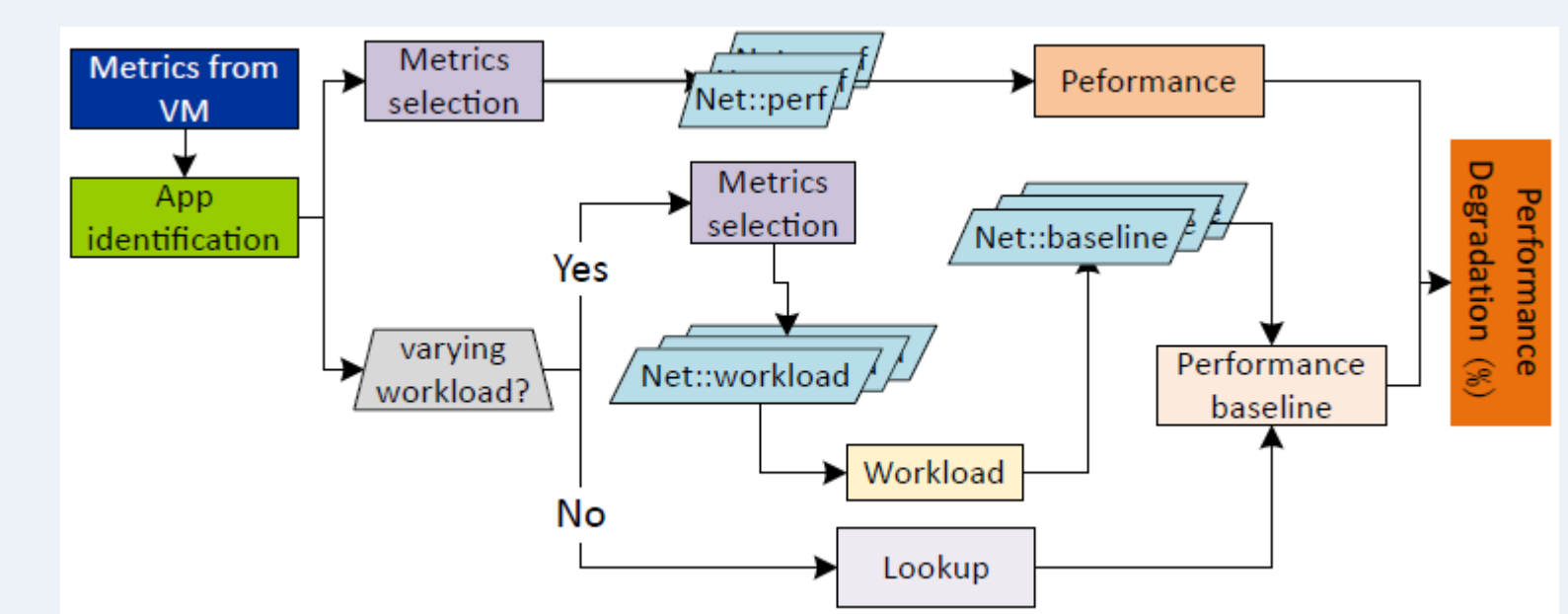
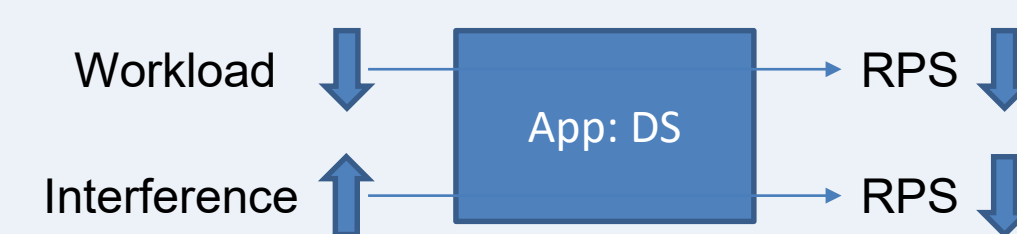
$$d(m, n) = \sqrt{\sum_{i \in t} (m_i - n_i)^2}$$

$$d_{\min}(m, n) = \min_{\pi \in A(m, n)} \left(\sum_{(i, j) \in \pi} \sqrt{(m_i - n_j)^2} \right)$$

Dynamic time warping

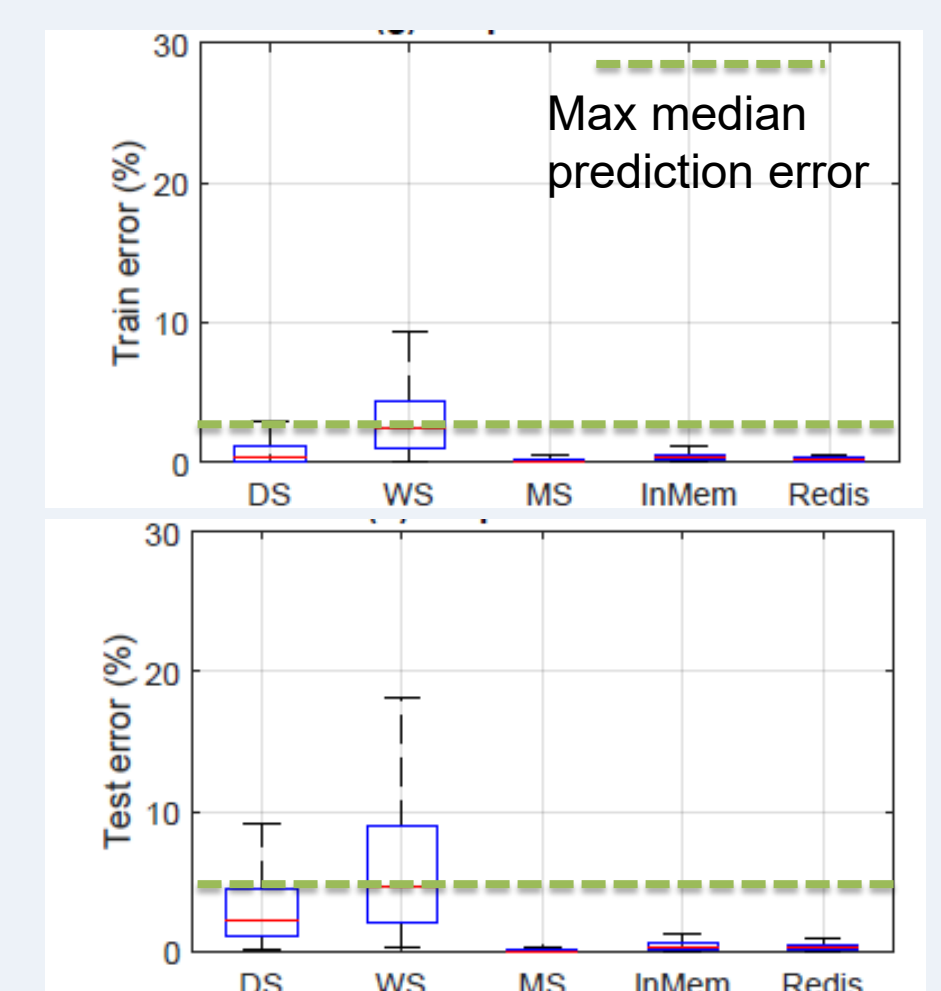
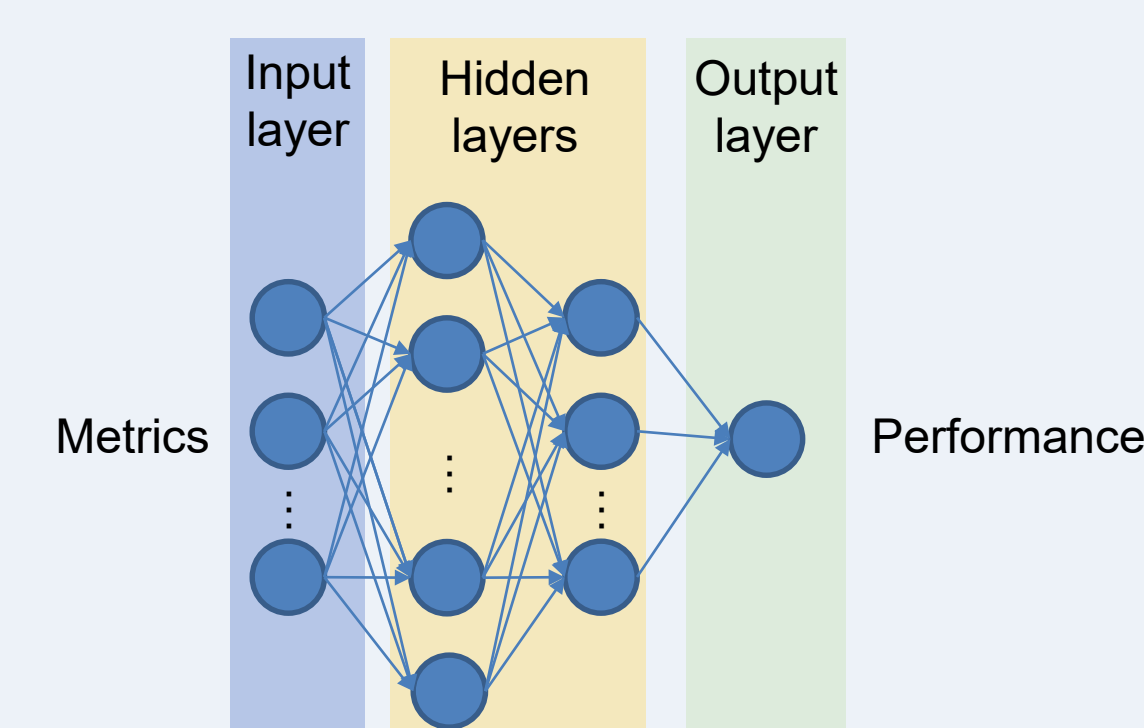
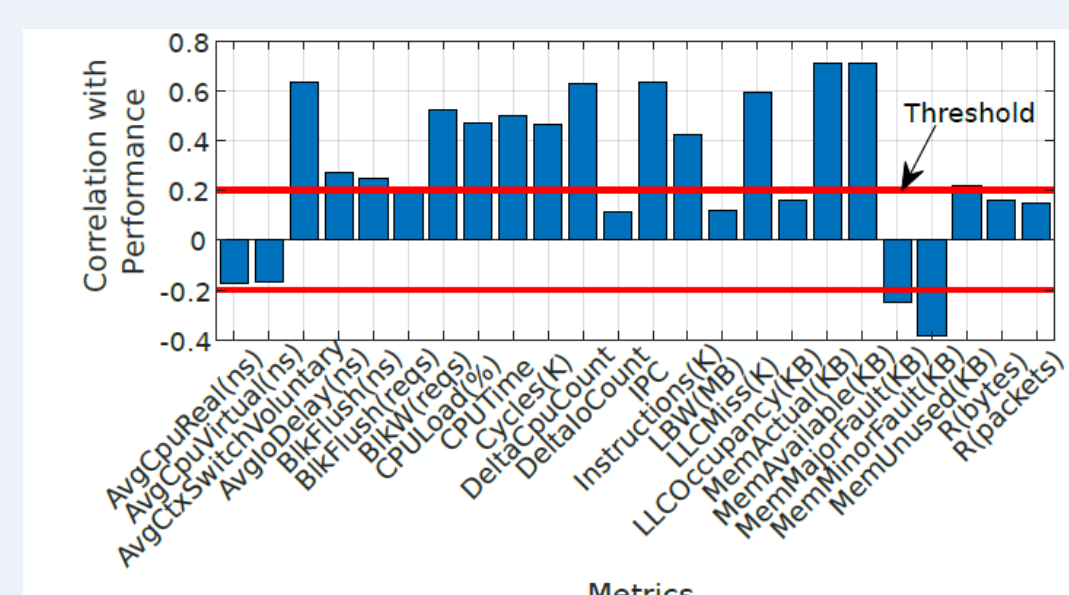
Performance degradation prediction

- **Problem:** Both workload and interference influence the performance level of the application!
- **Solution:** Alongside the **performance prediction**, two additional NNs are used for **workload prediction** and **performance baseline prediction**



I. Performance prediction

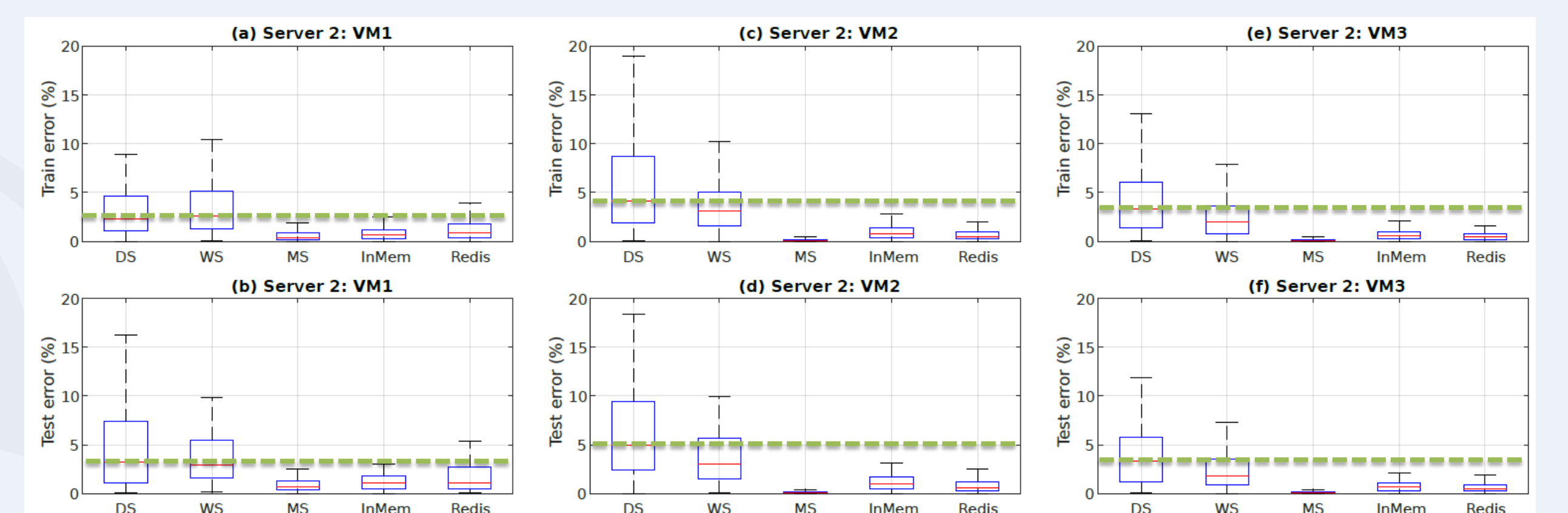
- Highly correlated metrics are selected for the prediction
- Proposed method achieves a mean error below 5% for all applications!



II. Performance degradation prediction

- NN-based performance degradation prediction method achieves a mean error below 7% for all applications
- Moreover, the proposed method can work with different server and VM configurations

App	Train error (%)			Test error (%)		
	mean	max	std	mean	max	std
DS	4.5	44.7	5.6	5.4	26.0	5.2
WS	4.6	27.0	4.3	6.8	32.1	6.1
MS	0.5	7.6	0.9	1.2	16.4	2.9
InMem	0.4	3.3	0.4	0.6	7.4	0.8
Redis	0.2	1.3	0.2	0.3	1.7	0.3



Server	CPU	Memory	OS
S1	2x Intel Xeon E5-2620 v3 @ 2.4GHz	128GB ECC	EulerOS 2.0
S2	2x Intel Xeon Gold 6242R @ 3.1GHz	384GB ECC	CentOS 7.9

VM name	CPU	Memory	OS
VM1	2x vCPU	4GB	Ubuntu 20.04
VM2	4x vCPU	4GB	Ubuntu 20.04
VM3	4x vCPU	8GB	Ubuntu 20.04

Summary

- Black box VM assumption
- Accurate performance prediction (7% mean error)
- DTW-based app. identification
- Support different server and VM configurations

Atli Kosson & Martin Jaggi
Machine Learning & Optimization Laboratory

Motivation

Problem: Neural network training is computationally expensive with the multiply-and-accumulate operations performed in matrix multiplications accounting for most of this cost. Compared to addition, multiplication is significantly more demanding in terms of hardware area and energy and is ultimately the primary contributor to the overall computational burden.

Questions:

- Can we replace matrix multiplications with theoretically cheaper piecewise-affine alternatives?
- Can neural networks be trained in a completely “multiplication-free” manner by extending this approach to other operations (see figure →)?

Background: Floating Point Numbers

- Binary equivalent of the scientific notation:
 $(-1)^S \cdot 2^E \cdot (1 + M)$ with $S \in \{0,1\}$, $E \in \mathbb{Z}$, $0 \leq M < 1$
- The bit representation of this using IEEE 754 float32 encoding is:
 $[S, e_0, \dots, e_7, m_0, \dots, m_{22}]$
with $\bar{E} = (e_0 \dots e_7)$, $E = \bar{E} - 127$, $\bar{M} = (m_0 \dots m_{22})$, $M = \bar{M}/2^{23}$
- Special interpretations for $\bar{E} = 0$ and $\bar{E} = 255$ give 0, NaN and $\pm\infty$

Piecewise-Affine Multiplication

Definition:

Let $A = (-1)^{S_A} \cdot 2^{E_A} \cdot (1 + M_A)$ and $B = (-1)^{S_B} \cdot 2^{E_B} \cdot (1 + M_B)$
Piecewise-Affine Multiplication (PAM) is then defined as $C = A \hat{B}$ with:

$$C = (-1)^{S_C} \cdot 2^{E_C} \cdot (1 + M_C)$$

$$S_C := (S_A + S_B) \bmod 2$$

$$E_C := E_A + E_B + 1\{M_A + M_B \geq 1\}$$

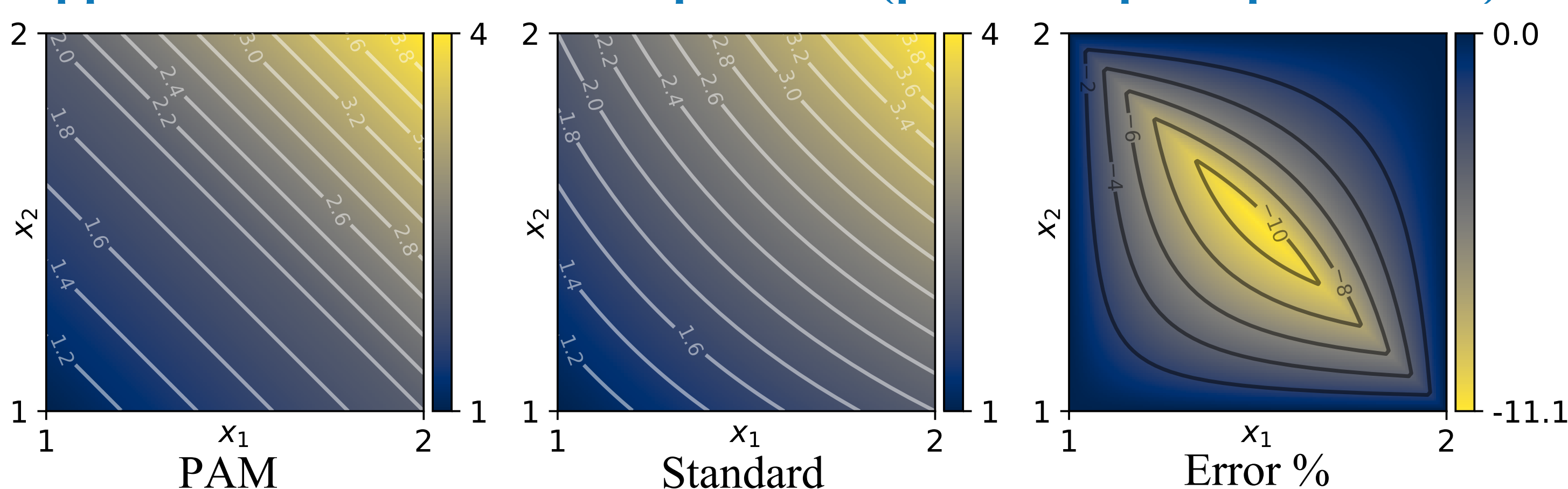
$$M_C := M_A + M_B - 1\{M_A + M_B \geq 1\}$$

where $1\{a\}$ is one when a holds and zero otherwise.

Implementation:

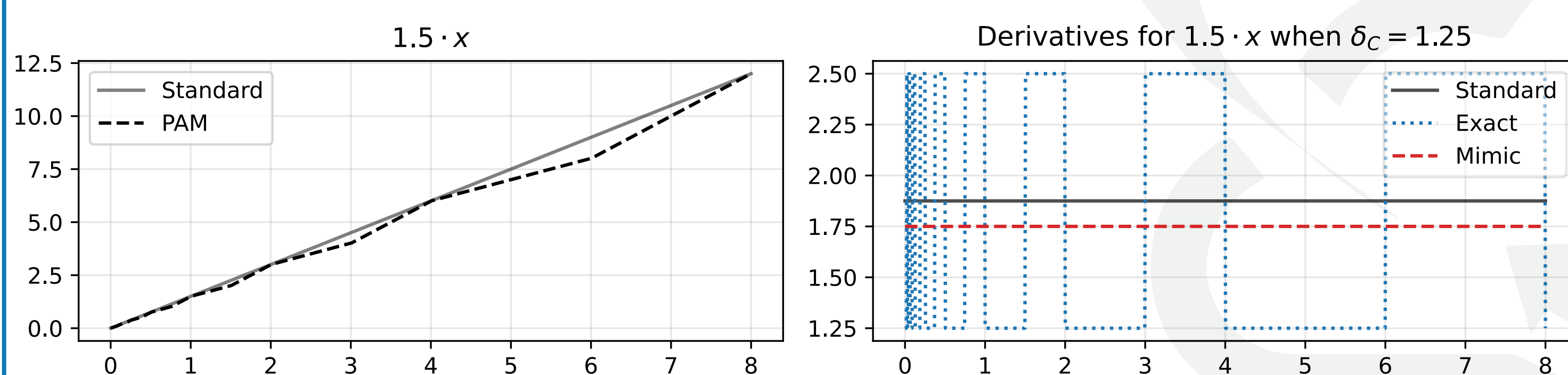
Adding the binary representations together as int32 which allows $\bar{M}$ to overflow into $\bar{E}$. Then correcting for the exponent offset of -127 and handling edge cases including exponent over/underflows and NaN/Inf.

Approximates standard multiplication (pattern repeats per binade):



Two types of derivatives:

- Exact bwd:** $\delta_A = 2^{E_B+1\{M_A+M_B \geq 1\}} \delta_C$, the true derivative
- Mimic bwd:** $\delta_A = B \hat{\delta}_C$, uses PAM in standard derivative formula
- Both can be performed without standard multiplications.
- The mimic derivative is biased but has less variance.



Other Piecewise-Affine Operations

- We can construct other piecewise-affine operations e.g., $\log/\exp$
 $\text{paexp}_2(A) = 2^{\lfloor A \rfloor} \cdot (1 + A - \lfloor A \rfloor)$, $\text{palog}_2(A) = E_A + M_A$, $A > 0$
- Operations like $\sqrt{x}$, $\exp$ etc. can be defined in terms of these
- Also have two types of derivatives, exact and a mimic version based on the formula for the derivative of the standard analog

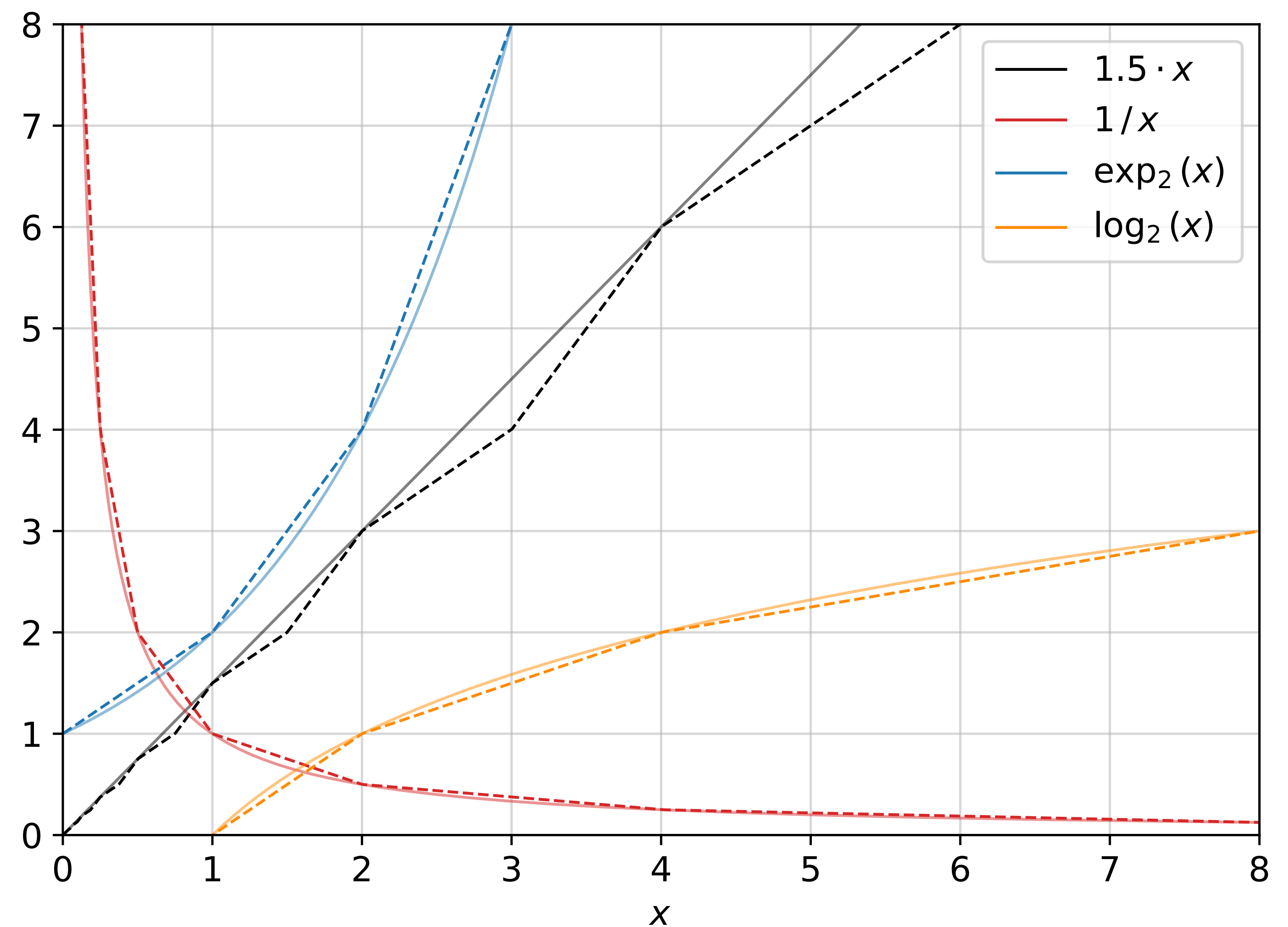


Figure: Elementary neural network operations (solid lines) along with cheaper piecewise affine alternatives (dashed lines). The piecewise affine functions periodically match the originals, while the slope of each segment is an exact power of 2.

Experiments

Piecewise-Affine Matrix Multiplications

	DATASET	BASELINE	Alternative method	
			PA-MATMUL	ADDER
DeiT-Tiny	CIFAR10	94.8%	95.0[+0.2]%	92.4[-0.2]%*
	IMAGENET	72.2%	72.2[+0.0]%	70.5[-1.7]%
ConvNets (CIFAR-10)	NETWORK	BASLINE	PA-MATMUL	
	VGG-13	92.9 ± 0.3%	92.9[+0.0] ± 0.2%	
	RESNET-20	92.1 ± 0.3%	92.0[-0.1] ± 0.2%	
	RESNET-110	94.2 ± 0.1%	94.1[-0.1] ± 0.2%	
	RESNEXT-20 4x16	93.7 ± 0.2%	93.8[+0.1] ± 0.2%	
	CONVMIXER-256/8	94.8 ± 0.2%	94.8[-0.0] ± 0.1%	

Narrow Mantissas

MATMUL TYPE	VGG-13	Transformer-Small
	CIFAR-10 TEST ACCURACY	IWSLT14 BLEU SCORE
FLOAT32	92.9 ± 0.3%	34.4 ± 0.1
PAM FLOAT32	92.9 ± 0.2%	34.2 ± 0.2
PAM BFLOAT	92.9 ± 0.2%	34.4 ± 0.2
PAM 4 BIT MANTISSA	92.9 ± 0.5%	34.2 ± 0.1
PAM 3 BIT MANTISSA	92.8 ± 0.2%	29.4 ± 0.5

Fully Multiplication-Free Training

PA OPERATION(S)	IWSLT14 DE-EN BLEU Score		
	EXACT BWD	MIMIC BWD	CUMULATIVE
MATMUL	33.4[-1.0] ± 0.1	34.2[-0.2] ± 0.2	34.2[-0.2] ± 0.2
ATTENTION SOFTMAX	3.0[-31.4] ± 0.9	34.3[-0.1] ± 0.7	34.2[-0.2] ± 0.2
LAYER NORM	33.7[-0.7] ± 0.2	34.4[-0.0] ± 0.1	34.4[-0.0] ± 0.1
LOSS	33.7[-0.7] ± 0.2	33.0[-1.4] ± 0.9	33.7[-0.7] ± 0.2
OPTIMIZER	34.2[-0.2] ± 0.1		33.5[-0.9] ± 0.3

Conclusion: Able to match the baselines with piecewise-affine matrix multiplications. Only minor degradation replacing all multiplications.

PAM Hardware Cost

FORMAT	ADDITION		MULTIPLICATION	
	ENERGY [pJ]	AREA [μm ²]	ENERGY [pJ]	AREA [μm ²]
INT32	0.1	137	3.1	3495
INT16	0.05	67		
INT8	0.03	36	0.2	282
FLOAT32	0.9	4184	3.7	7700
FLOAT16	0.4	1360	1.1	1640

PAM vs standard multiplication for float32:

- ~20x energy savings, ~30x area savings
- Special hardware required to realize gains



arXiv:2305.17190

POLLUX and CASTOR: Enforcing Separation for Mixed-Language Environments

Andrés Sánchez, Flavio Toffalini, Trent Jaeger, Mathias Payer

The menace of unsafe external libraries in safe code

```
extern "C" { fn f_c(addr_c: i64); }
fn f_rust() {
  let vals: [i64;3] = [1,2,3];
  let cb: fn(&mut i64) = fp1; // Function pointer
  let addr_r = &vals as *const i64 as i64;
  unsafe{ f_c(addr_r) }
  (cb)(&mut vals[0]);
}
```

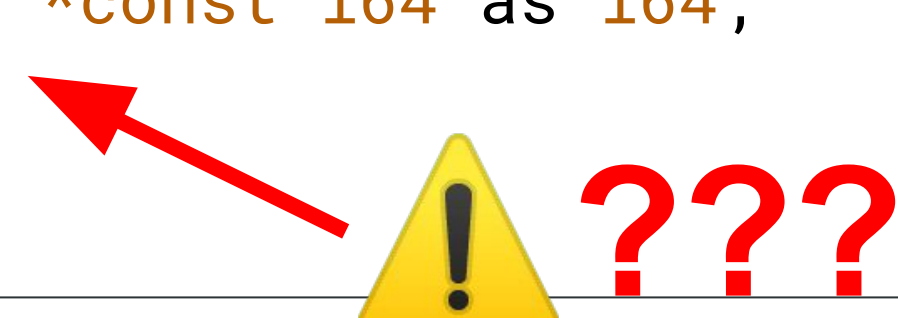


Fig 1. Call from Rust to external C function

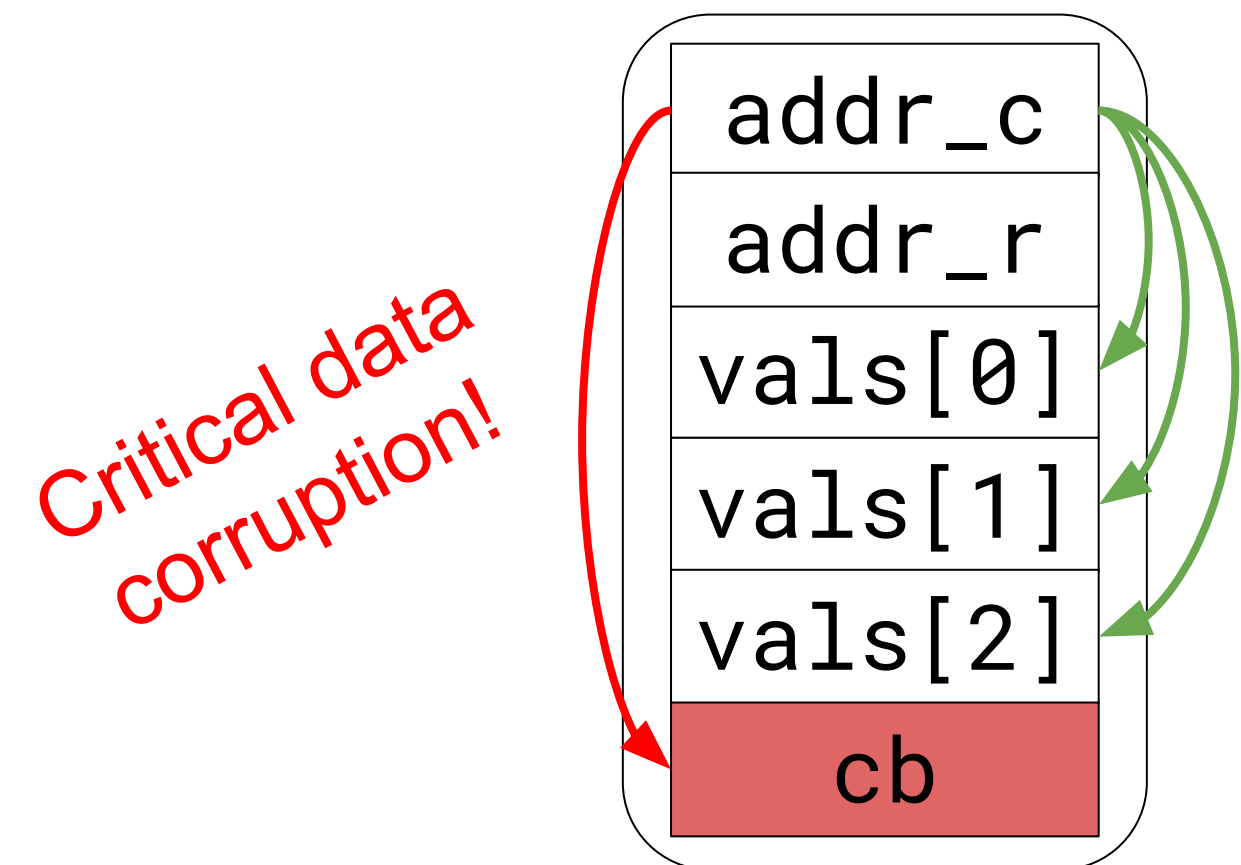
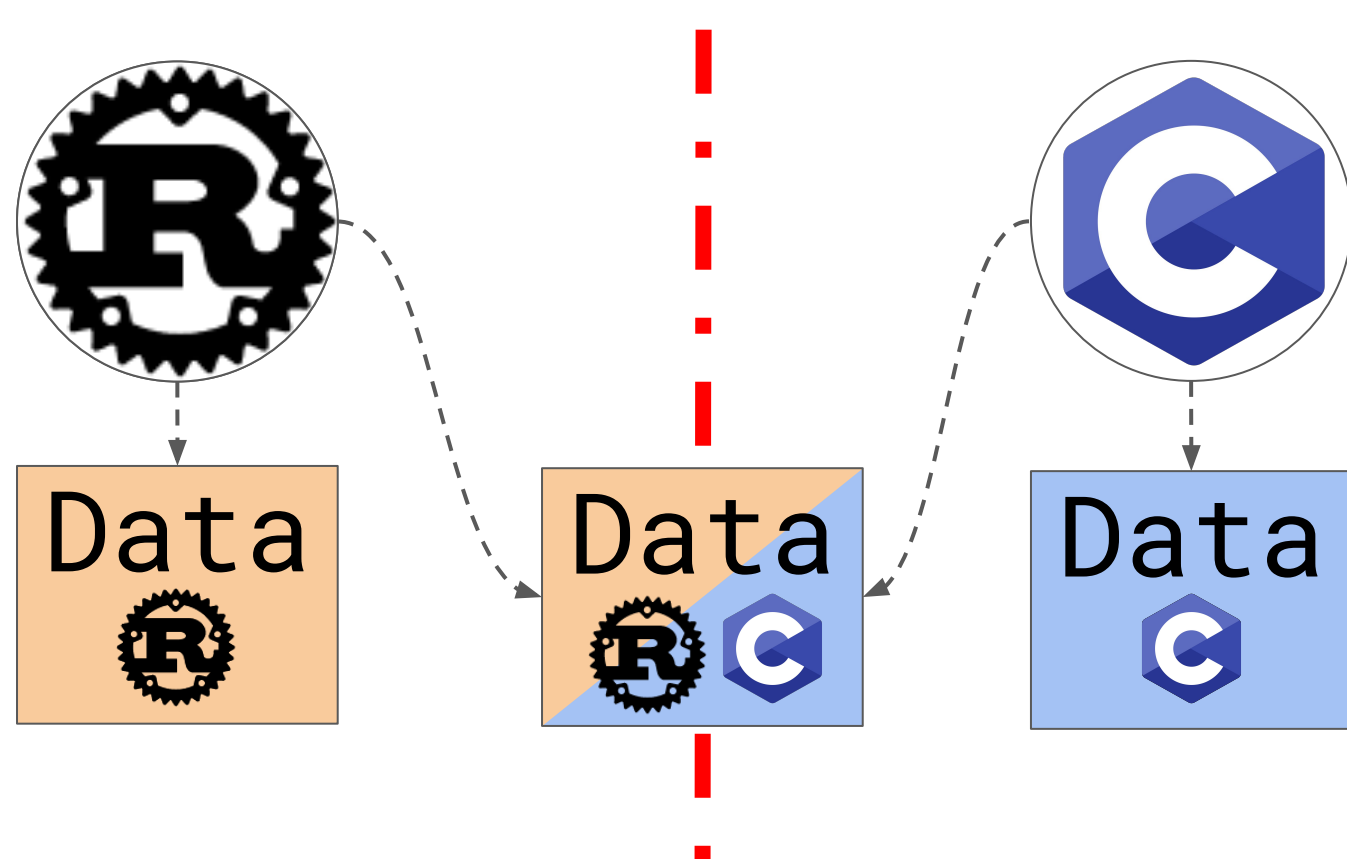


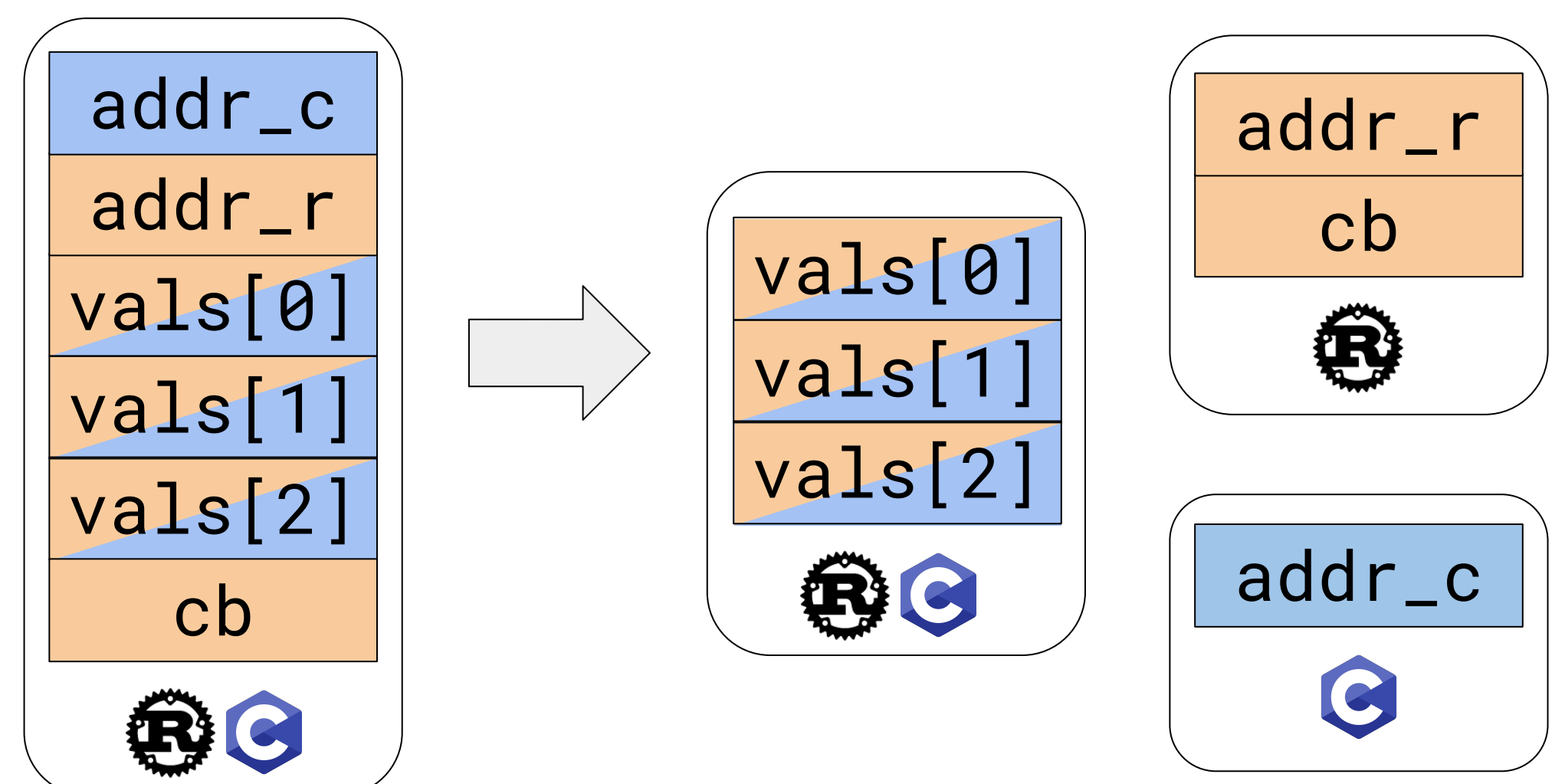
Fig 2. Stack-based **cb** corruption from C

Apply compartmentalization between languages data

Data access is restricted to languages
(e.g., data shared between Rust and C)



Stack address space is split in
two private stacks and a shared stack



Evaluation

14% (180) Rust projects call external code

8% (102) pass references,
need manual intervention

6% (78) protected
without intervention

2.2% (29) compile
custom C code

2% (26) compile
custom C code

41.1% of stack allocations tagged as shared
40.3% of shared stack space
5.8% runtime overhead

Impact

- Compiler warns about unsafe patterns
- Developer informed about critical shared data (e.g., code pointers)
- No need to analyze unsafe code
- Leverages existing hardware isolation extensions (e.g., MPK, MTE, TEEs)
- Applies to other languages than C/Rust
- Challenge: performance cost due to frequent inter-language transitions

Georgios Sarantakos, Enea Figini, Prof. Dr. Mario Paolone
Head of the Distributed Electrical Systems Laboratory

MOTIVATION

Data centers (DCs) use **2% of global energy** and will reach **10% by 2030**.
Switzerland is one of the European countries with the **highest number of DCs (77)**.
IT and energy supply represents today **44% of the EPFL CO₂eq emissions**.
The DCs expansion is associated with key energy and climate-related challenges:

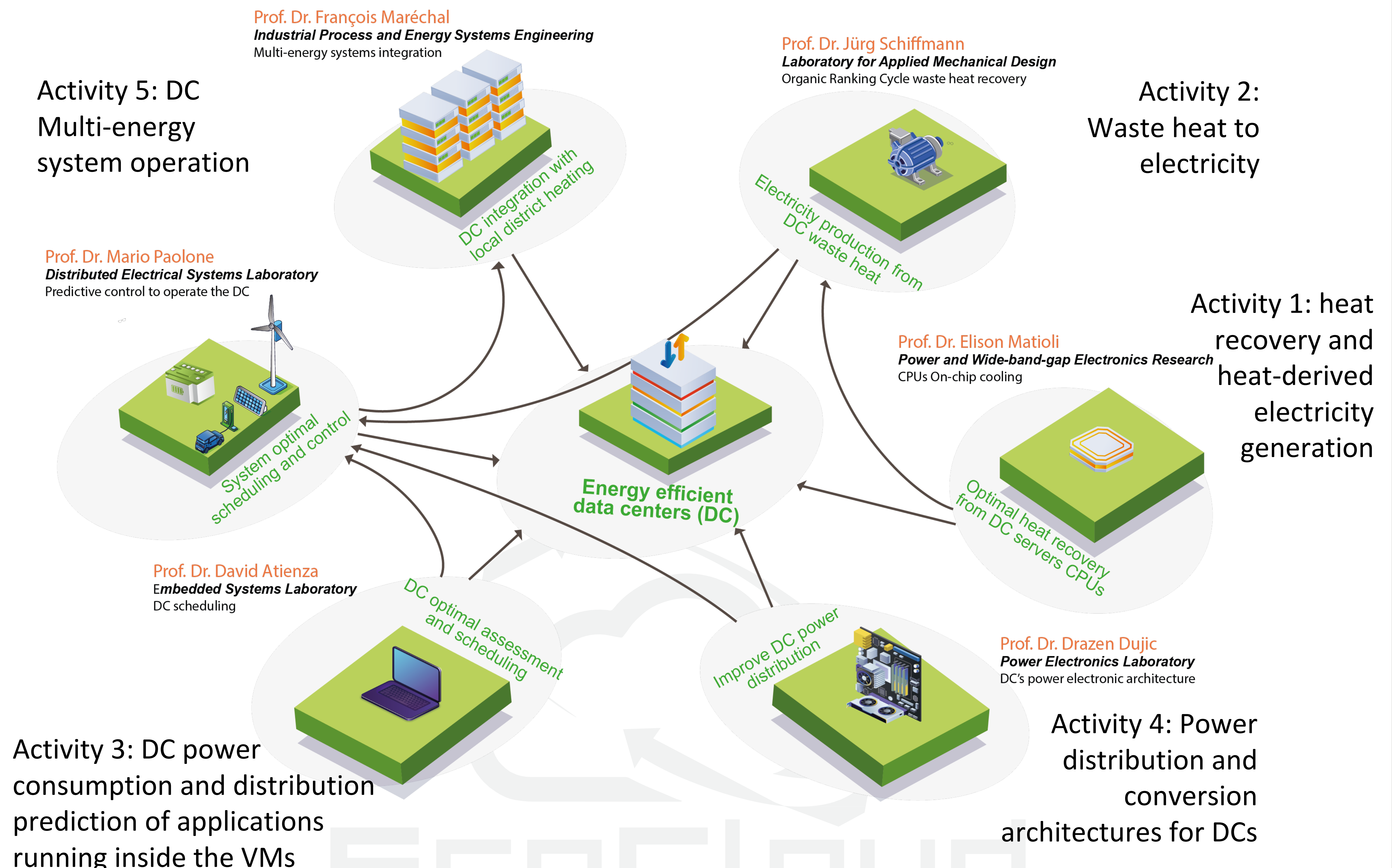
1. impact on the power grid;
2. inefficient energy usage;
3. electricity-end-use carbon footprint.

OBJECTIVES

- Develop a solution to **improve DCs' energy efficiency and reducing their carbon footprint** with special focus on the new EPFL DC
- Validate this solution in a **first-of-its-kind demonstrator** benefiting the entire EPFL campus leveraging EPFL research and tech transfer towards our startups.
- Scale up through the EPFL EcoCloud to industry-grade DCs operated by Amazon, Google, Microsoft, Meta.

ACTIVITIES

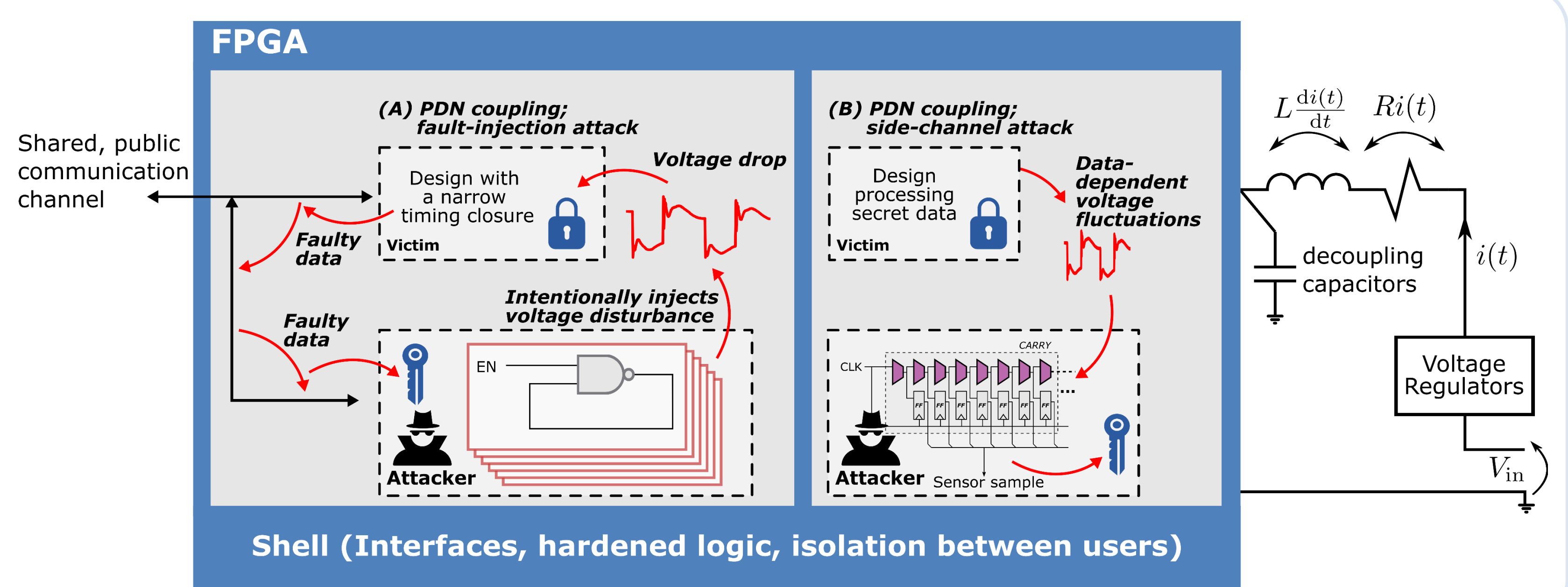
The project involves 6 EPFL research groups undertaking 5 interrelated activities, each of them contributing to the improvement of the energy efficiency of the data centers.
Particularly a predictive control layer will be developed that allows to operate the DC with the least carbon footprint by leveraging different control variable, namely Data center workload (up to 50Kw), BESS (xxxx), forecasted PV generation, ORC (50 kW).



ACKNOWLEDGEMENT

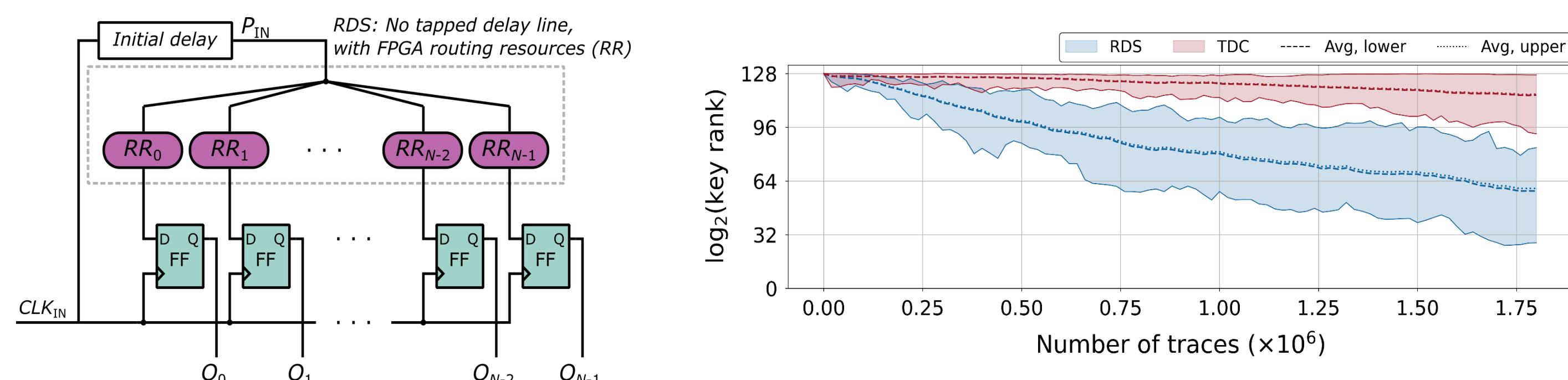
The project is financed by the EPFL Solutions 4 Sustainability initiative

- Cloud providers (AWS, Azure, Alibaba) deploy FPGAs in servers
- FPGAs allow access to **low-level logic** (LUTs, FFs)
 - To build voltage-drop sensors (time-to-digital converters (TDCs))
 - To build power wasters
- Remote FPGA access creates security risks
 - Remote **power side-channel** attacks (SCAs)
 - Remote **fault injection** (FI) attacks
- In a multitenant user scenario
 - Tenants **logically** and **physically** separated
 - Connected through the power delivery network (PDN)



Routing Delay Sensors (RDS) [TCHES'23]

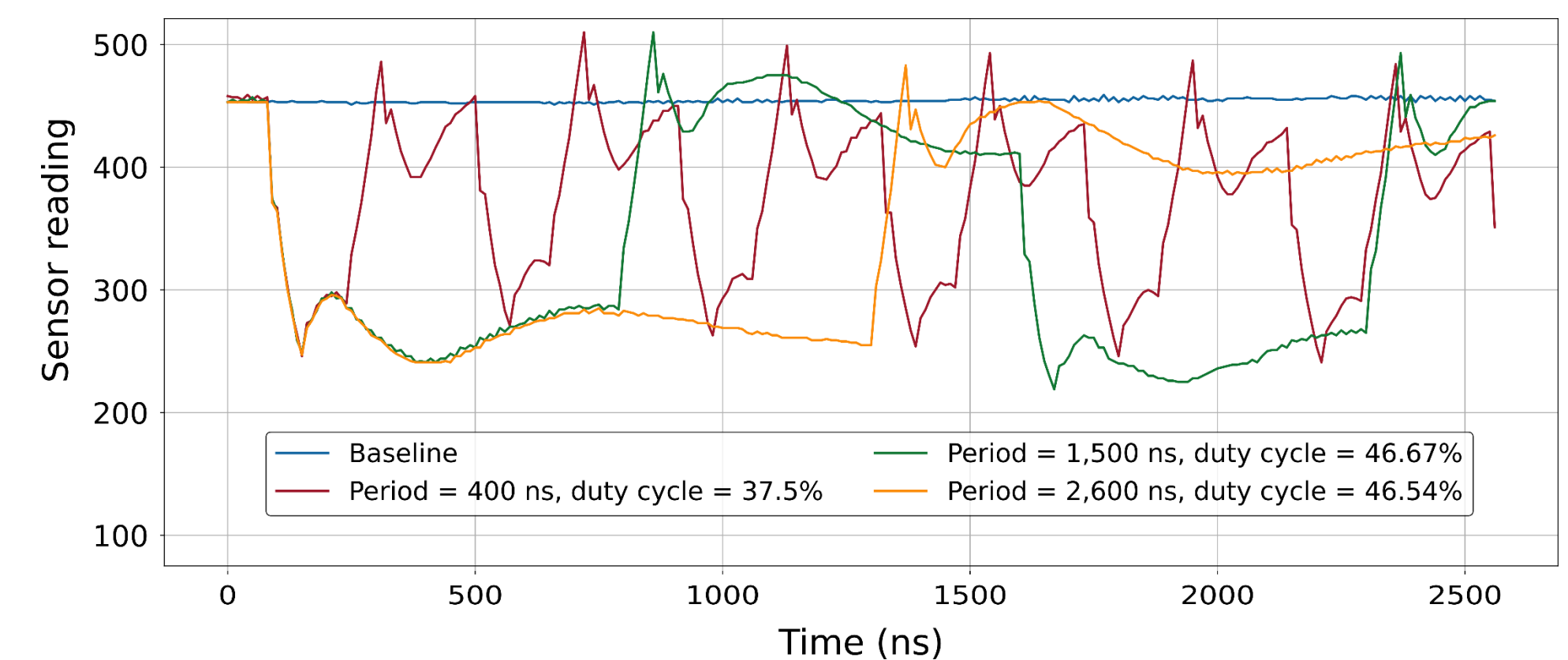
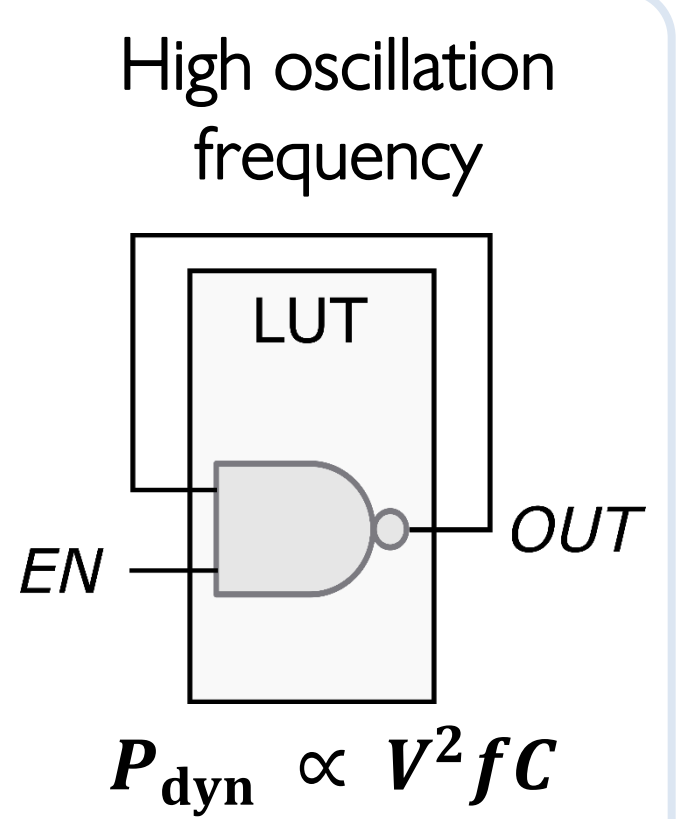
- TDC sensors use CARRY elements detectable by bitstream scanners
- We present the first voltage-drop sensor that leverages routing resources
 - The observable delay line replaced with routing for a stealthier sensor
- RDS also **outperforms the state-of-the-art** TDC
 - 4× larger peak-to-peak ratio and 2× larger SNR
 - Breaks 3× more key bits on a datacenter-scale FPGAs (Alveo U200)



Stealthy routing-based voltage sensor, outperforming state-of-the-art TDCs

Fault Injection Attacker [DATE'22]

- Remote fault injection relies on violating timing constraints
 - Undervolting increases the delays in the circuit
 - High power consumption $\rightarrow$ undervolting
- Limit the duration to avoid denial-of-service
- Divide the attacker into independently-controlled blocks
- Carefully choose the period and the duty cycle of the enable signals

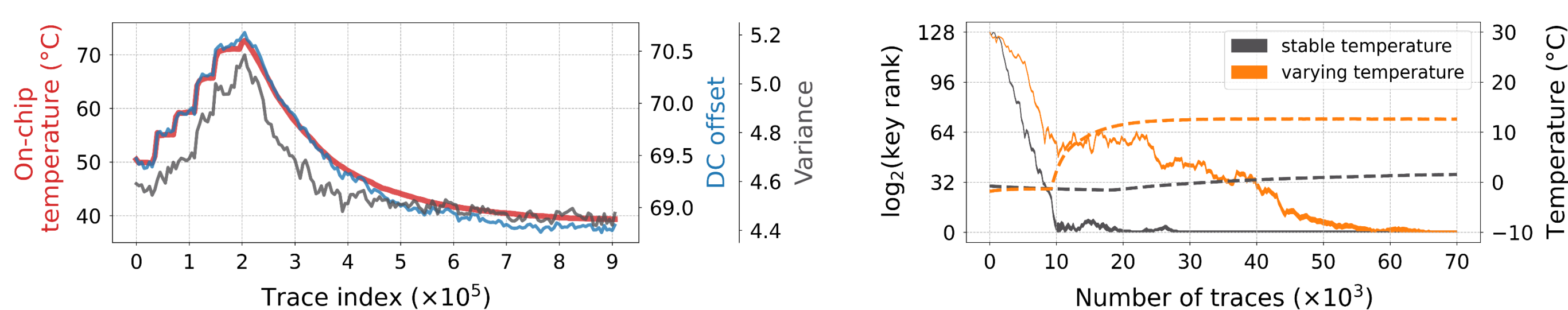


217 blocks of 124 LUTs each

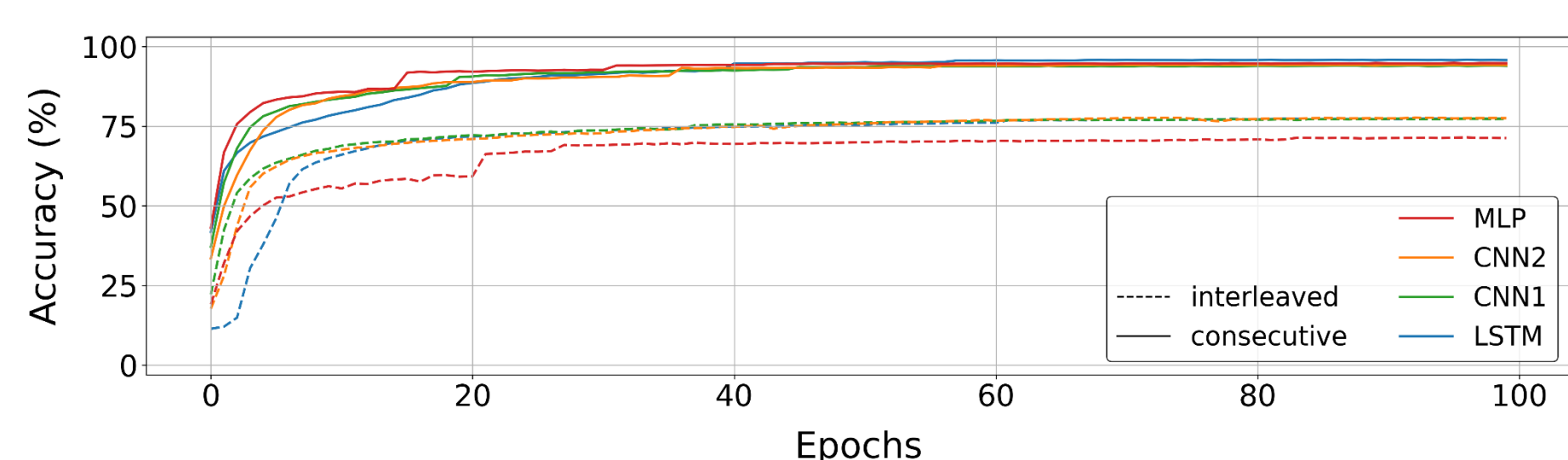
Shape and amplitude of voltage drop controlled by activation pattern of attacker

Temperature Impact on Remote Power Analysis Attacks [DATE'23]

- On shared FPGAs, we present the first analysis of temperature impact on voltage-drop sensors and statistical and ML-based power analysis attacks
- TDC sensors are impacted by temperature, affecting power SCA results



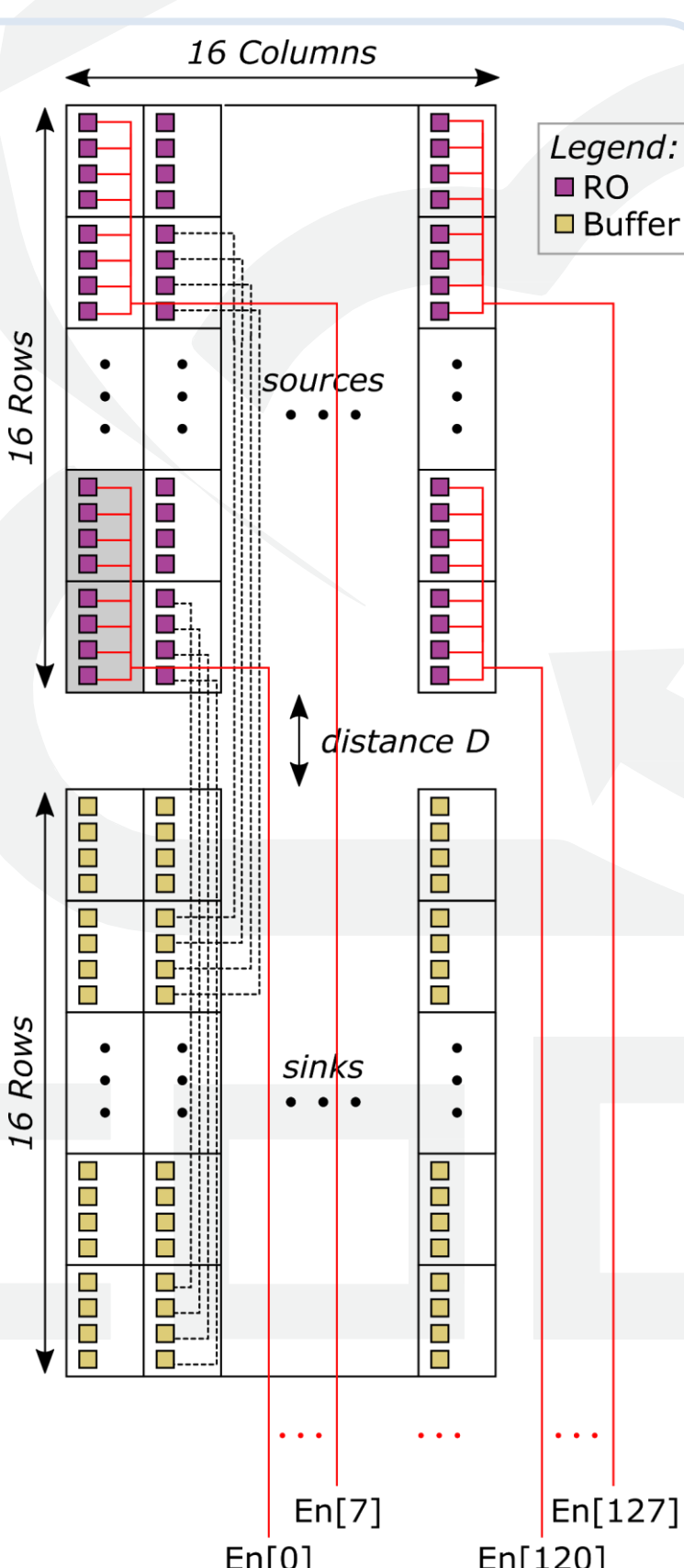
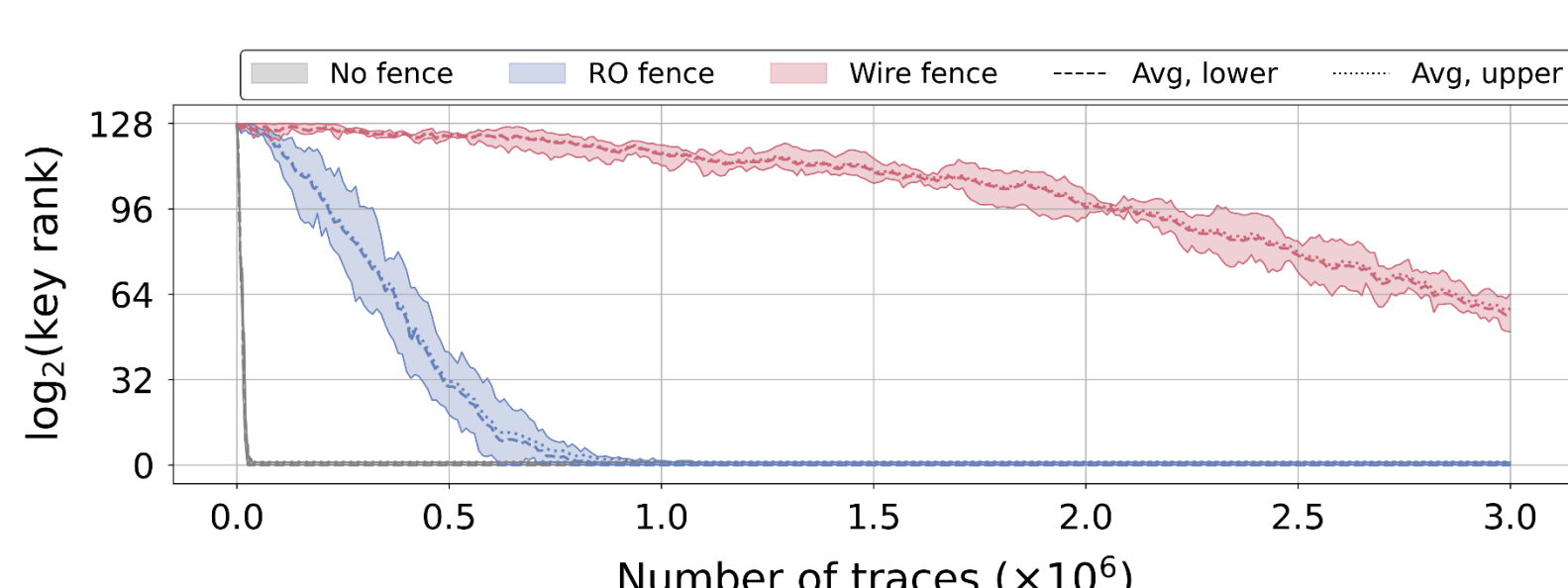
- We quantify the impact of trace acquisition techniques on ML attacks
 - Correct trace acquisition is important even in shared FPGAs
 - Consecutively recorded traces can show a **misleadingly high accuracy**



First analysis of the temperature impact on remote power analysis attacks

Active Wire Fences [DDECS'23]

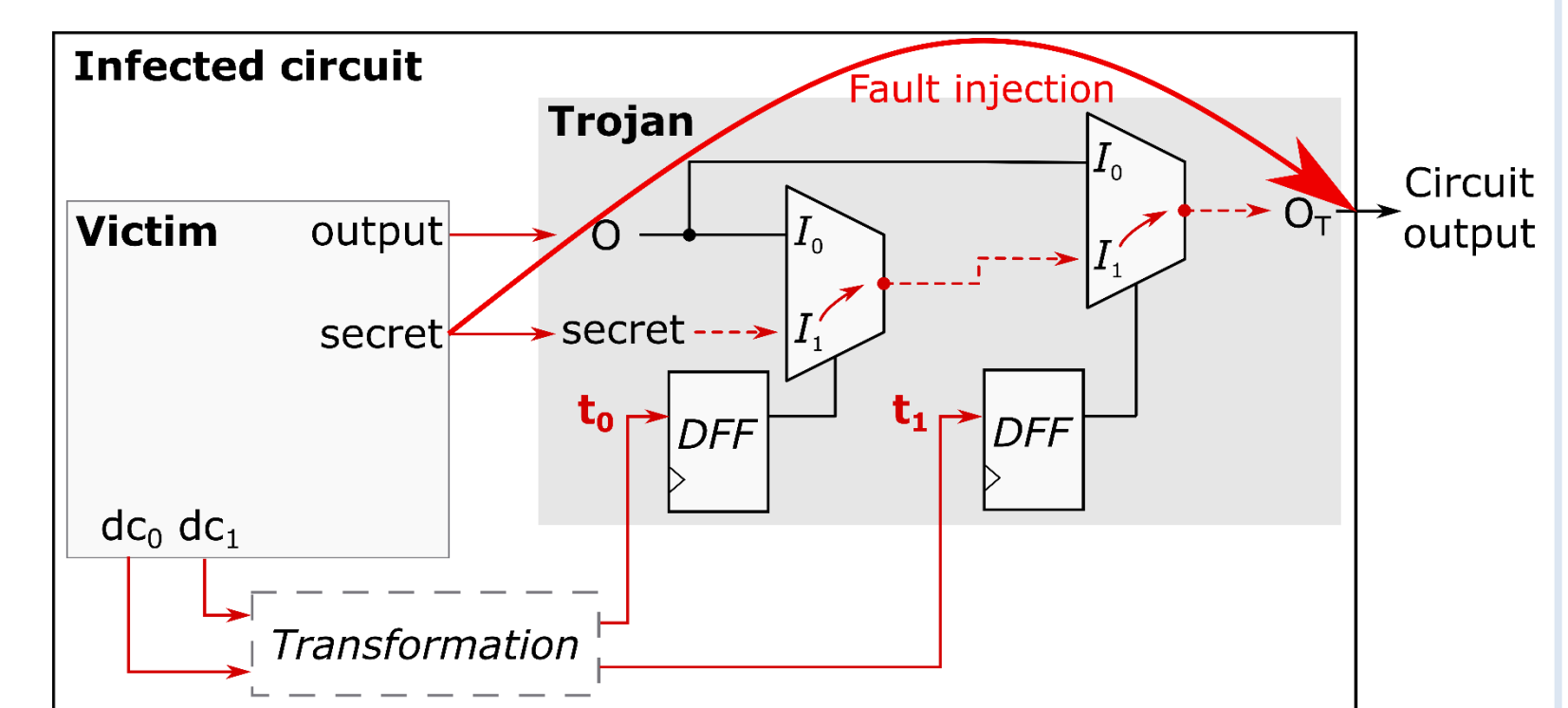
- We propose novel wire-based power wasters
 - Sources (ROs) drive wires ending in sinks (buffers)
- Wire-based wasters consume more power than ROs
 - 2× voltage drop without extra resource overhead
- Active wire fences **protect better** from power SCAs
 - Surpassing RO fences when activated with a PRNG



New active wire fences for 6× improved power side-channel security

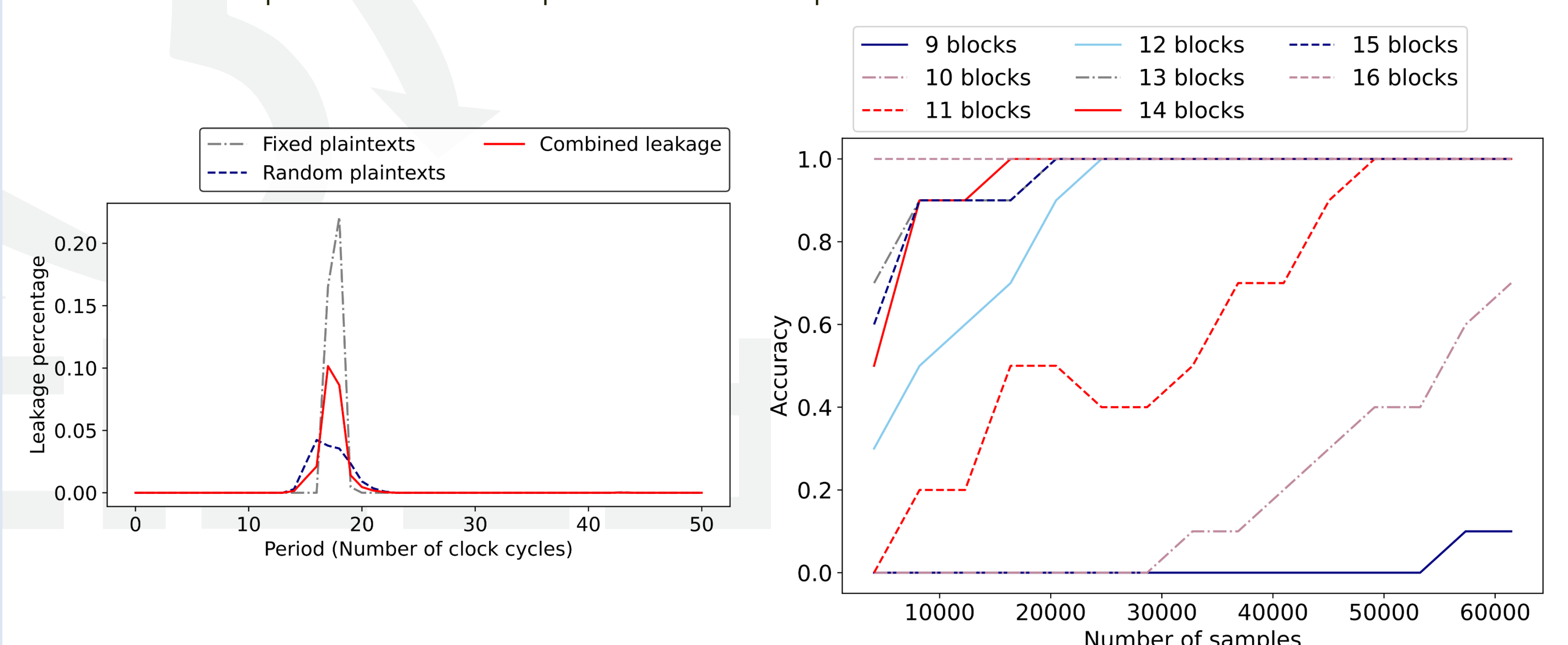
Fault Injection Victim [FPL'20]

- Satisfiability don't-cares (SDCs): Internal design states which are never reached
 - e.g., two signals cannot both be logic '1' at the same time
- Use SDC signals as Trojan activation signals
 - Stealthy Trojan only activated in faulting conditions**
- Hide Trojan into in the Sbox of an AES circuit to leak AES key



Fault Injection Results

- Using blocks of 4,096 ROs each
- Tested on commercially available cloud FPGA instances
- Success of the attack is measured by the leakage of the AES key
 - Percentage of ciphertexts at the output that are actually the key
- The optimal activation frequency is centered around a specific frequency
 - Most likely the resonance frequency of the PDN
- With enough samples, the key is the most occurring value (accuracy = 100%) at the output when the plaintexts are pseudorandom



Controlled fault injection is a threat to cloud FPGAs



SecureCells: A Secure Compartmentalized Architecture

Atri Bhattacharyya et al., EPFL



44th IEEE Symposium
on Security & Privacy

The Need for Secure Compartmentalization

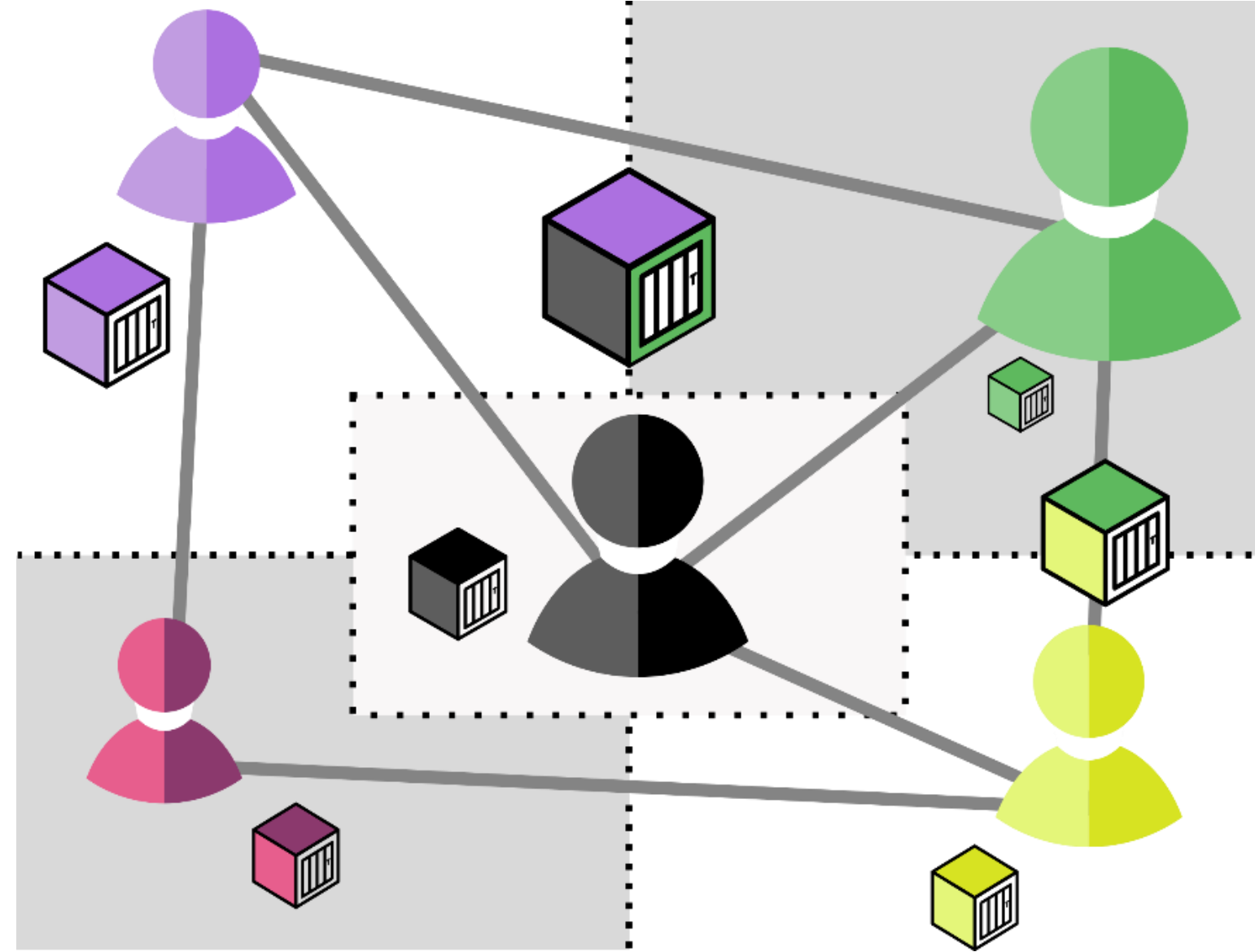
Problem: Applications have parts

- Lack of mutual trust
- Need to communicate

Solution: Compartmentalize programs

Why:

- Principle of Least Privilege
- Defense in Depth



Compartments

- Application-level abstraction of isolation
- Have conceptually different roles
- Have private/shared data and code
- Communicate: control flow and data flow

SecureCells Abstractions

Protects:

- Compartments (**S**ecurity **D**ivision)
- Memory VMAs (**C**ells)

By

- Cell-granularity access control
- Checked user insts. for compartment switch, perm. transfers

Goals:

- Security: Prevents Privilege Escalation
- Performance: Low overhead compartmentalization
- Flexibility: Compatible with generic isolation policies

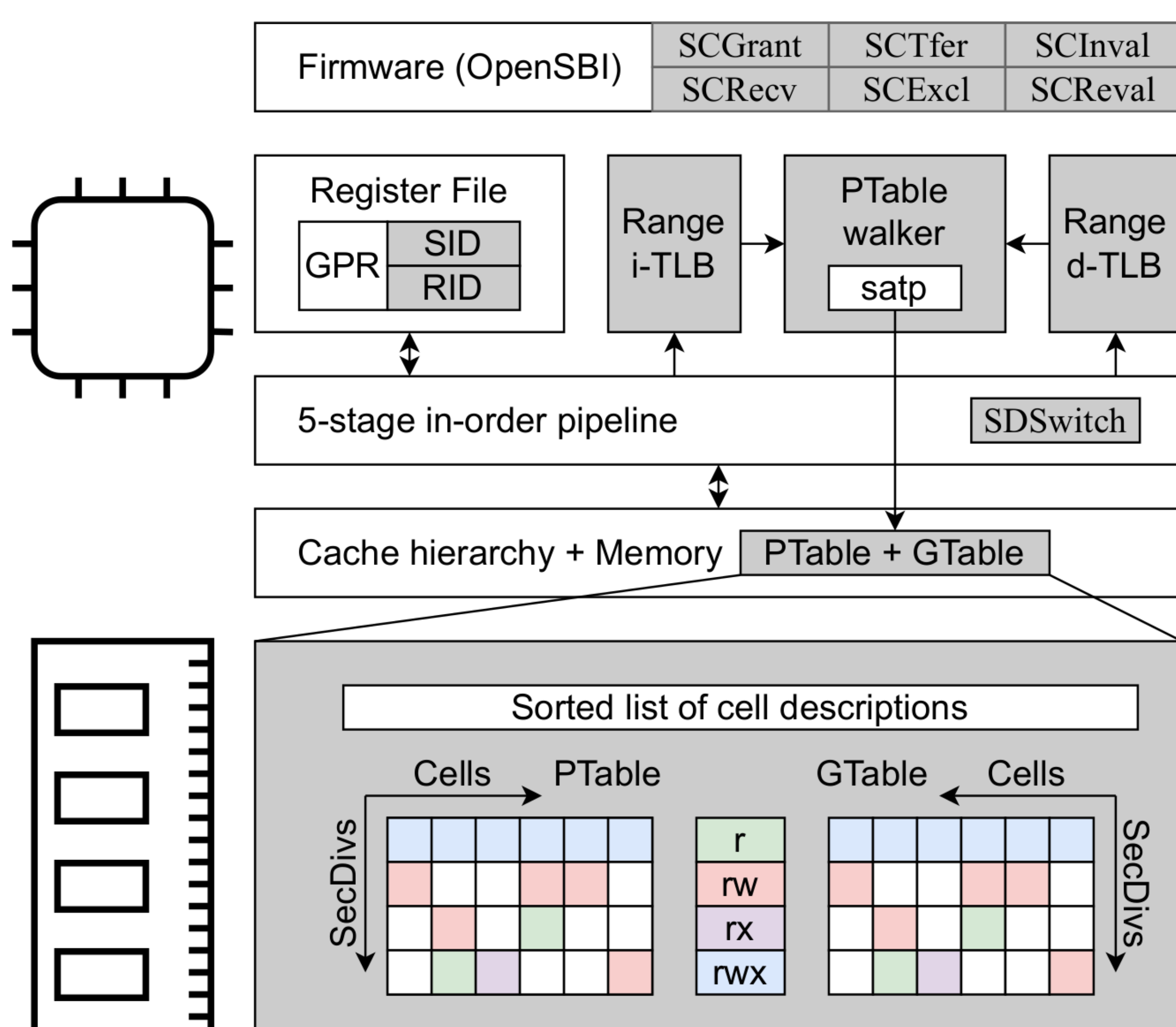
Memory Views for Compartments

SecureCells PTable (replaces traditional page tables)

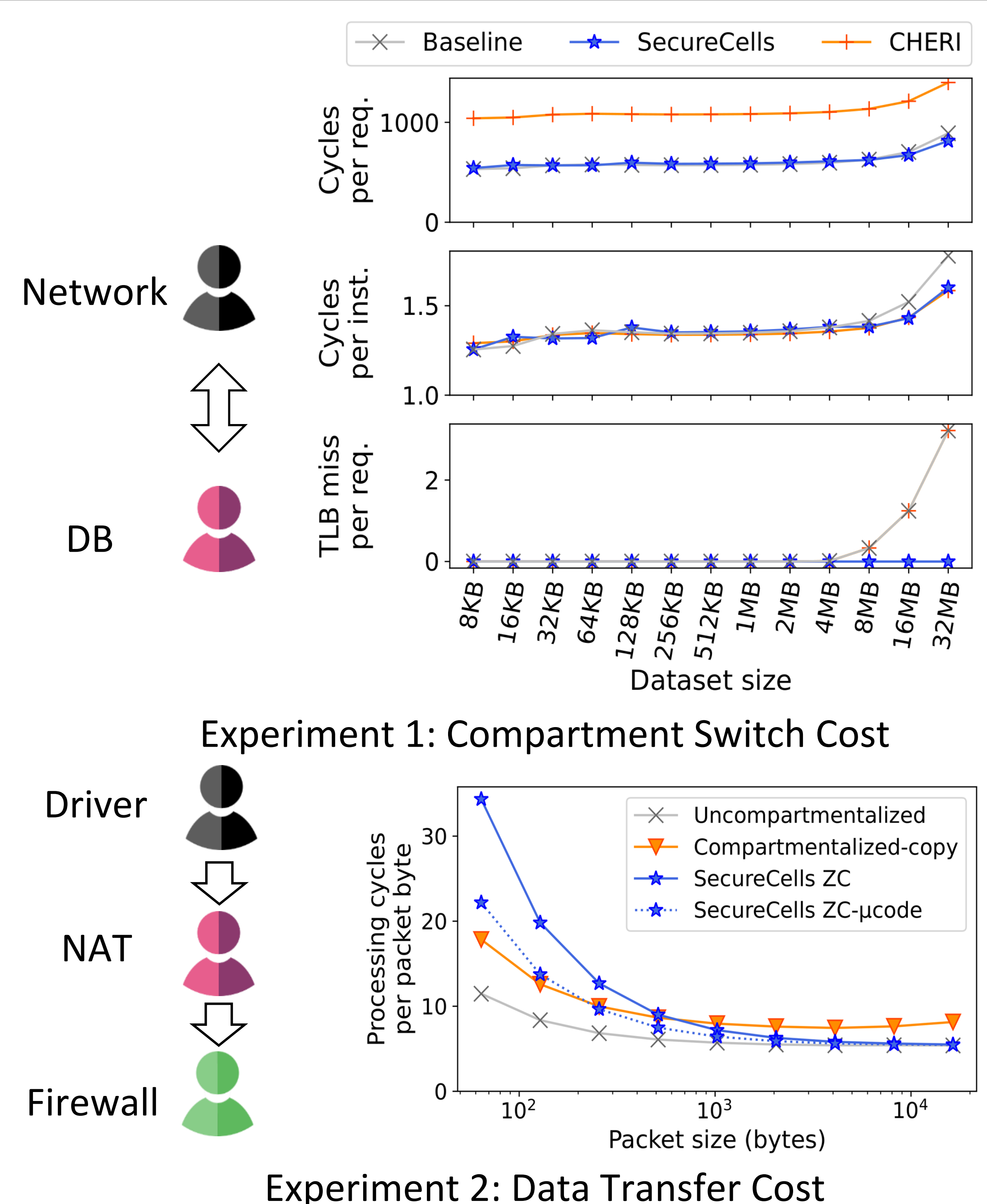
SecureCells Implementation

Prototyped on in-order RISC-V Rocketchip

- 8-cycle compartment switch
- 32x fewer TLB entries with 99.9%+ hit rate
- Permission transfers implemented in firmware (OpenSBI)



Performance



Gaëtan Frusque, Olga Fink
Intelligent Maintenance and Operations Systems Lab, EPFL

Motivations: High-frequency data processing

- High frequency (HF) signals from acoustic sensors and accelerometers are widely used in predictive maintenance.
- HF signal are high dimensional, often noisy with many artifacts.
- End-to-end deep learning model remain a challenge for HF signals.

L-WPT for data-adapted time-frequency representation of high-frequency signals

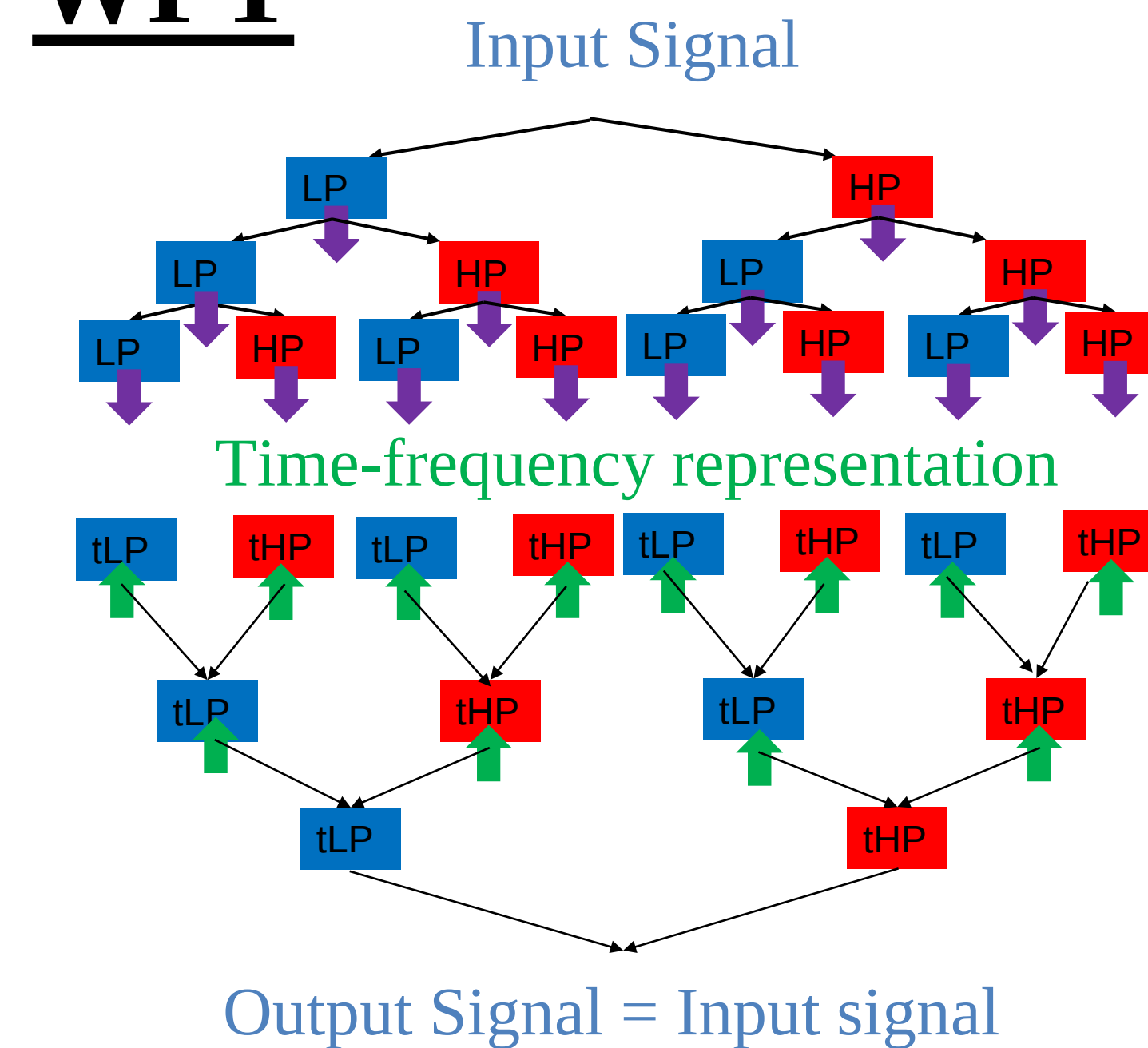
Deep Learning inspired by signal processing:

- Inspired by Wavelet Packet Transform (WPT) Architecture. The WPT is an invertible transformation that convert a temporal signal into a time-frequency signal.
- Learnable denoising inspired by wavelet shrinkage.

Advantages:

- Smart initialisation to behave as a WPT.
- Few parameters for high learning capabilities.
- Interpretable parameters.

WPT

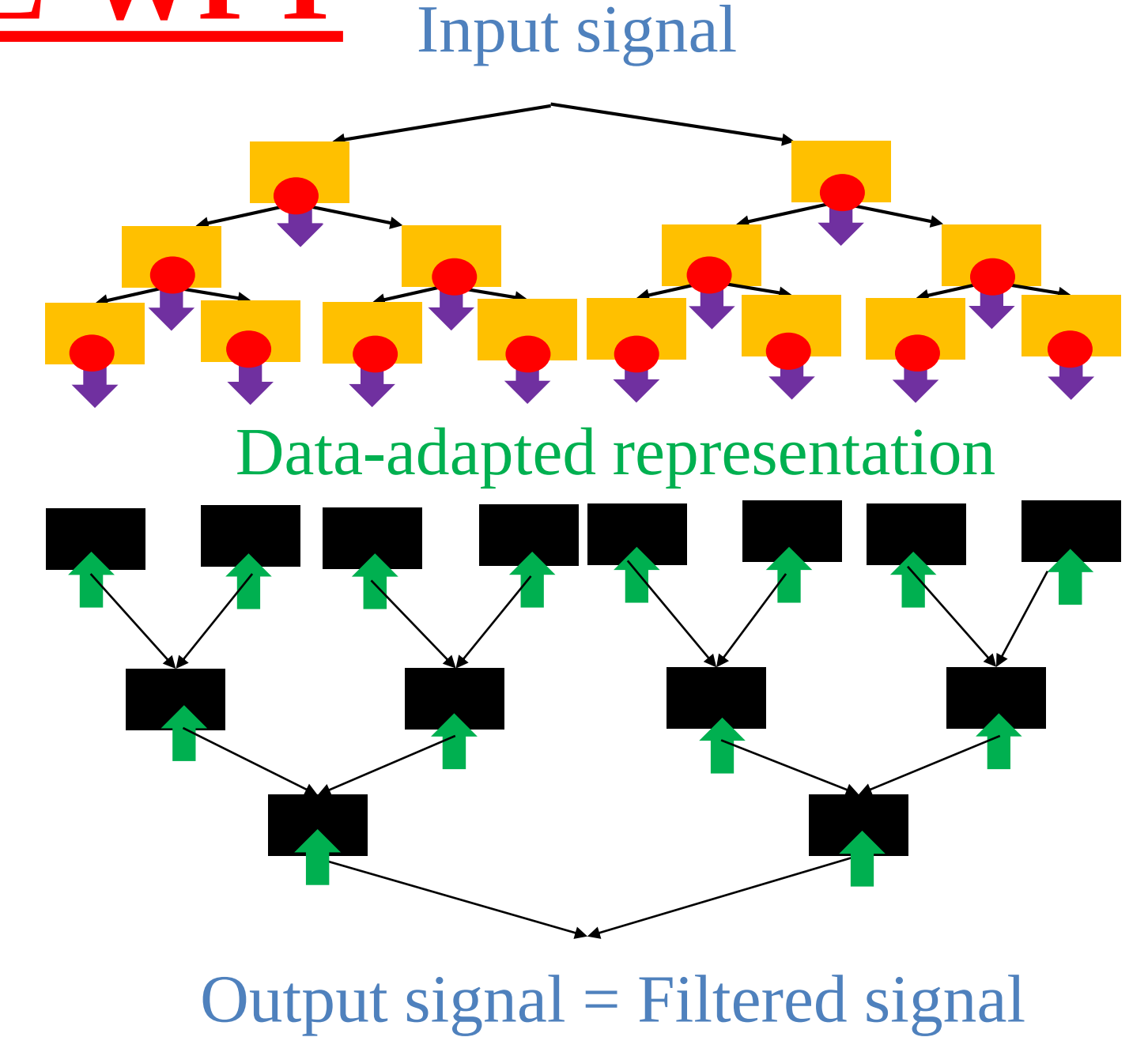


LP = Low-pass filter ↓ = Subsampling by 2 HP = High-pass filter ↑ = Upsampling by 2 = learnable CNN layer = transpose CNN layer = Learnable denoising

Frusque, Gaëtan, and Olga Fink. "Learnable wavelet packet transform for data-adapted spectrograms." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

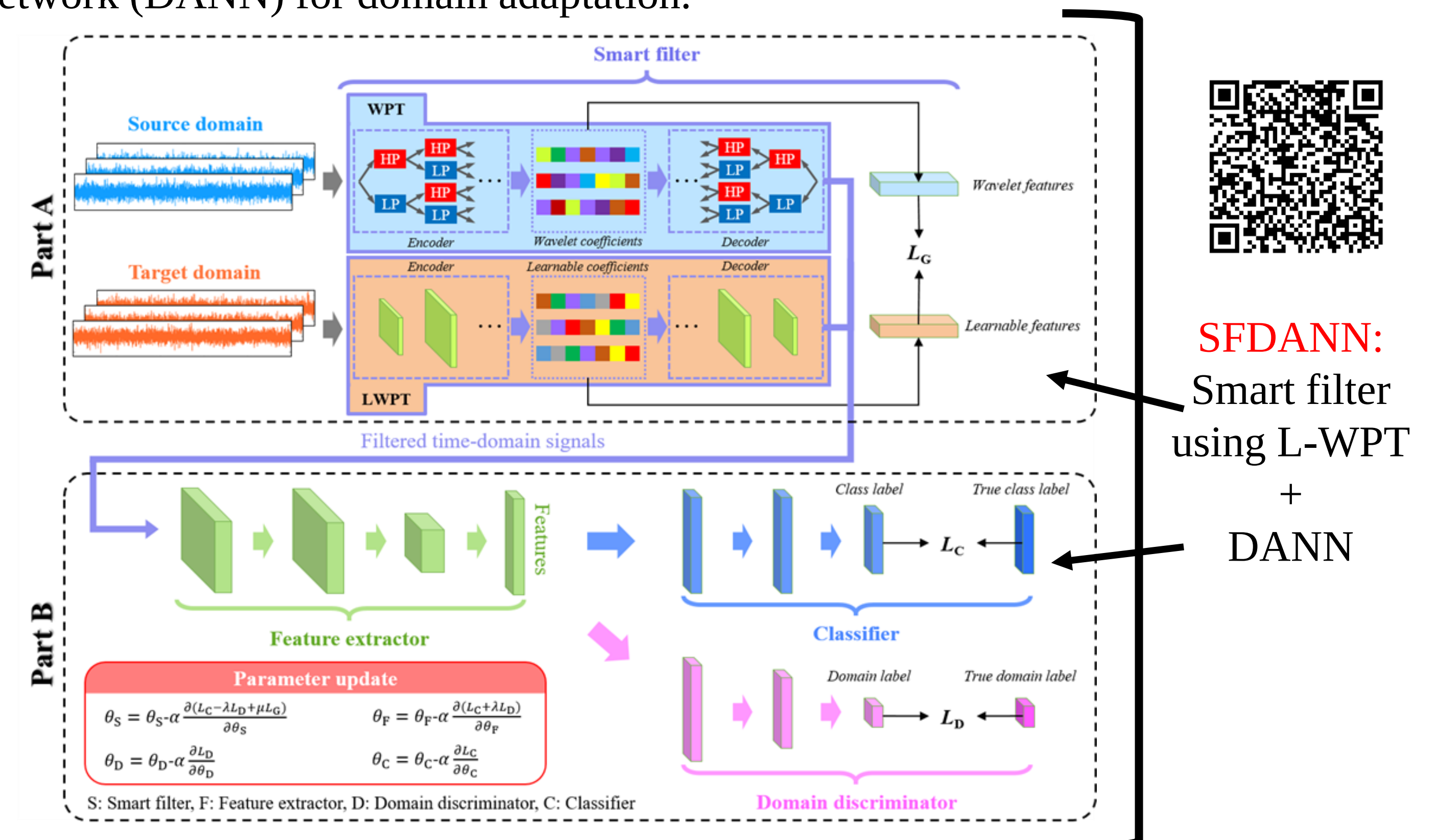


L-WPT



Application 2: Domain adaptation

- **Problem:** Unlabeled target dataset are often more noisy than the labeled source.
- **Proposition:** L-WPT is used as a filter to denoise the target signal and make it more similar to the source signal. It is combined with Domain Adversarial Neural Network (DANN) for domain adaptation.

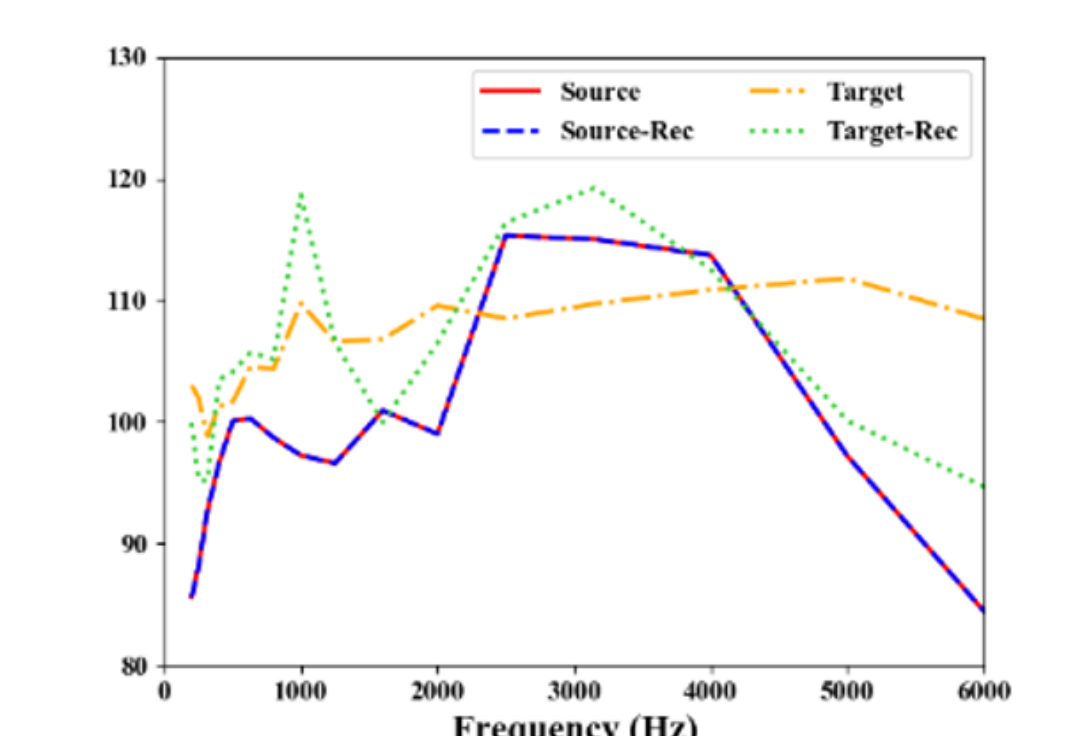


- **Dataset:** Case Western Reserve University bearing dataset and a real application on slab track monitoring.

The smart filter using L-WPT is able to improve the domain adaptation performances of DANN.

Noise	CNN	DANN	SFDANN
No noise	96,8	100	99,9
SNR=0	45,5	92,2	96,3
SNR=5	31	52,1	83,9

Average accuracy CWRU dataset

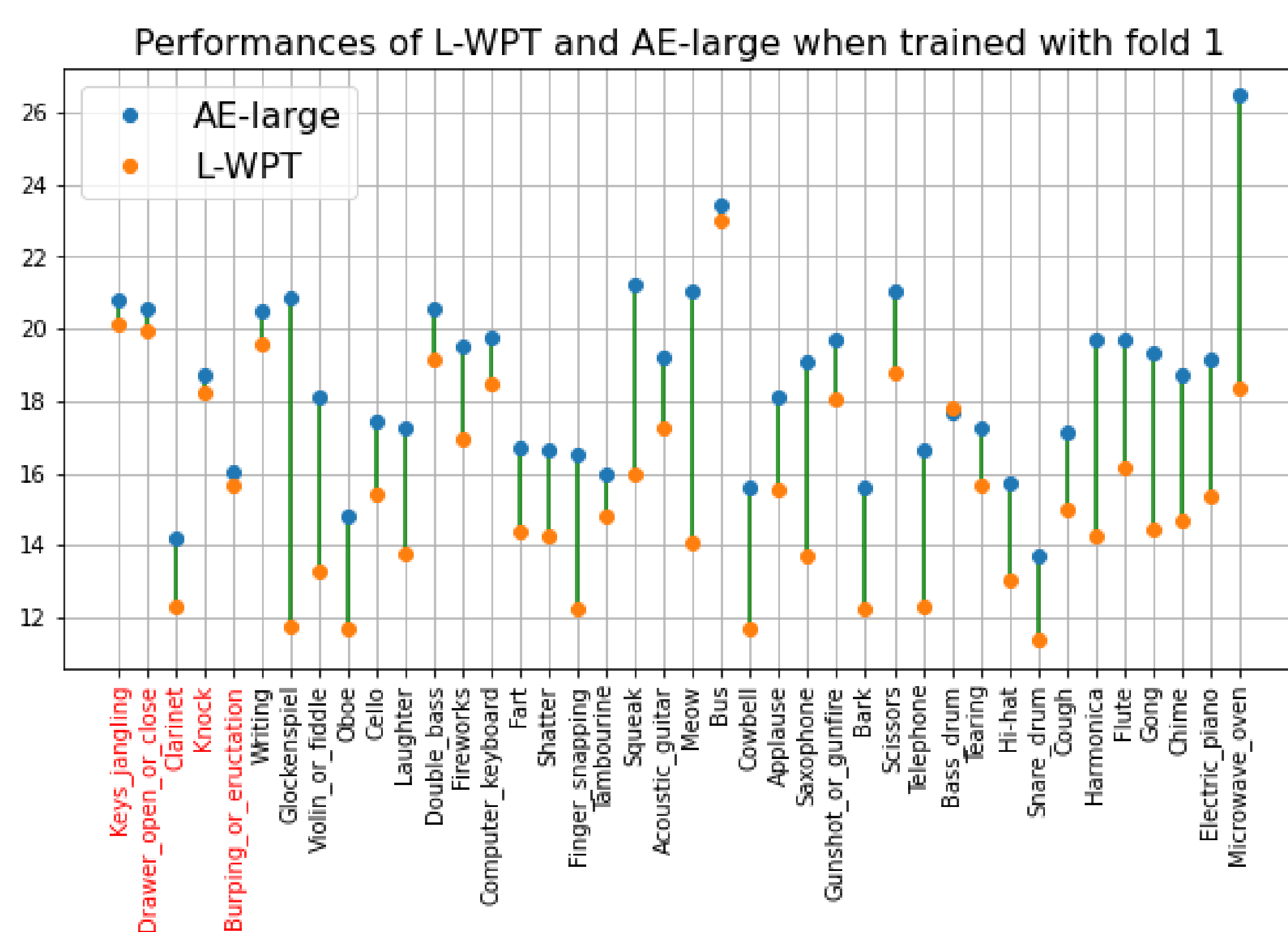


Frequency content of the input and output of the smart filter

Dai, Baorui, et al. "Smart filter aided domain adversarial neural network for fault diagnosis in noisy industrial scenarios." *Engineering Applications of Artificial Intelligence* 126 (2023): 107202.

Application 1: Background denoising

- **Dataset:** 40 sound classes + Aeroport background noises from the DCASE 2018 challenge
- **Loss function:** Supervised denoising, Mean Squared Error (MSE) between pure and noisy signal.
- **Training:** 5 classes from the 40 classes + Barcelona aeroport background noises
- **Test:** All 40 classes + all aeroport background noises.



L-WPT outperforms other methodologies and generalises better to sound classes that are unseen during the training.

Model	MSE Score test dataset	Number of parameters
WPT denoising	28,2	-
AutoEncoder (AE-large)	17	550K
DeepLab-V3 denosing	16,1	>1M
L-WPT	14,8	8K

Best performances with 1000 times less trainable parameters than the second best performing method.

Frusque, Gaetan, and Olga Fink. "Robust time series denoising with learnable wavelet packet transform." *arXiv preprint arXiv:2206.06126* (2022).



Conclusion:

- L-WPT is a deep model inspired by signal processing.
- Shows good performances in multiple applications.

Other works:

- Anomaly detection, classification, DWT: Michau, Gabriel, Gaetan Frusque, and Olga Fink. "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series." *Proceedings of the National Academy of Sciences* 119.8 (2022): e2106598119.
- Slab track monitoring, domain adaptation: Dai, Baorui, et al. "Slab Track Condition Monitoring Based on Learned Sparse Features from Acoustic and Acceleration Signals." *arXiv preprint arXiv:2205.05365* (2022).

Future directions:

- Combine L-WPT forward module and neural network.
- Optimize the training time of the L-WPT.

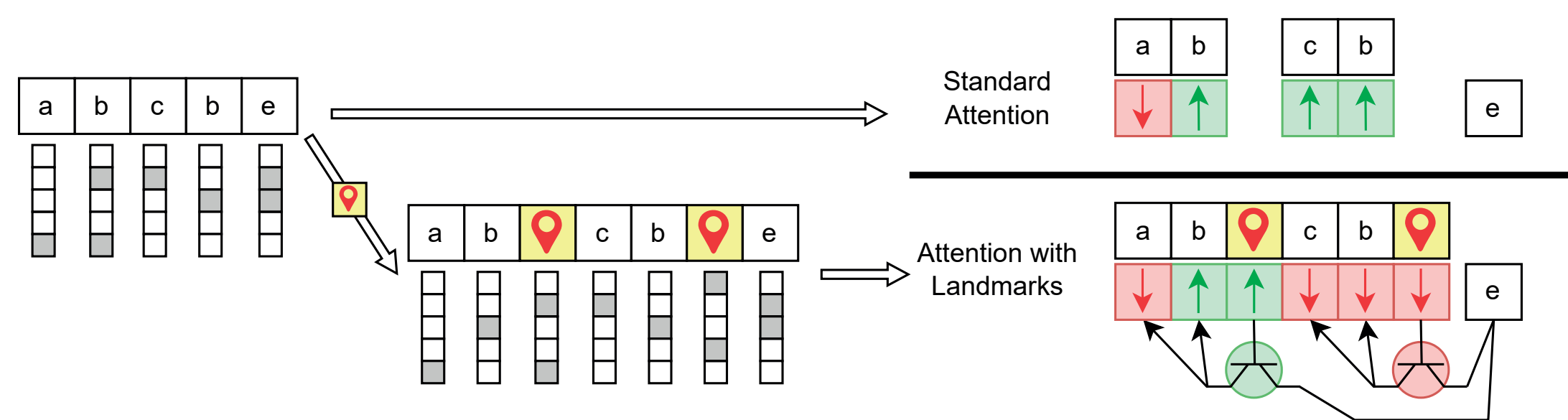
Overview

Problem

- Attention offers the flexibility of accessing the representation of any other token in each layer.
- This flexibility comes with quadratic computational cost and highly problematic memory footprint, limiting the context length.
- Previous approaches to remove this limit often sacrifice the random-access flexibility of attention.

Solution

- We break the input into blocks of fixed length and introduce a special token for each block, called a landmark, which acts as a gate for attending to its corresponding block.
- The gating mechanism is controlled by the attention score to the landmark token.
- At inference time, we compute the attention scores of the landmarks and retrieve the blocks corresponding to the highest scoring landmarks (active gates), integrating them into the attention.
- Our proposed approach maintains the random-access flexibility of attention and offers an alternative solution to the recurrent memory approaches.



Training

- We add a landmark token after every l_{block} tokens. The initial embedding for the landmark token is trained (similar to any other token), thus the vocabulary size is increased by one.
- Each landmark token acts as a gate for attending to the preceding l_{block} tokens.
- Landmark tokens are passed through the transformer as any other token while their representations are updated using the self-attention mechanism.
- To have the gating mechanism, we want the attention weight for a token to depend on the similarity of the query vector with both the token's key and with the key of its block's landmark token.
- First we replace the softmax function in the attention layer with **Grouped Softmax** which computes the softmax separately over each group.

$$\sigma_G(\mathbf{v}, \mathbf{g})_i := \text{GroupedSoftmax}(\mathbf{v}, \mathbf{g})_i := \frac{e^{\mathbf{v}_i}}{\sum_{j: \mathbf{g}_j = \mathbf{g}_i} e^{\mathbf{v}_j}}.$$

Grouping Scheme

- For each block, we create a separate group and put its regular (non-landmark) tokens in that group.
- When computing the attention weights for the i -th token, landmark tokens for other blocks are placed in the same group as the i -th token.
- The landmark token for the i -th token's block is **ignored** when computing the attention weights for the i -th token.

Landmark Attention

- Using the above grouping, the attention weight for each token can be computed as the product of the token's attention weight and its corresponding landmark token's attention weight.
- Under this scheme, the attention weights sum to one same as in the standard softmax function.
- Since tokens in the same block and the landmark tokens share the softmax group, the model has to choose between attending to other blocks and current tokens.
- Intuitively, the grouping forces the model to only attend to relevant blocks because of this trade-off.

$$S_{i,j} := \text{SoftmaxScore}(\mathbf{Q}, \mathbf{K})_i$$

$$:= \text{GroupedSoftmax}\left(\frac{\mathbf{Q}_i^\top \times \mathbf{K}}{\sqrt{d_{\text{head}}}}, \mathbf{G}_i\right)$$

$$\text{Att}(\mathbf{Q}, \mathbf{K})_{i,j} := \begin{cases} 0 & p_j = j \\ S_{i,j} & \mathbf{G}_{i,j} = \mathbf{G}_{i,i} \wedge p_j \neq j \\ S_{i,j} \cdot S_{i,p_j} & \mathbf{G}_{i,j} \neq \mathbf{G}_{i,i} \wedge p_j \neq j \end{cases}$$

i	0	1	2	3	4	5	6	7	8
$\text{mask} + \frac{Q_i^\top K}{\sqrt{d_{\text{head}}}}$	1	1	1	1	1	1	1	mask	mask
p_i	2	2	2	5	5	5	8	8	8
$G_{6,j}$	2	2	8	3	3	8	8	8	-1
$S_{6,j}$	0.5	0.5	0.33	0.5	0.5	0.33	0.33	0	ign
$W_{6,j}$	0.167	0.167	0	0.167	0.167	0	0.33	0	0

Results

Pass Key Retrieval Task

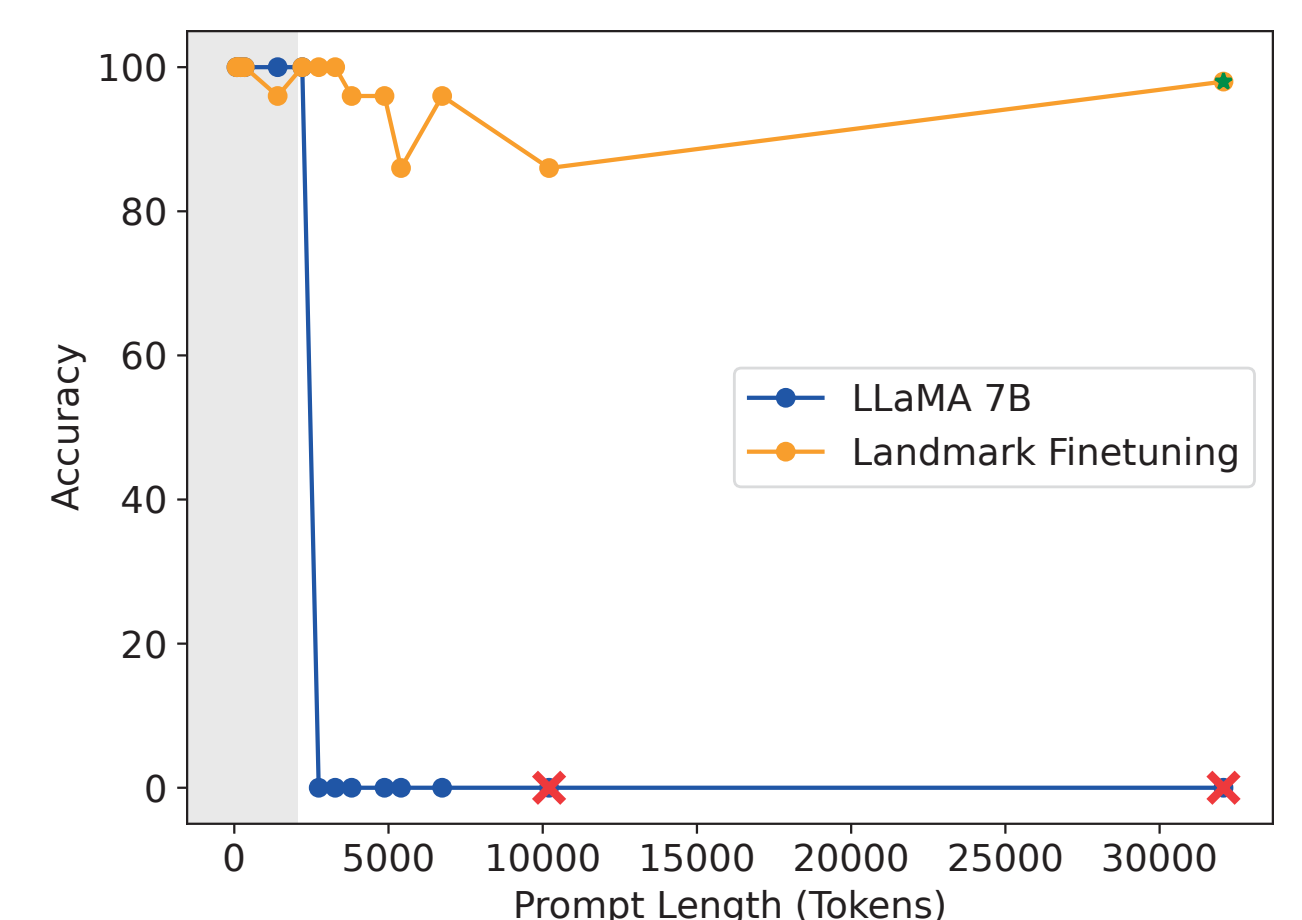
- The task is to find a pass phrase hidden in a long piece of text.
- Instances of this task are generated by randomly choosing the pass key (a number between 1 and 50000) as well as its location.
- Prefix and suffix strings filled with a repetitive text are added to the pass key to ensure the desired length.

There is an important info hidden inside a lot of irrelevant text. Find it and memorize them. I will quiz you about the important information there.
<prefix filler by continuously repeating: The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again.>
The pass key is <PASS KEY>. Remember it.
<PASS KEY> is the pass key.
<suffix filler>
What is the pass key? The pass key is

- We use the success rate as the evaluation metric.

Evaluation

- We fine-tune LLaMA 7B using our method.
- The original model fails to retrieve the pass key and even runs out of memory as the prompt gets longer (marked with a red cross)
- In contrast, the fine-tuned model using landmarks can successfully identify the pass key with a high success rate.
- A more efficient inference mechanism allows offloading KV-Cache to CPU which resolves the memory problem further (a green star marks this mechanism's deployment).
- The results demonstrate that our method can be successfully used even during fine-tuning to extend the context length limit to arbitrary large values.

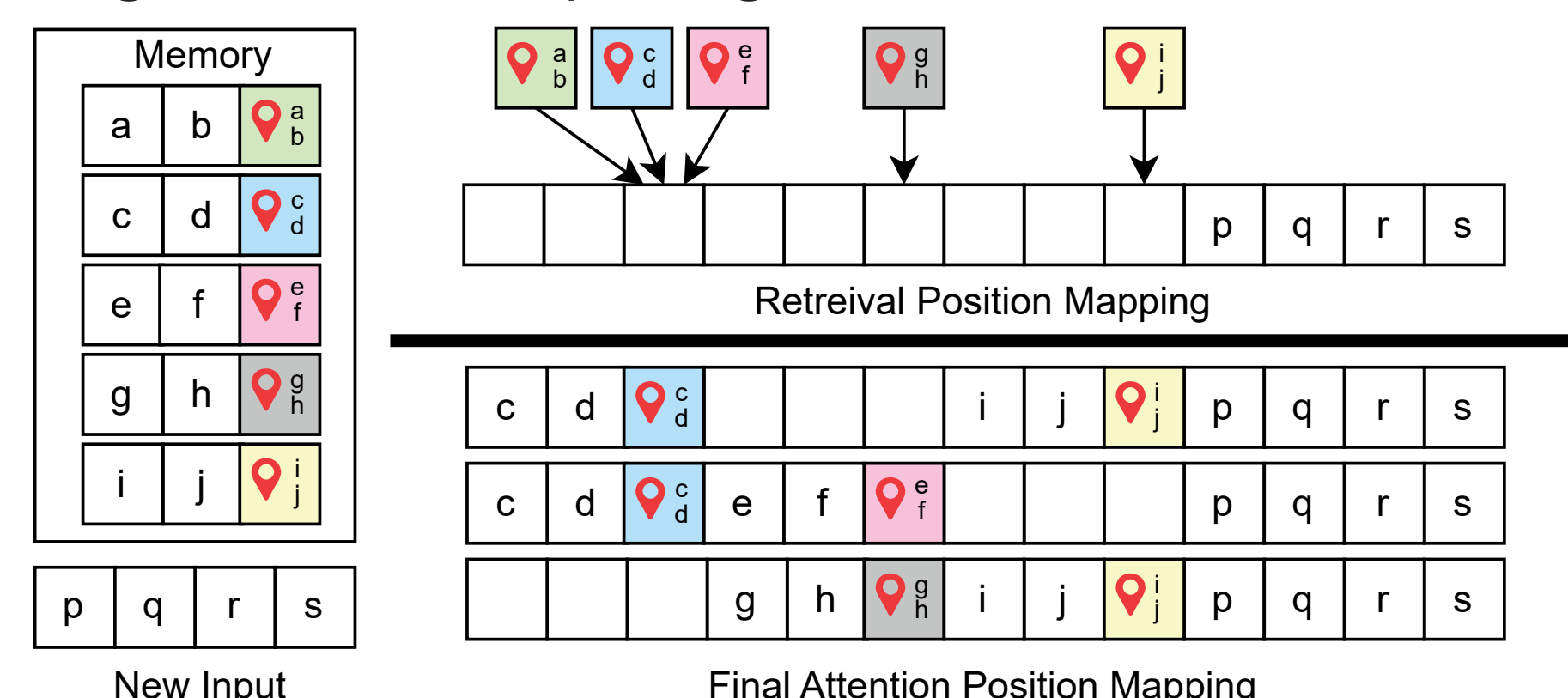


Inference

- Similar to training, the input gets augmented by a landmark token after every l_{block} tokens.
- Then the input is broken into chunks of l_{local} length. The chunks are iteratively fed to the model from the beginning to the end.
- To retrieve relevant blocks, each attention layer has access to a cache of previous blocks. The cache stores the key-value vectors for all tokens of those blocks, including the landmark tokens.
- When processing each chunk at each layer, we first compute the attention score of each token with the landmark tokens currently in the cache.
- For each token, we compute the attention score for all the tokens in the top-k highest scoring blocks and prepend the obtained attention matrix to the local attention matrix.
- We finish by applying GroupedSoftmax with the same grouping as in training to the attention matrix and computing the weighted average of the value vectors to obtain the token's representation.

Sting Position Mapping

- Optimally, the tokens are encoded with their actual position index. However, a known flaw of Transformers is their inability to extrapolate to lengths not observed during training.
- Previous methods proposed to alleviate this flaw usually penalize or prevent attention to long distance tokens.
- We instead use a special approximate position mapping scheme since our main goal is to facilitate attending to large contexts. We allocate a segment with length $(k+1) \cdot l_{\text{block}}$ in the beginning of the context. We index the current chunk starting after this segment.
- We map the index for the latest k blocks to the corresponding place within the last k blocks in the allocated prefix segment. All other blocks in the cache have their position mapped to the first block in the allocated prefix segment.
- Once we decide which blocks to retrieve, we map those among the latest k blocks to the right of pre-allocated segment and map the rest of the blocks to the left of the pre-allocated segment, while respecting the order of the blocks.



Martijn de Vos, Akash Dhasade, Anne-Marie Kermarrec,
Rafael Pires, Mathis Randl, Rishi Sharma

Motivation

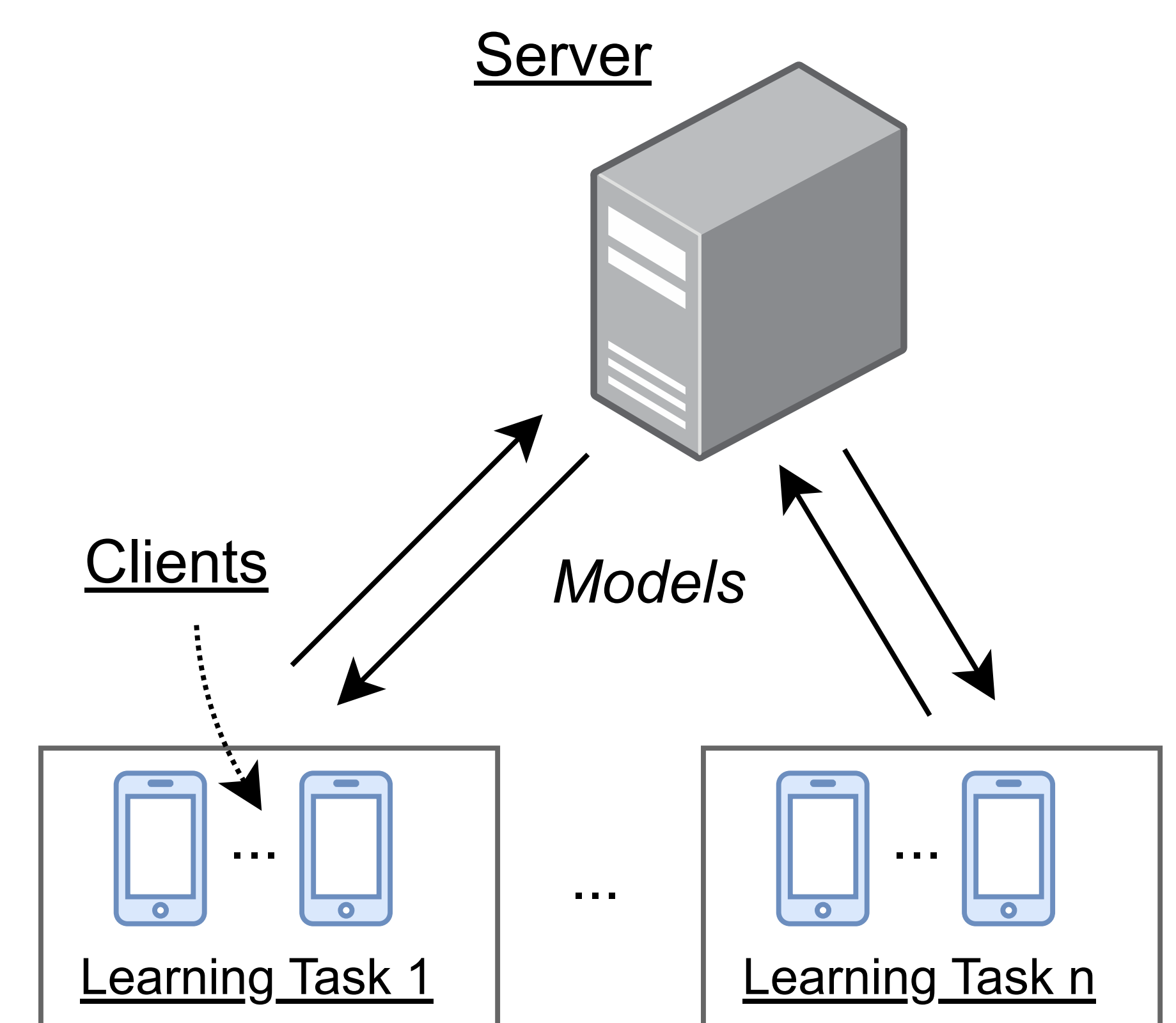
Server in FL Setups ^[1]

1. **Selecting** clients for model training
2. **Aggregating** model updates
3. **Tracking** availability and capabilities of devices

Challenges using Server

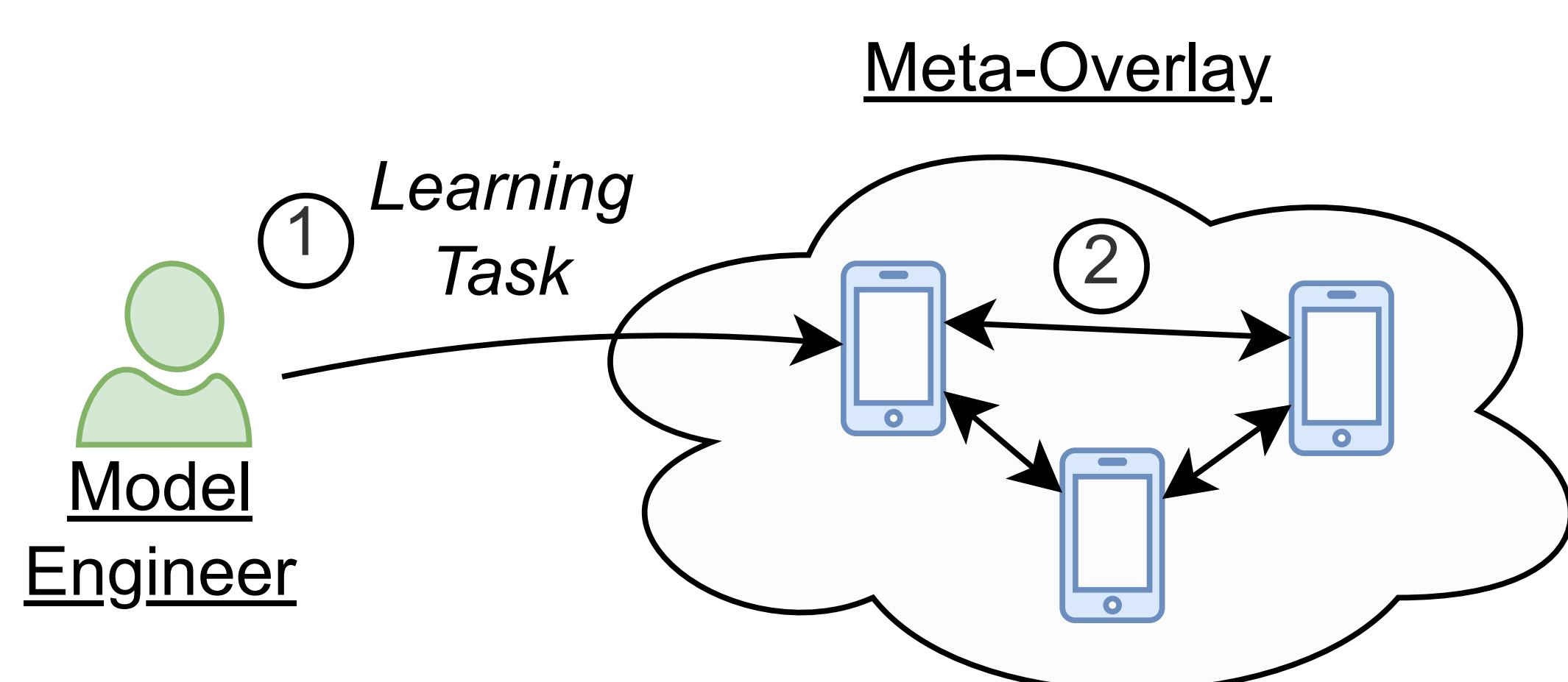
1. Single point of **failure**
2. Infrastructure and maintenance **costs**
3. Risk of **censorship**

RQ: How can we design a distributed learning setup without any critical dependency on a server?



Approach

Creating/Discovering Learning Tasks

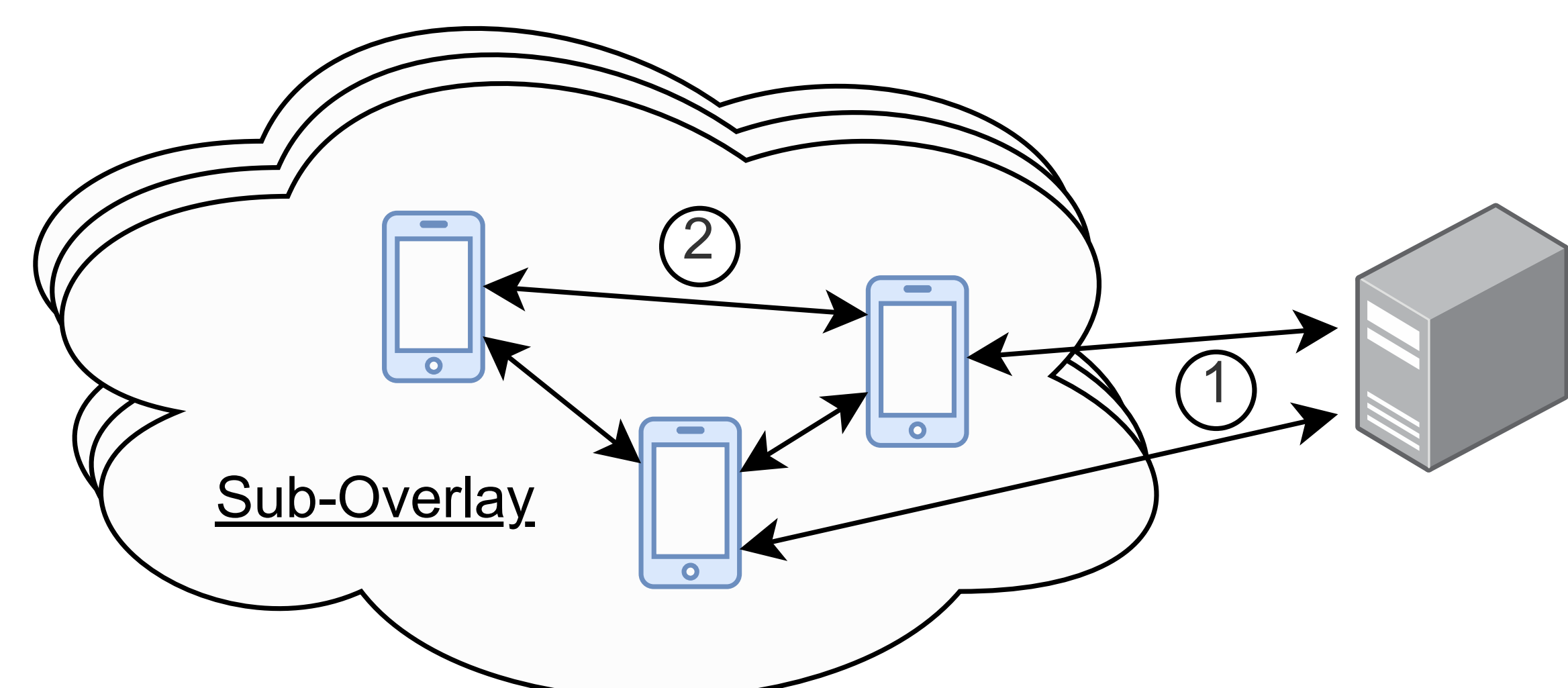


1. **Model Engineers** create learning tasks and publish in a decentralized meta-overlay
 - **Peer-reviewed** to prevent abuse
2. **End-users** join a decentralized meta-overlay to discover learning tasks
 - Peers gossip available learning tasks

Solving Learning Tasks

Federated Learning vs. Gossip Learning ^[2]

1. **Server available**
 - server-side aggregation (asynchronously)
2. **Server unavailable**
 - gossip aggregation
3. **Doubly weighted averaging**
 - server model vs. model age



A central server is optional in our design and not critical for progression of training!

Benefits

1. **Open infrastructure** for creating and solving learning tasks
2. **Autonomy** for end-users to work on the learning tasks they want
3. **No dependency** on central server
 - But a server can speed up the model training

Evaluation

Plan to evaluate using:

FedScale

- Realistic models
- Real-world device traces from FedScale ^[3]
- What is the model convergence speed compared to FL?
- How energy-efficient is DecentraLearn?

[1] Bonawitz, Keith, et al. "Towards federated learning at scale: System design." *Proceedings of machine learning and systems* 1 (2019): 374-388.

[2] Hegedűs, István, Gábor Danner, and Márk Jelasity. "Gossip learning as a decentralized alternative to federated learning." *DAIS*, 2019.

[3] Lai, Fan, et al. "Fedscale: Benchmarking model and system performance of federated learning at scale." *International Conference on Machine Learning*. PMLR, 2022.

Energy-Aware Decentralized Learning with Intermittent Model Training

Martijn de Vos¹, Akash Dhasade¹, Paolo Dini², **Elia Guerra**^{1,2}, Anne-Marie Kermarrec¹, Marco Miozzo², Rafael Pires¹, and Rishi Sharma¹

¹Scalable Computing Systems Laboratory, École polytechnique fédérale de Lausanne (EPFL)

²Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)

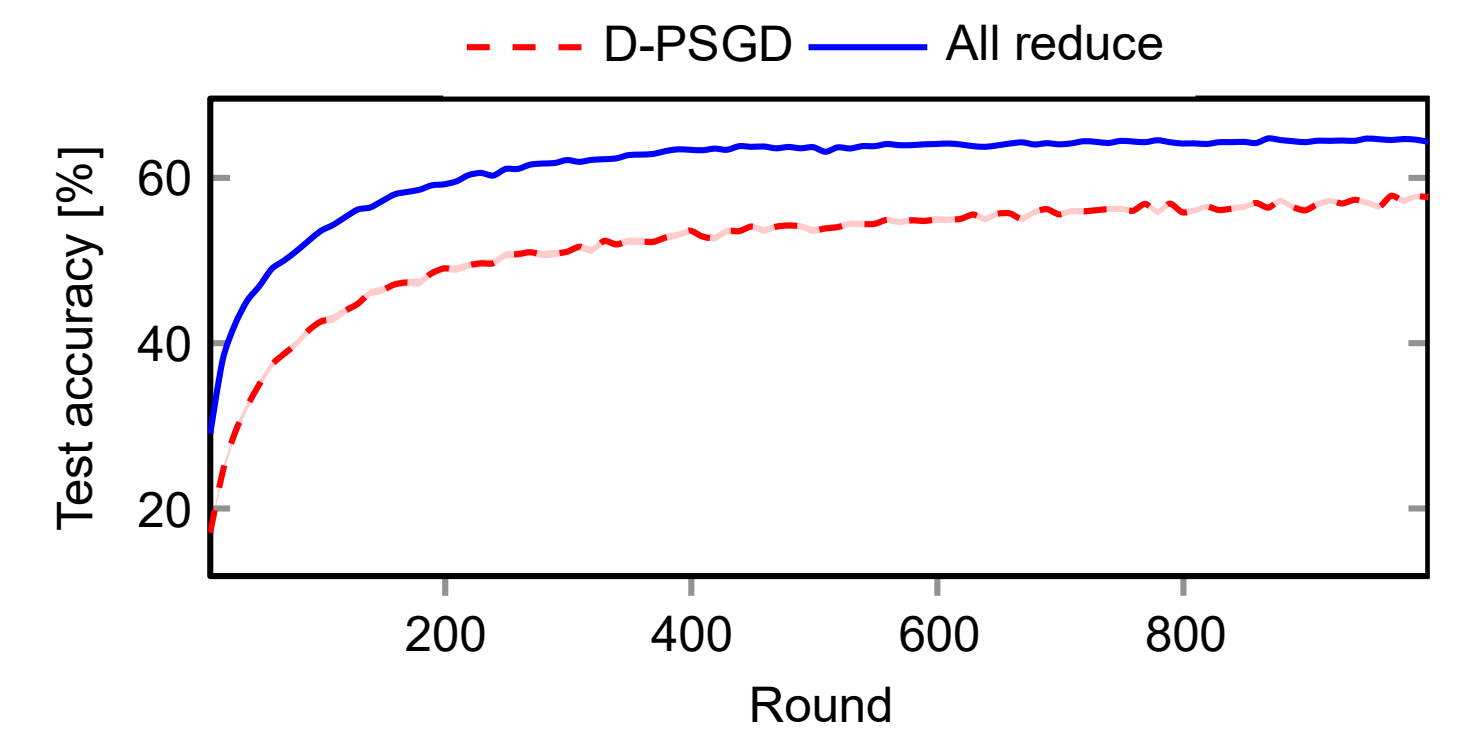
Introduction

Setting

- In decentralized learning (DL), nodes have intermittent network and constrained energy budgets
- Several efforts in the literature to improve communication efficiency
- Energy spent on the training process is often overlooked

Insights

- Training is more energy-demanding than sharing and aggregation
- DL convergence is faster with all-reduce aggregation
- Energy budgets distributed over the duration of the learning process



SkipTrain

Synchronization rounds

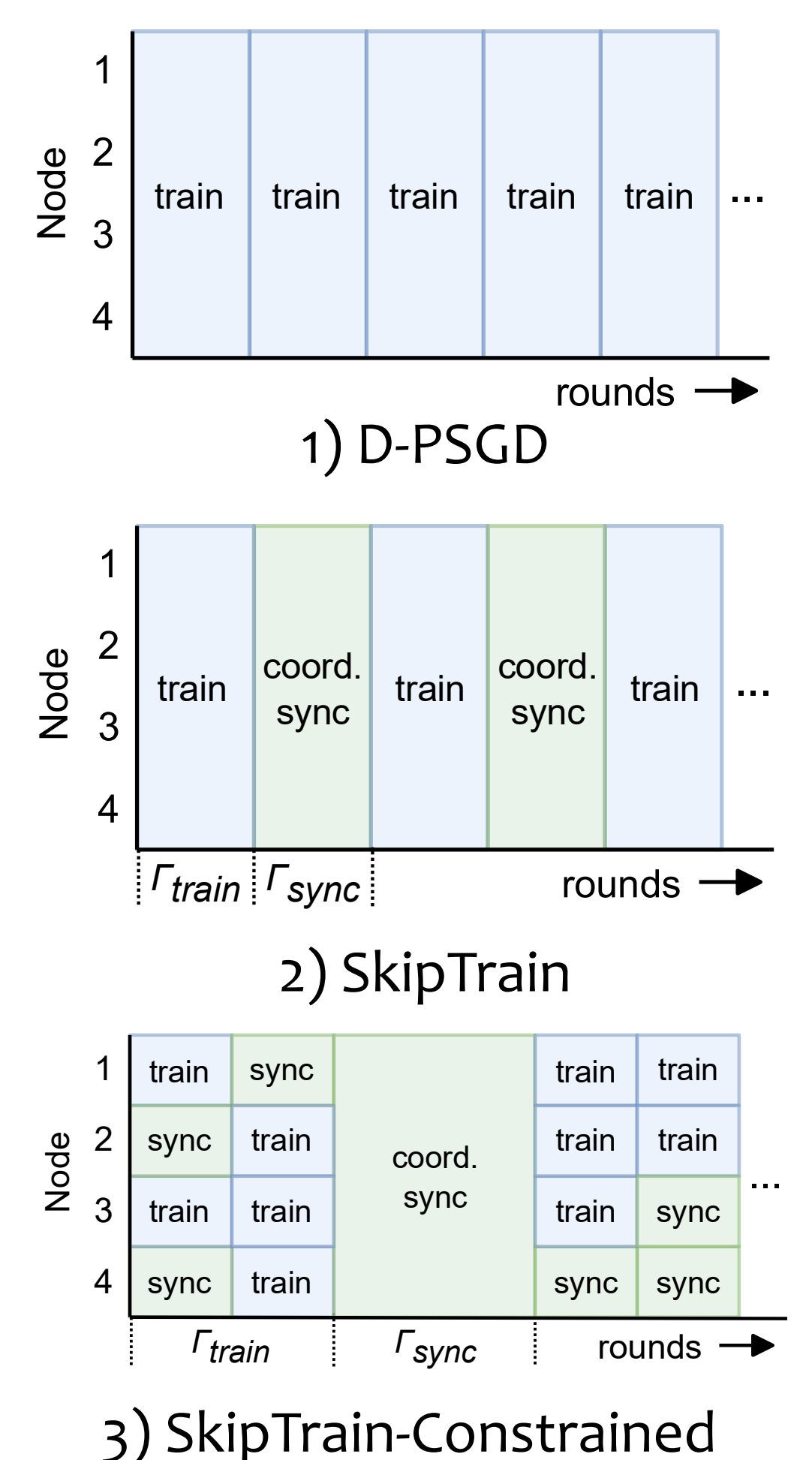
- In D-PSGD [1] each node iteratively executes training, sharing, and aggregation
- **Synchronization rounds:** sharing and aggregation
 - Each model converges towards the global average model [2]
- **Training rounds:** training, sharing, and aggregation
 - Train the model on the local dataset

Participation probabilities

- Each client has an upper bound on the percentage of battery that can be used for training
- Participation probabilities distribute the energy budgets (τ_i) across total number of rounds (T)

$$T_{\text{train}} = \frac{\Gamma_{\text{train}}}{\Gamma_{\text{train}} + \Gamma_{\text{sync}}} T$$

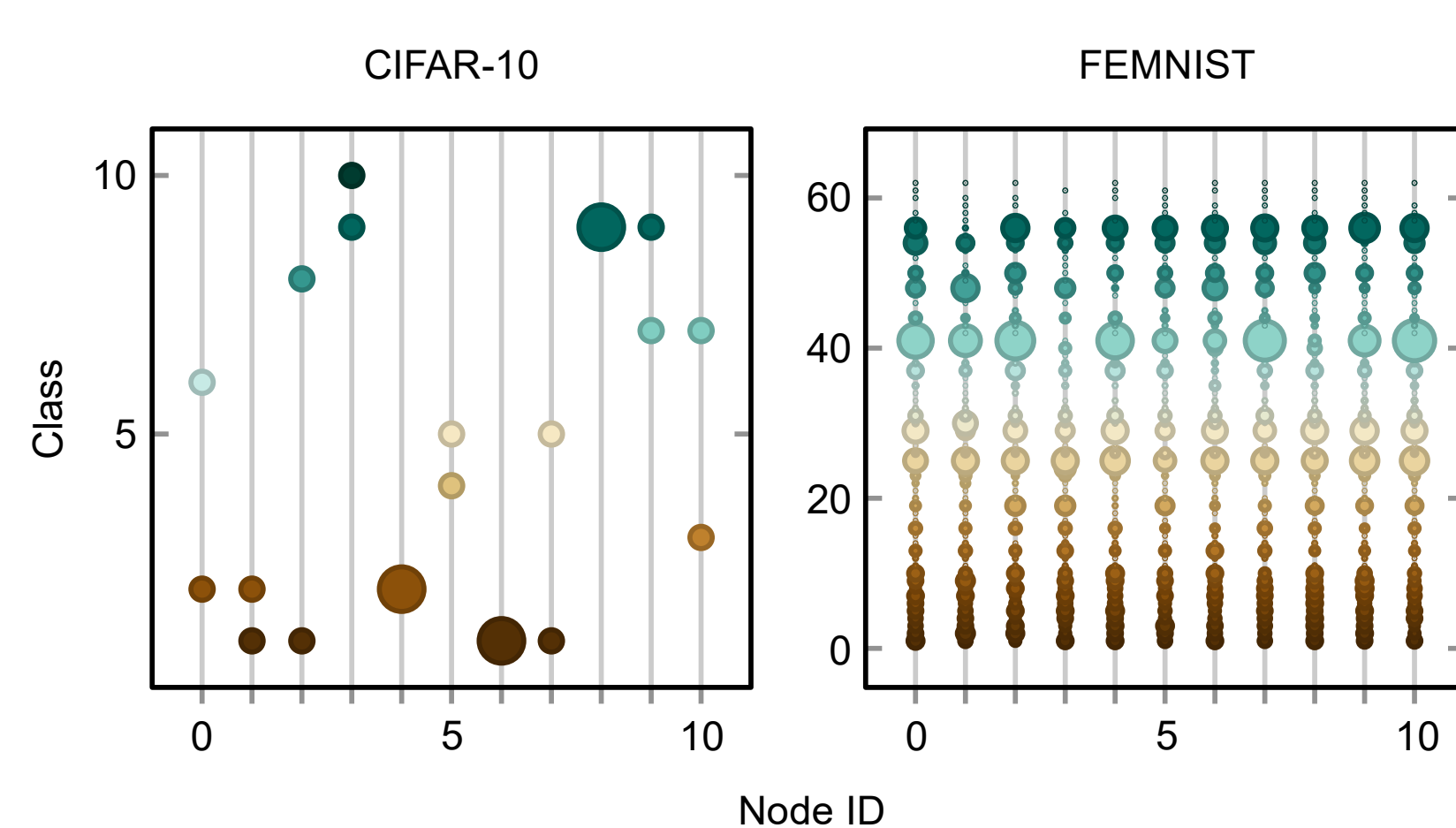
$$p_i = \min\left(\frac{\tau_i}{T_{\text{train}}}, 1\right)$$



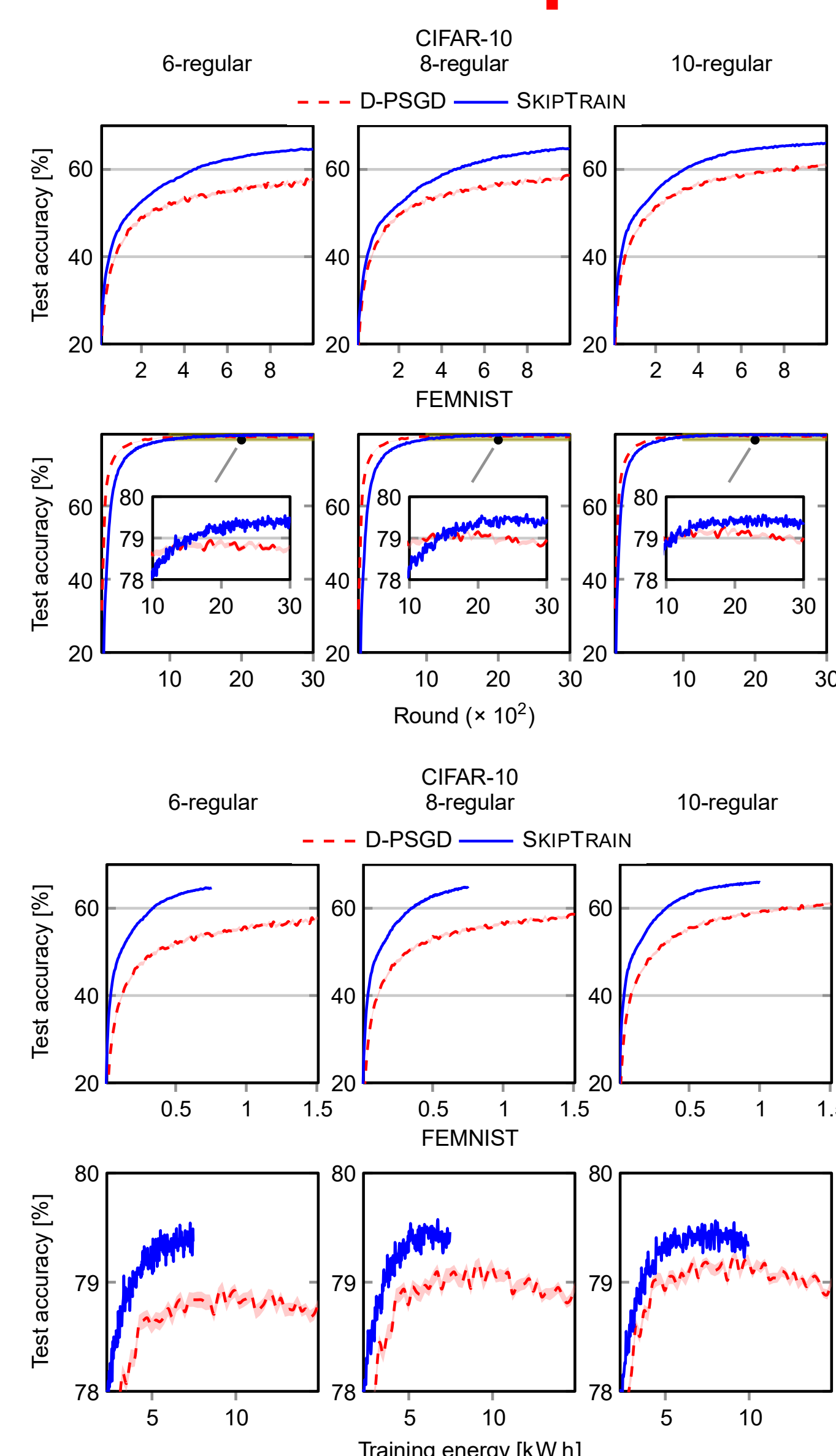
Evaluation

Experimental setting

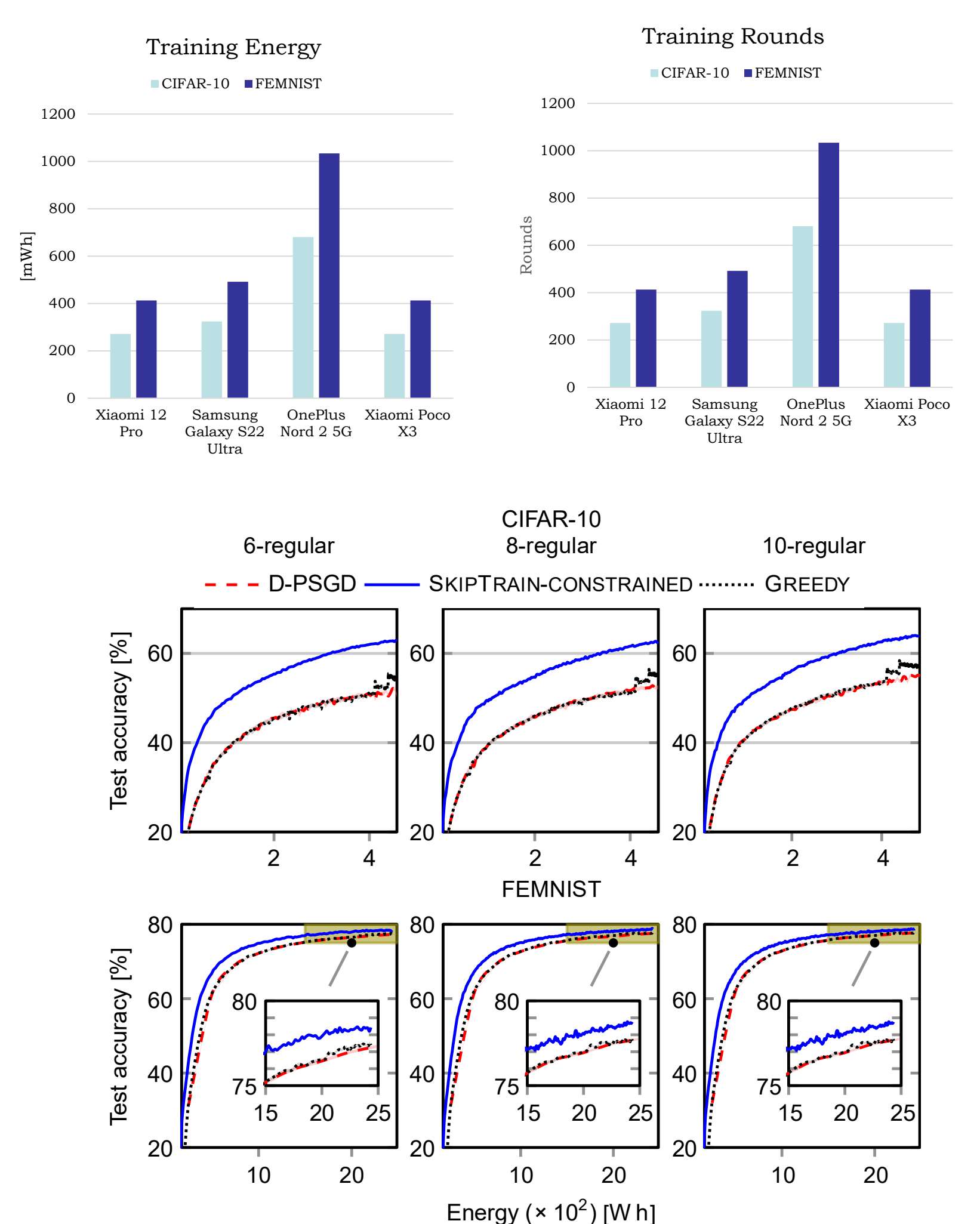
- Datasets: CIFAR-10 and FEMNIST
- Models: CNN
- **SkipTrain:** fixed number of rounds
- **SkipTrain-Constrained:** fixed energy budgets to compare accuracy



Results: SkipTrain



Results: SkipTrain Constrained



[1] X. Lian, et al. “Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent,” Advances in neural information processing systems, vol. 30, 2017.

[2] L. Xiao, et al. “Fast linear iterations for distributed averaging”, Systems & Control Letters, vol. 53, no. 1, pp. 65–78, 2004.

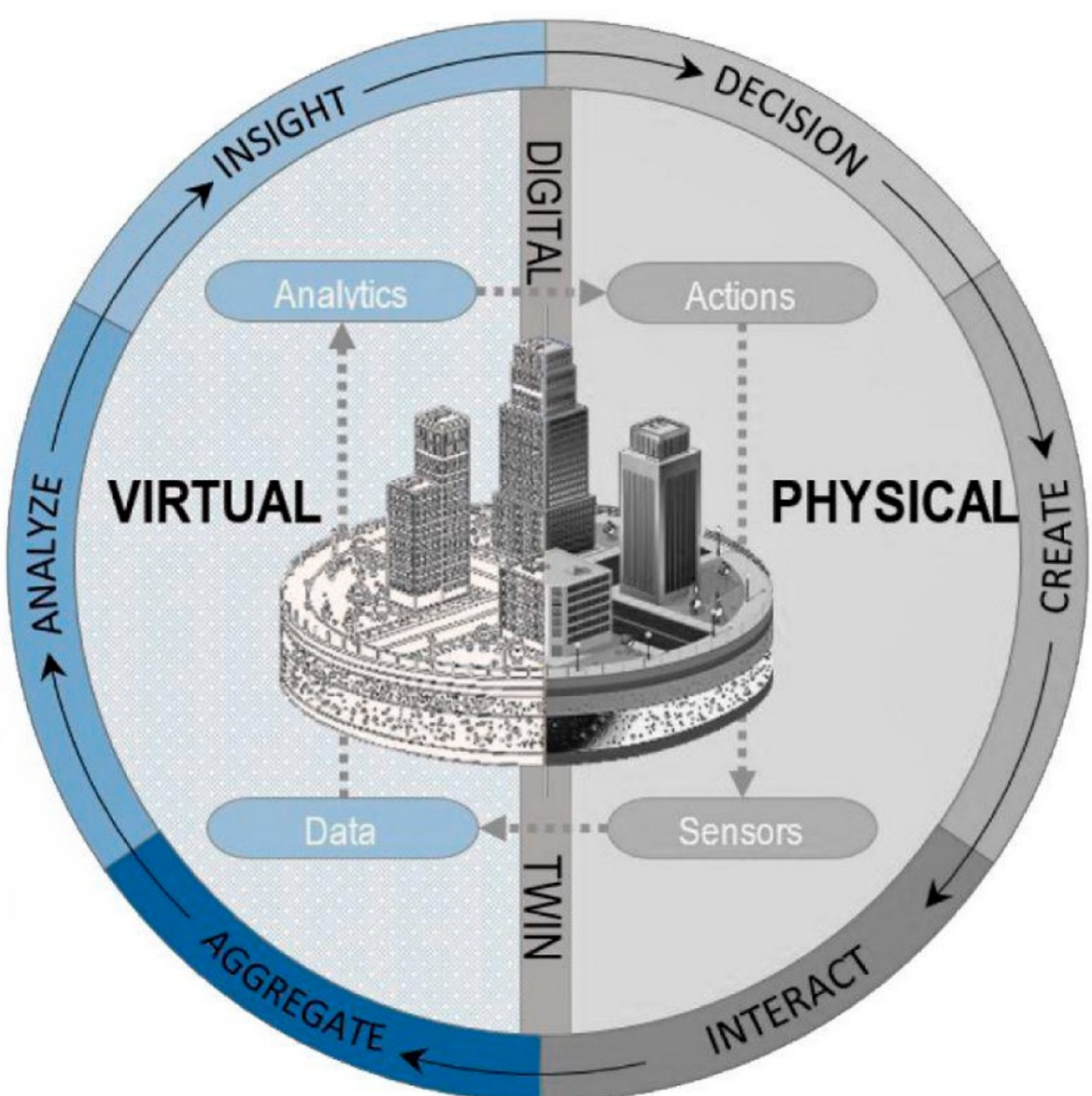


MOTIVATION

Urban areas are ideal to implement a coordinated, multi-sectoral response to climate change since they are:

- responsible for 75% of GHG emissions
- their livability significantly impacted by climate change
- represent a natural integrator of many systems (energy, transport, clean/waste- water, and buildings)

This can be achieved with the use of a **digital twin**, which is an algorithm designed to faithfully reflect a physical object, system or system of systems, along with all its dynamics.



OBJECTIVES

Develop and validate an **integrated tool to support stakeholders** in making data-driven and science-backed decisions for the achievement of climate and energy goals.

This will **enable them** to:

- ✓ probe the effectiveness of potential solutions,
- ✓ head off problems,
- ✓ prevent downtime, and
- ✓ develop new opportunities and planning strategies for the future.

The tool will be **validated** at the new EPFL datacenter experimental facility, the EPFL campus, and the demo sites in Aigle (VD) and in Lausanne (VD).



ACTIVITIES

This UrbanTwin tool **will involve**:

- a detailed model of critical urban infrastructure, such as energy, water, buildings and mobility,
- underlying socio-economic, environmental and climate drivers
- inter-dependencies among various systems
- multi-sectoral data acquisition from critical urban infrastructure to create a multi-scale digital twin running on a large-scale IT infrastructure.

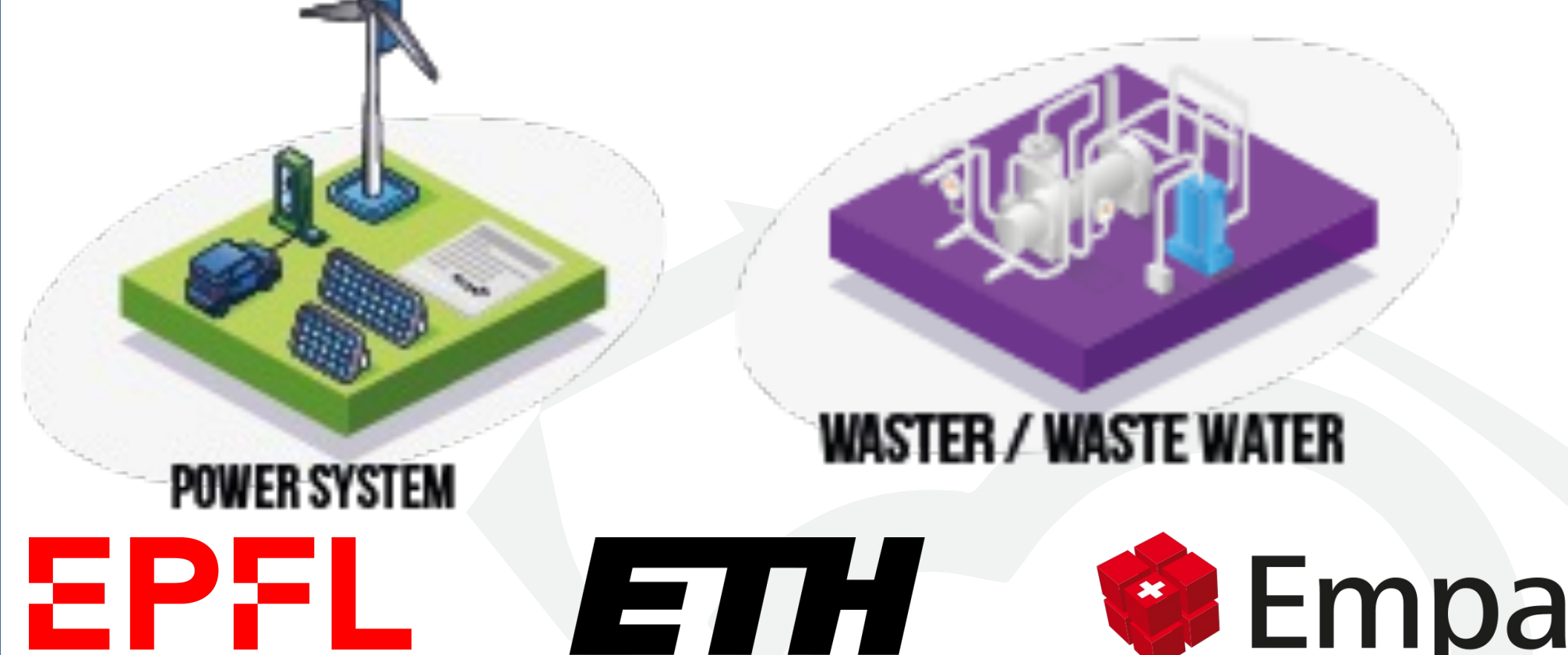
For the achievement of the project goal, 19 **partners** from EPFL, ETHZ, EAWAG, WSL and EMPA are involved. Furthermore, the UrbanTwin capitalizes on the network of 3 EPFL centers, notably CLIMACT, CIS, and EcoCloud.

Model development

WP2: Analyze and model socio-economic and environmental linkages with energy.



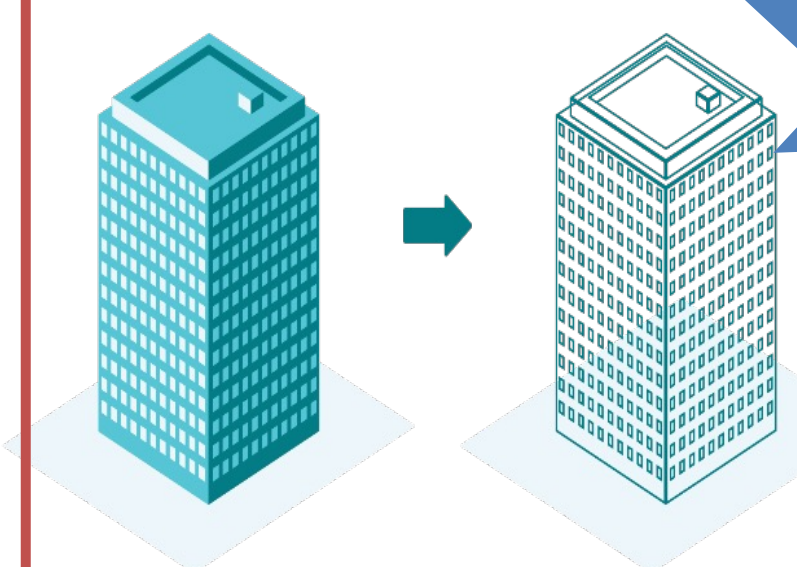
WP1: Analyze and model the joint management of energy and water for an urban demonstration site



Examples of Interrelationships Between Water and Energy (source: U.S. DOE (2006) Energy demands on water resources, U.S. Department of Energy_

Digital tool development

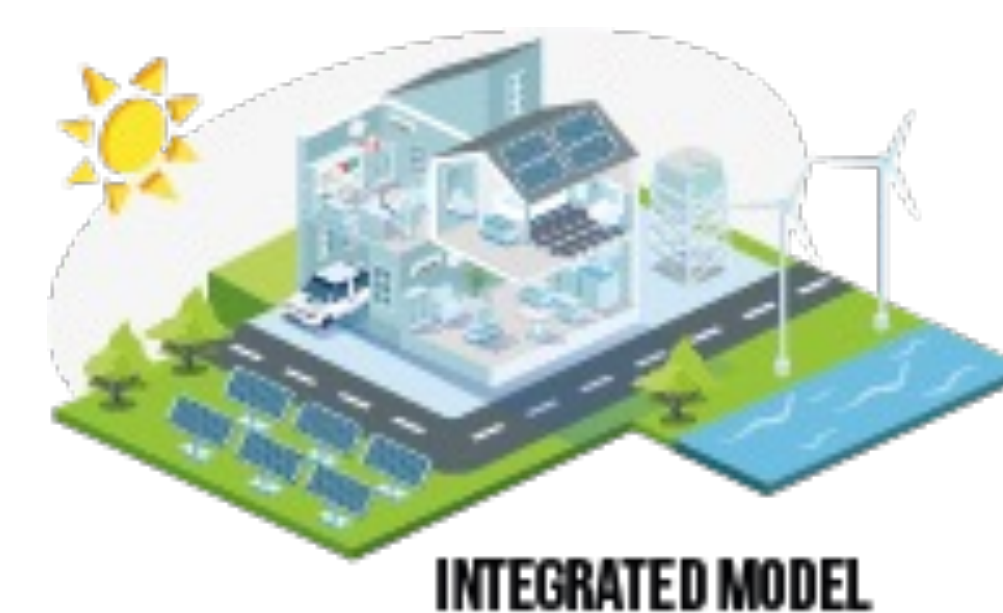
WP3: Develop edge-to-Cloud and multi-scale digital twin technologies for data gathering and application-driven optimization of resources



EPFL
ETH
Empa

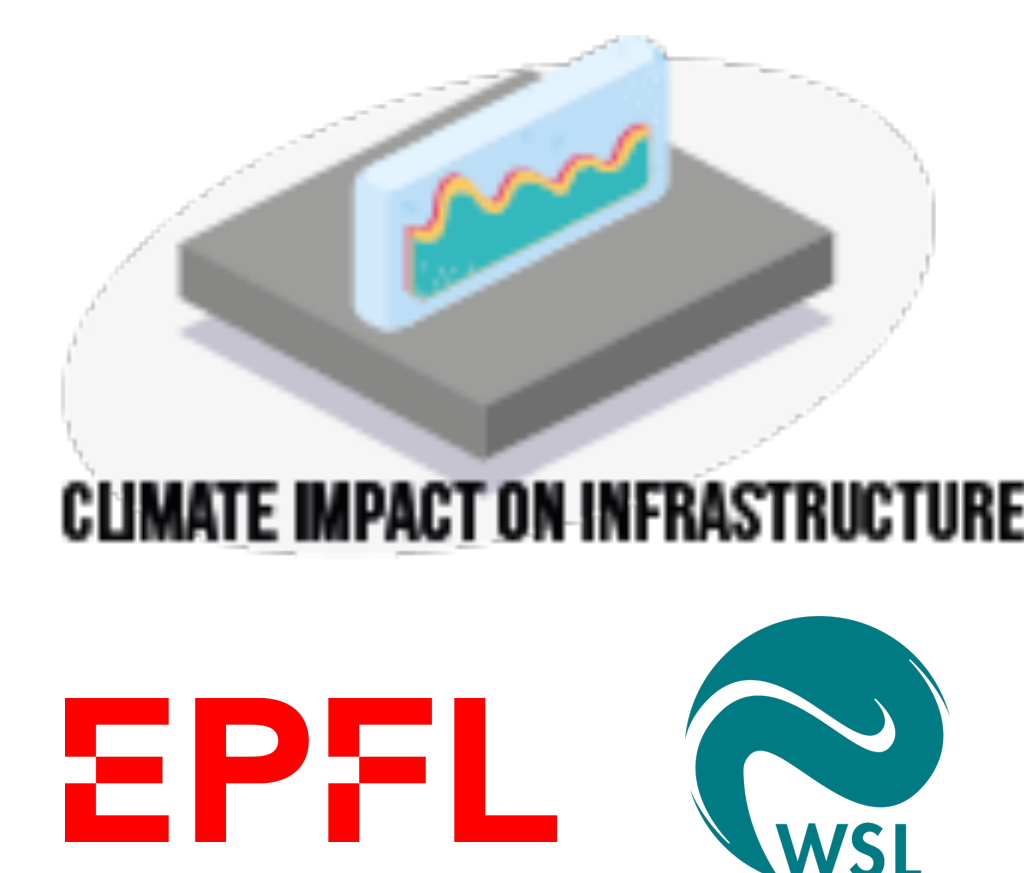
Case studies

WP4: Analyze the RE potential and the promotion of efficient energy use in local communities



EPFL ETH

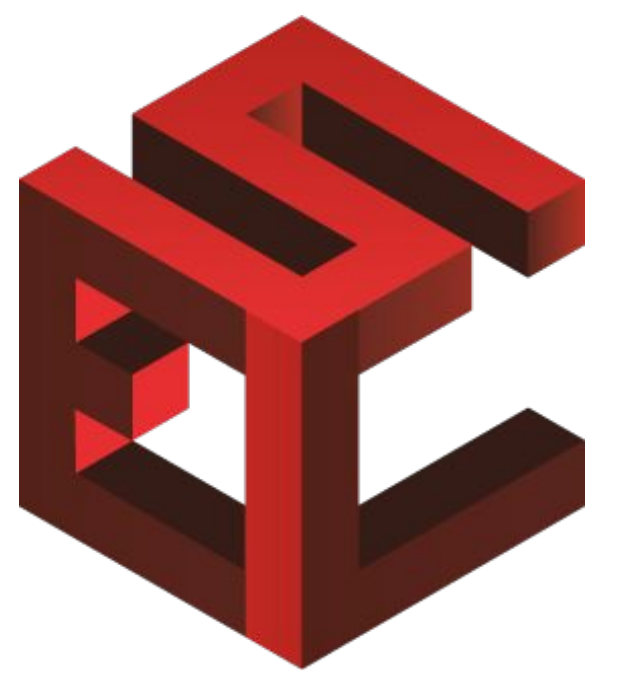
WP5: Assessment of climate impact on infrastructure and pathway feasibility



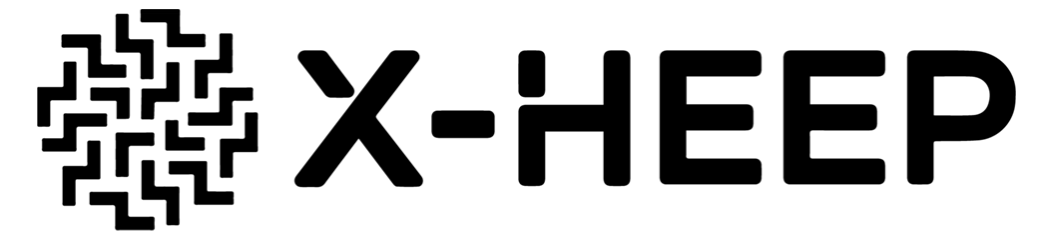
ACKNOWLEDGEMENT

This research is carried out in the frame of the “UrbanTwin: An urban digital twin for climate action: Assessing policies and solutions for energy, water and infrastructure” project with the financial support of the ETH-Domain Joint Initiative program in the Strategic Area Energy, Climate and Sustainable Environment

X-HEEP: An Open-Source, Configurable and Extendible RISC-V Microcontroller for the Exploration of Ultra-Low-Power Edge Accelerators

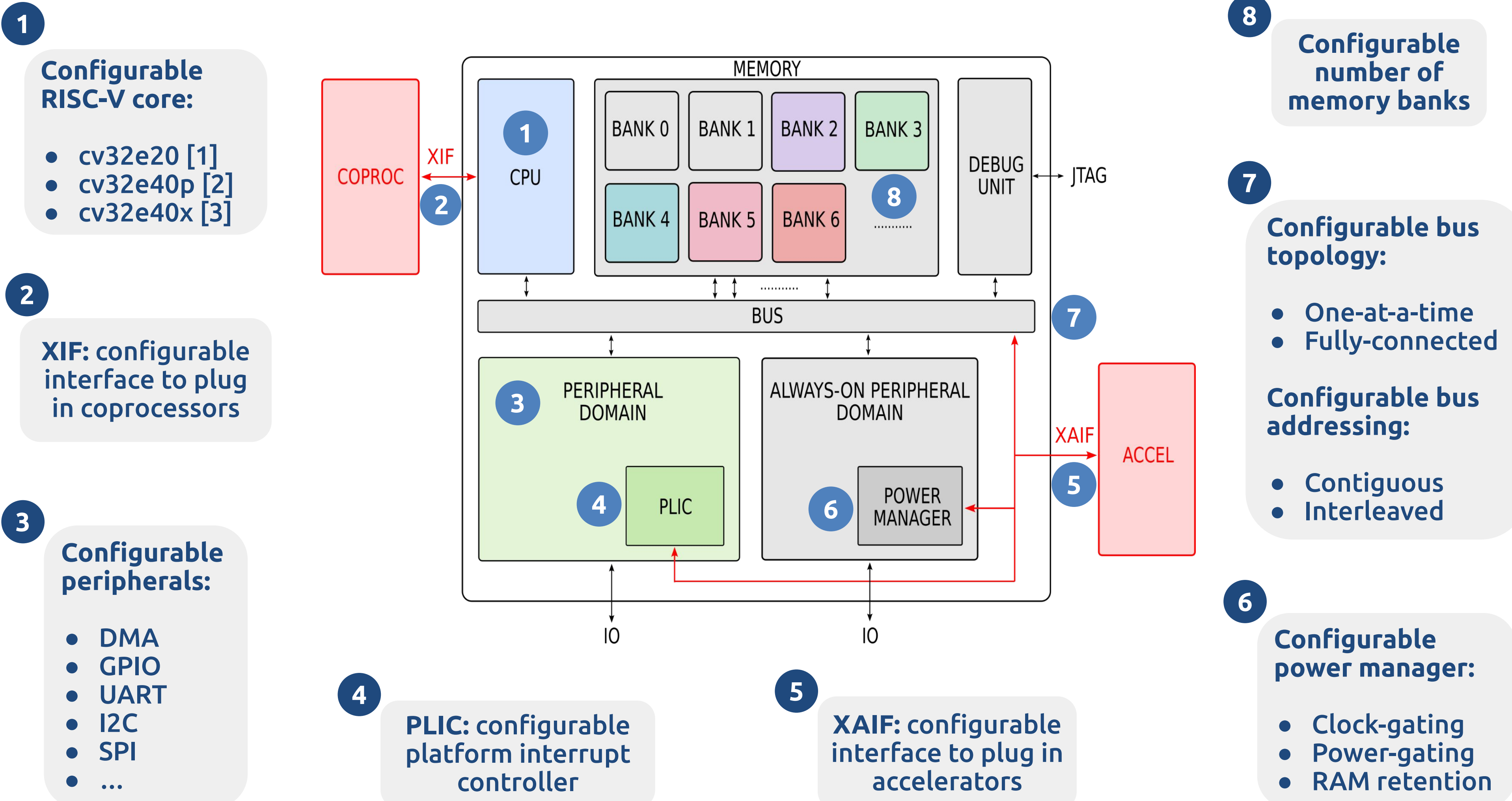


Simone Machetti, Pasquale Davide Schiavone, Thomas Christoph Müller, Miguel Peón-Quirós, David Atienza



X-HEEP: eXtensible Heterogeneous Energy Efficient Platform

Configurable and Extendible Architecture



[1] Low-power, 32-bit, in-order RISC-V core with a 2-stage pipeline that implements RV32[EI][M]C instruction set architecture.
 [2] High-performance, 32bit, in-order RISC-V core with a 4-stage pipeline that implements RV32IMFC[Xpulp] instruction set architecture.
 [3] High-performance, 32-bit, in-order RISC-V core with a 4-stage pipeline that implements the RV32[I,E][M]Zmmul[A]Zca_Zcb_Zcmp_Zcmt[Zba_Zbb_Zbs]ZictrZihpmZicsrZifencei[X] instruction set architecture and offers extensions through the CORE-V-XIF interface.

Design Flow Support

CONFIGURATION → SystemVerilog templates

Automatic tools to configure the RTL code of X-HEEP thanks to customizable parameters

SW IMPLEMENTATION → HAL and FreeRTOS

- Includes an hardware abstraction layer (HAL) for easy access to peripherals' functionalities
- Supports FreeRTOS to enable concurrent execution and enhance resource management

SIMULATION AND HW IMPLEMENTATION → FuseSoC

Automatic tool to generate the required script to simulate and implement the design

Supported tools:

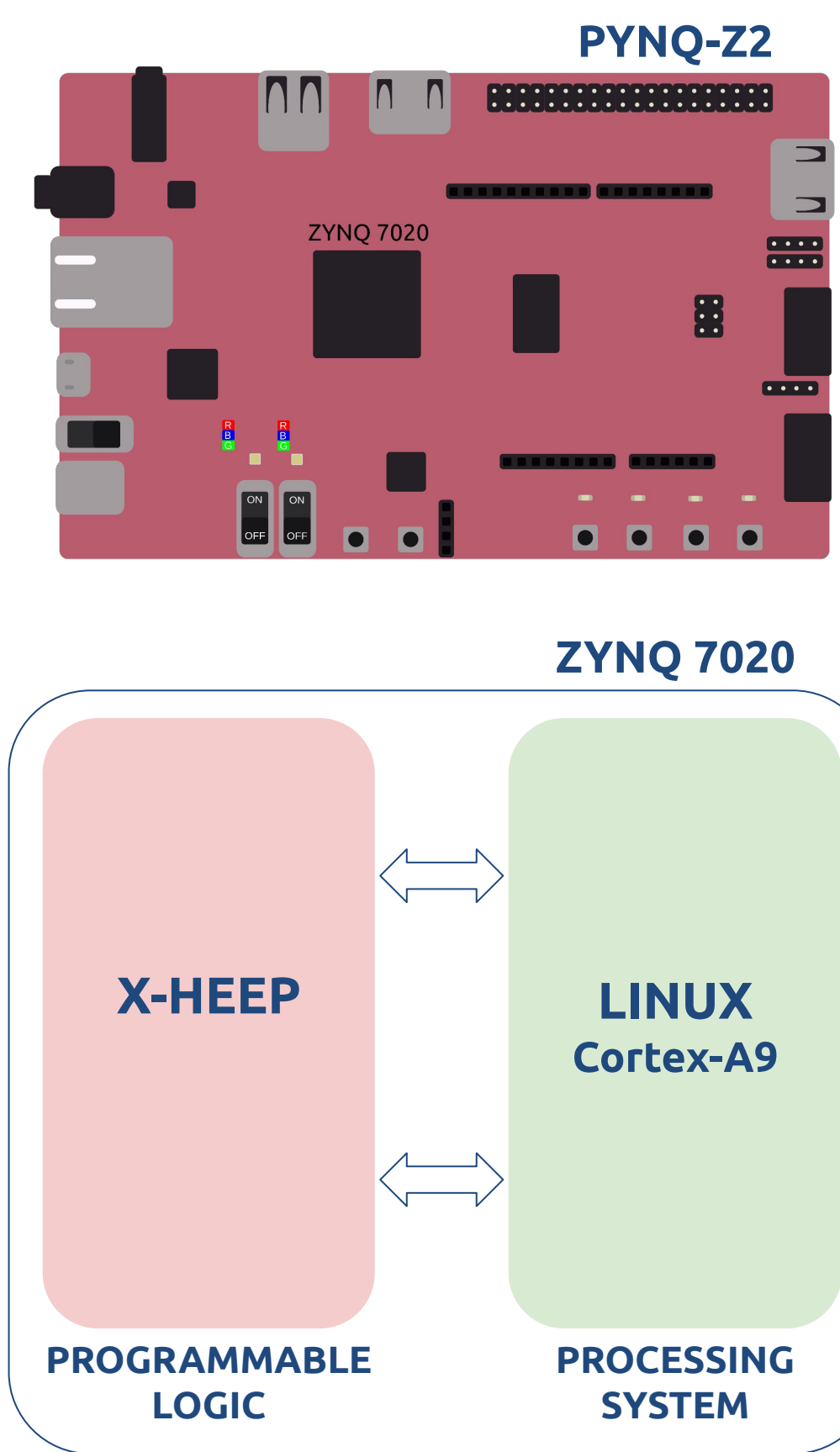
- Simulation:** Verilator, Questasim, etc.
- ASIC:** Design Compiler, Innovus, etc.
- FPGA:** Vivado, etc.

FPGA: Fully-virtualized Platform for Healthcare Applications

Useful Links



Implementation and Features



IMPLEMENTATION → Zynq 7020 on Pynq-Z2 board

X-HEEP implemented on the programmable logic side communicates with the Linux system on the processing system side

FULL VIRTUALIZATION → No off-board I/O connectivity required

All peripherals virtualized under Linux

Examples: JTAG virtualized towards Linux GPIOs, UART virtualized towards a Linux serial port, FLASH virtualized on the board DRAM, etc.

ESTIMATIONS → Performance and energy

- Performance estimation through dedicated performance counters
- Energy estimation based on TSMC 65nm CMOS technology

USER INTERACTION → Jupyter notebook

Dedicated Python class to support interaction through Jupyter notebook

Open-source Repositories

X-HEEP



FPGA-HW



FPGA-SW



X-HEEP → Main repository of X-HEEP

<https://github.com/esl-epfl/x-heep>

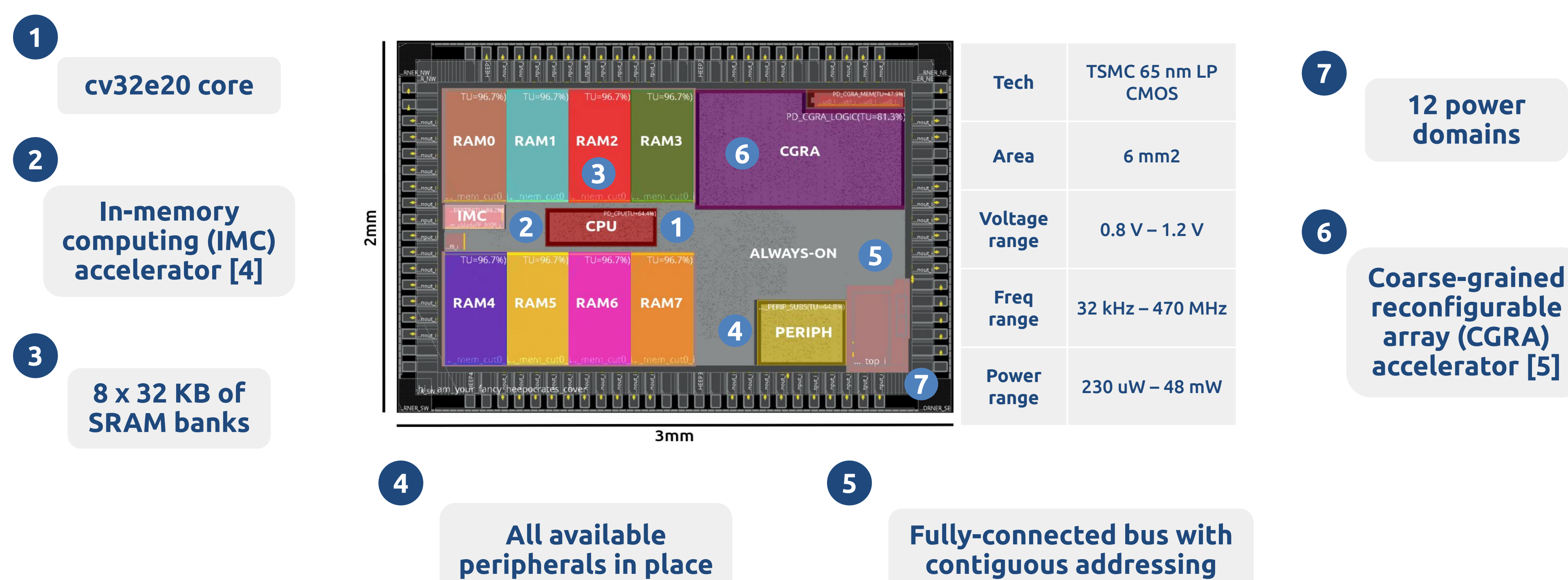
FPGA → Repositories of the FPGA emulation platform

<https://github.com/esl-epfl/x-heep-femu>

<https://github.com/esl-epfl/x-heep-femu-sdk>

HEEPOCRATES: X-HEEP-based Chip for Healthcare Applications

Post-PnR Layout and Silicon Characterization

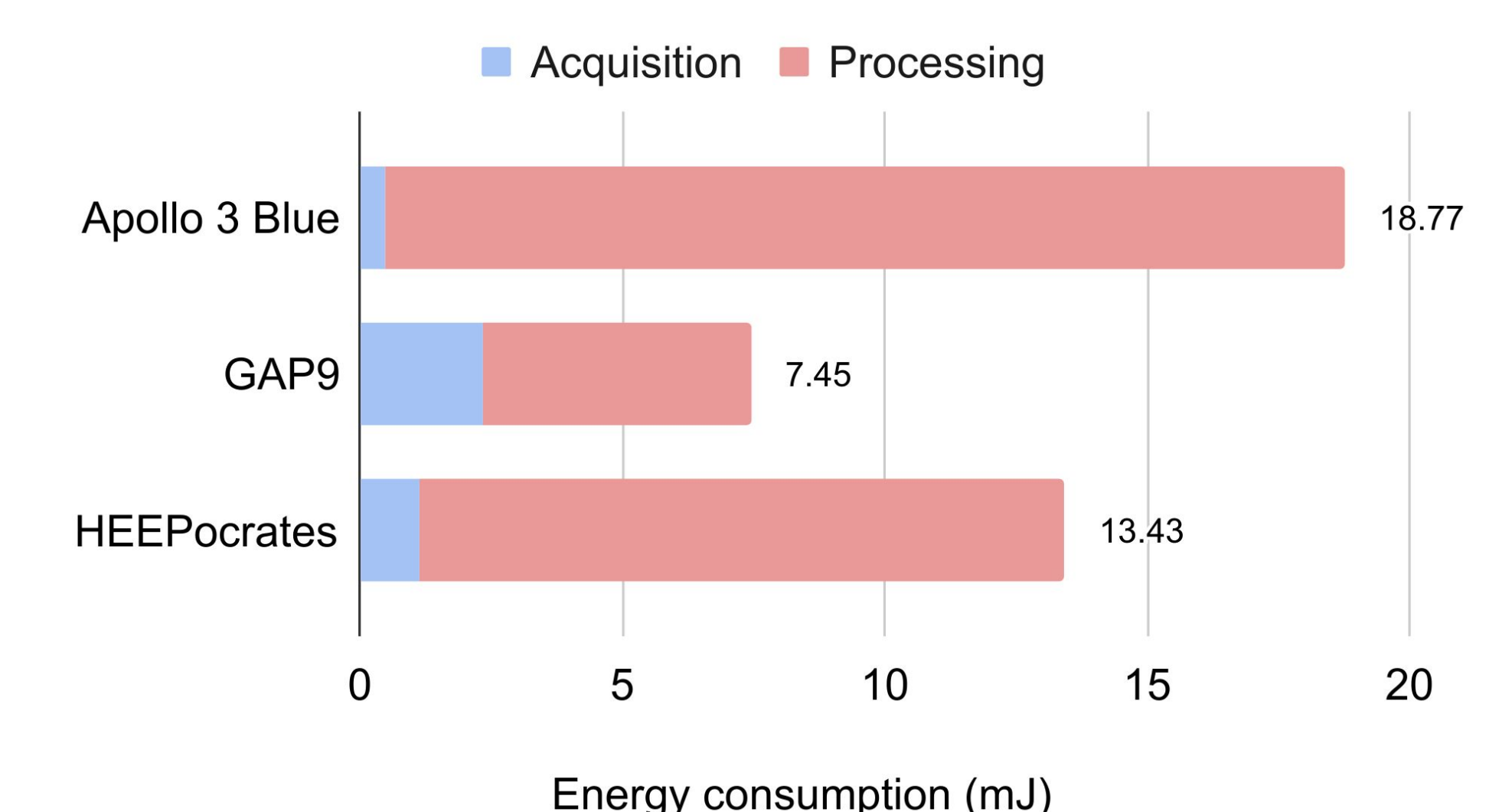


[4] William Andrew Simon, Yasir Mahmood Qureshi, Marco Rios, Alexandre Levisse, Marina Zapater, and David Atienza. 2020. Blade: an in-cache computing architecture for edge devices. IEEE Transactions on Computers, 69, 9, 1349–1363.
 [5] Loris Duch, Soumya Basu, Rubén Braojos, David Atienza, Giovanni Ansani, and Laura Pozzi. 2016. A multi-core reconfigurable architecture for ultra-low power bio-signal analysis, 416–419.

Energy Consumption

APPLICATION → EEG Seizure Detection [6]

Convolutional neural network (CNN) used to detect seizures in electroencephalography (EEG) signals.

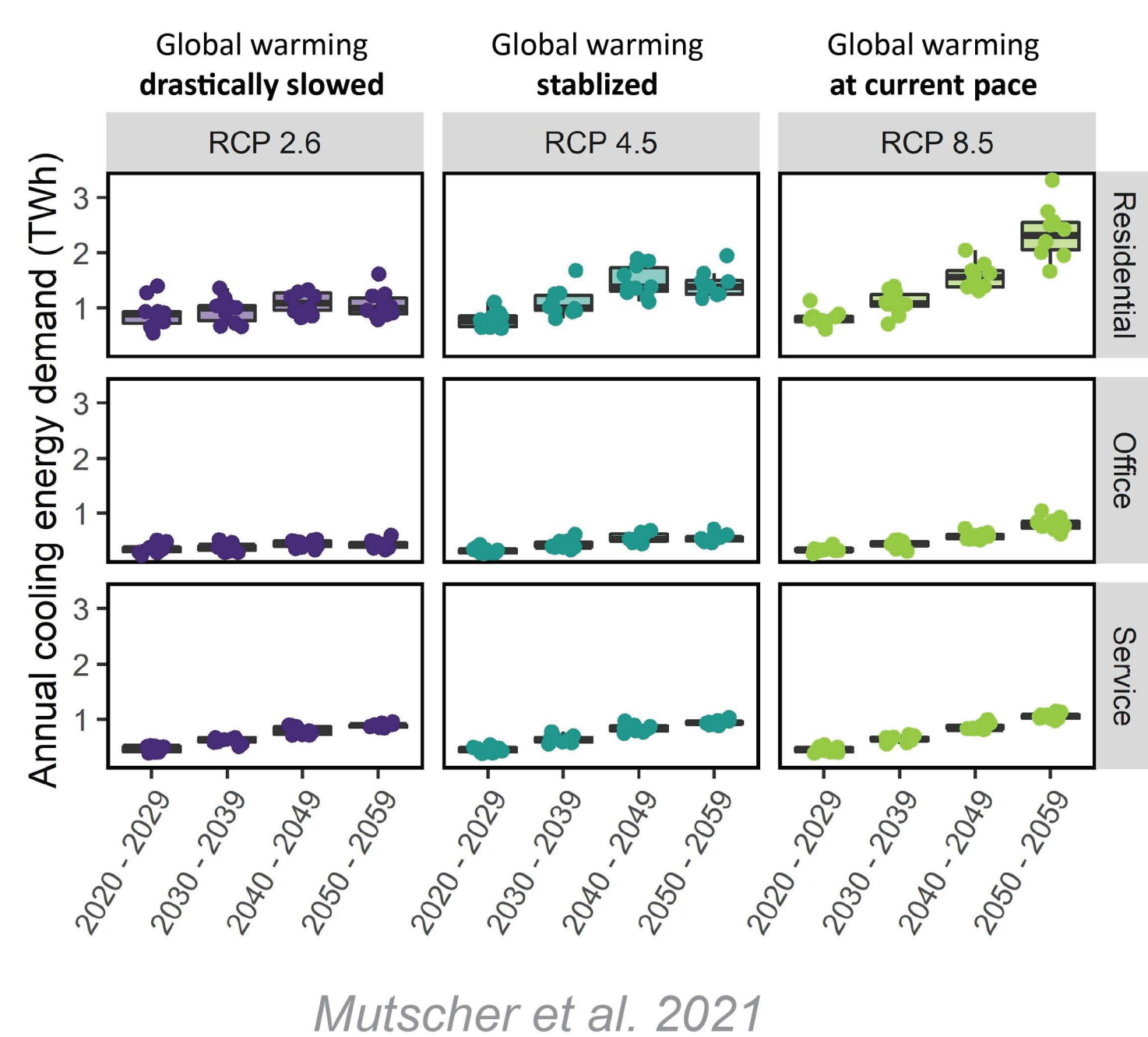


[6] Catalina Gómez, Pablo Arbeláez, Miguel Navarrete, Catalina Alvarado-Rojas, Michel Le Van Quyen, and Mario Valderrama. 2020. Automatic seizure detection based on imaged-eeeg signals through fully convolutional networks. Scientific reports, 10, 1, 1–13.

Gautier Rouaze, Zhengmao Lu
Energy Transport Advances Laboratory (ETA-Lab), EPFL, Lausanne

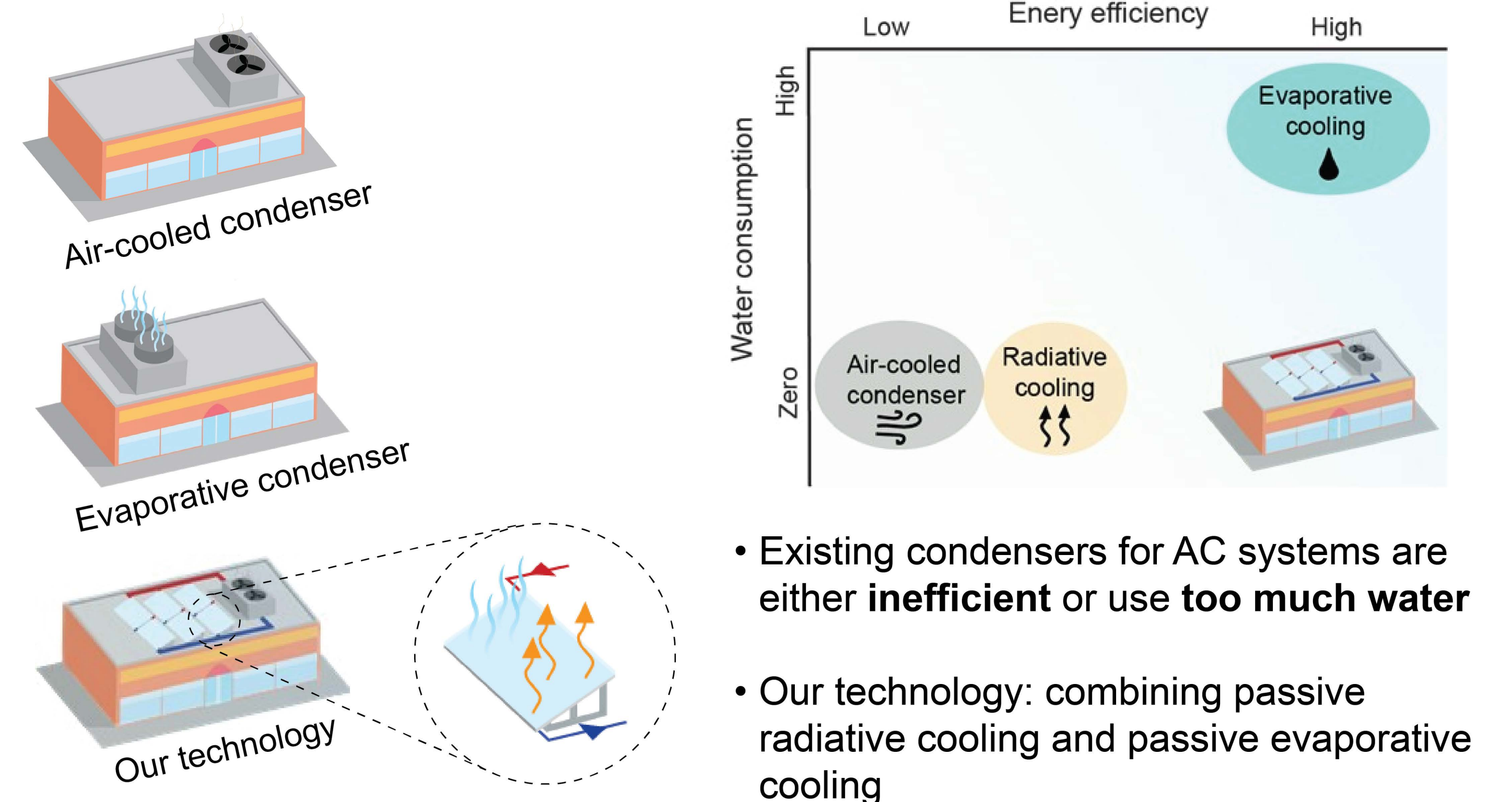
Motivations

Cooling trends in Switzerland

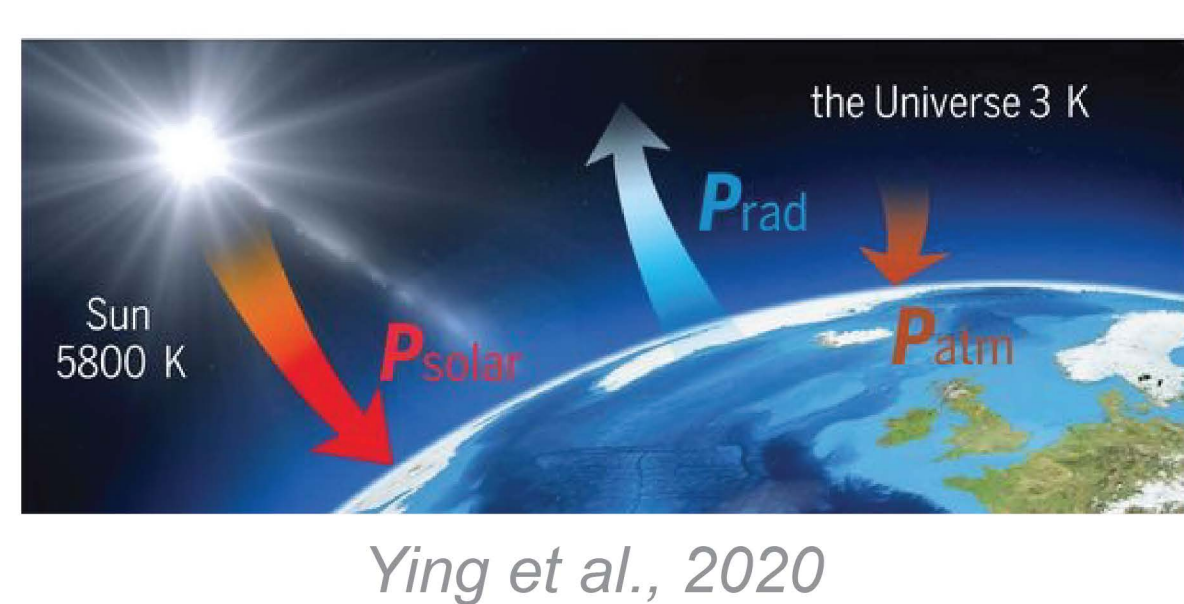


- **Air conditioning technology uptake** will significantly contribute to future cooling demands
- Need to “**limit cooling energy demand impacts** on the future energy systems particularly in countries with **temperate climates**”

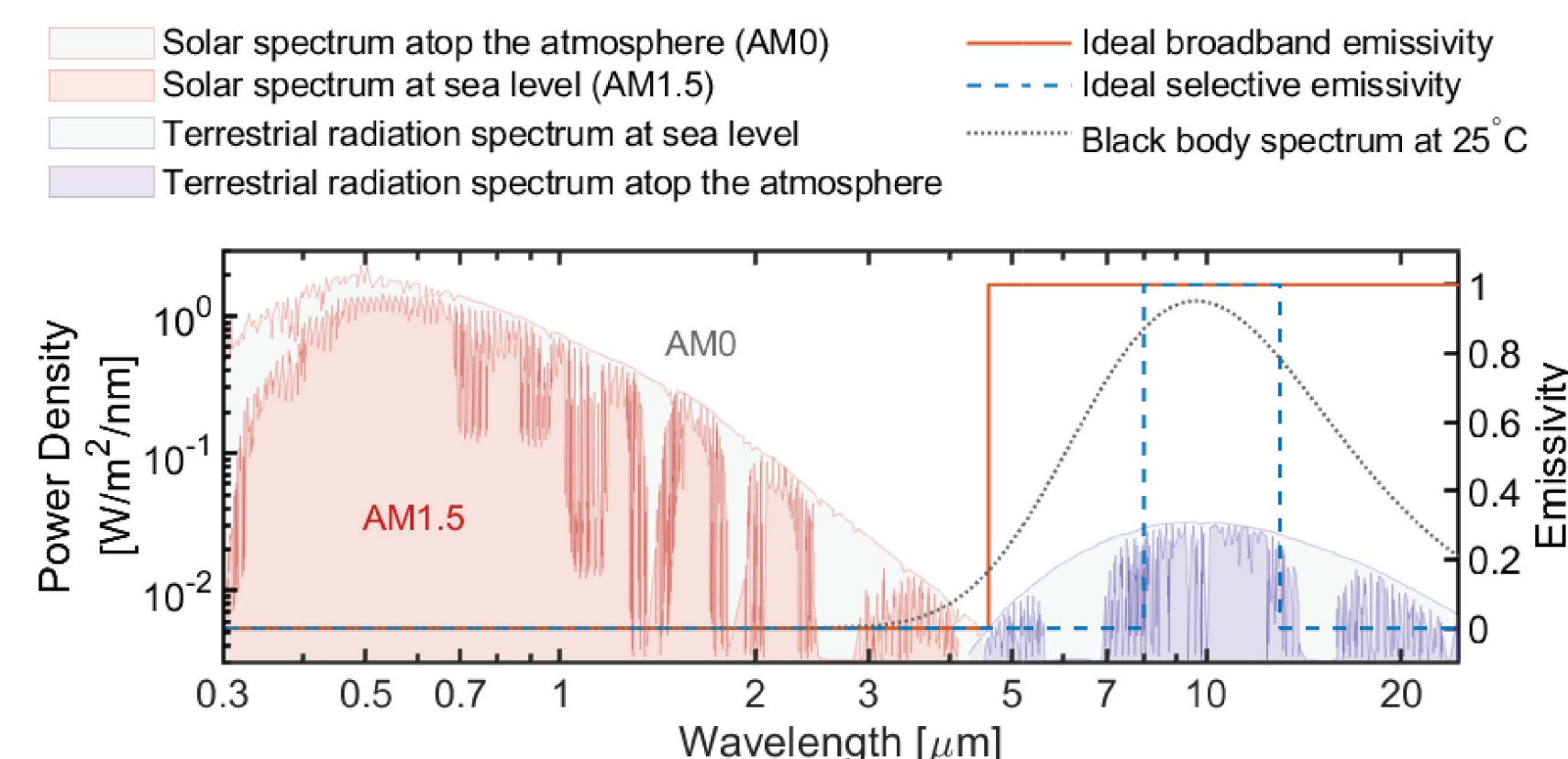
State of the Art / Opportunity space



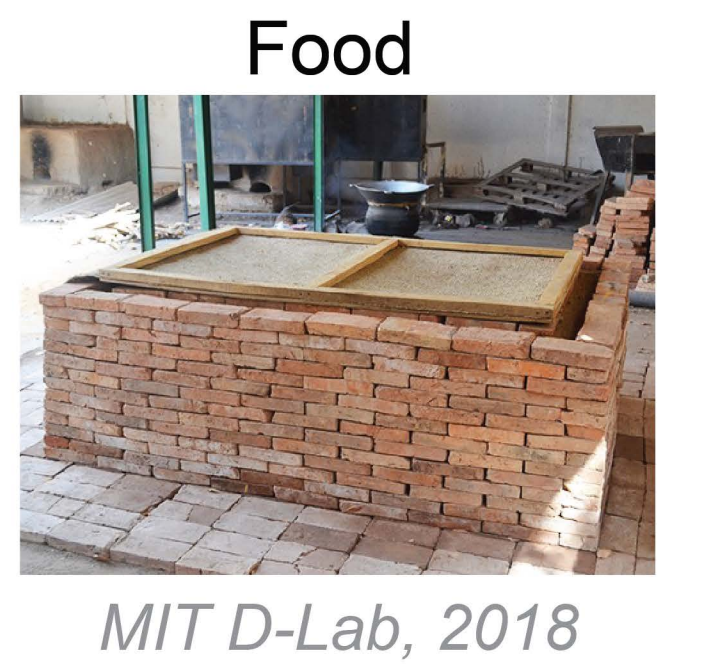
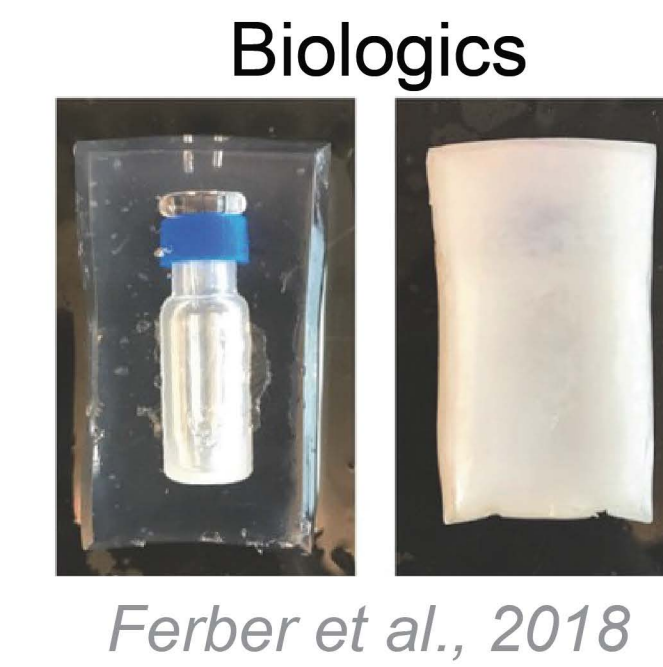
Passive Radiative Cooling



Challenges: Max. cooling power ($<120 \text{ W/m}^2$), parasitic heating, RH

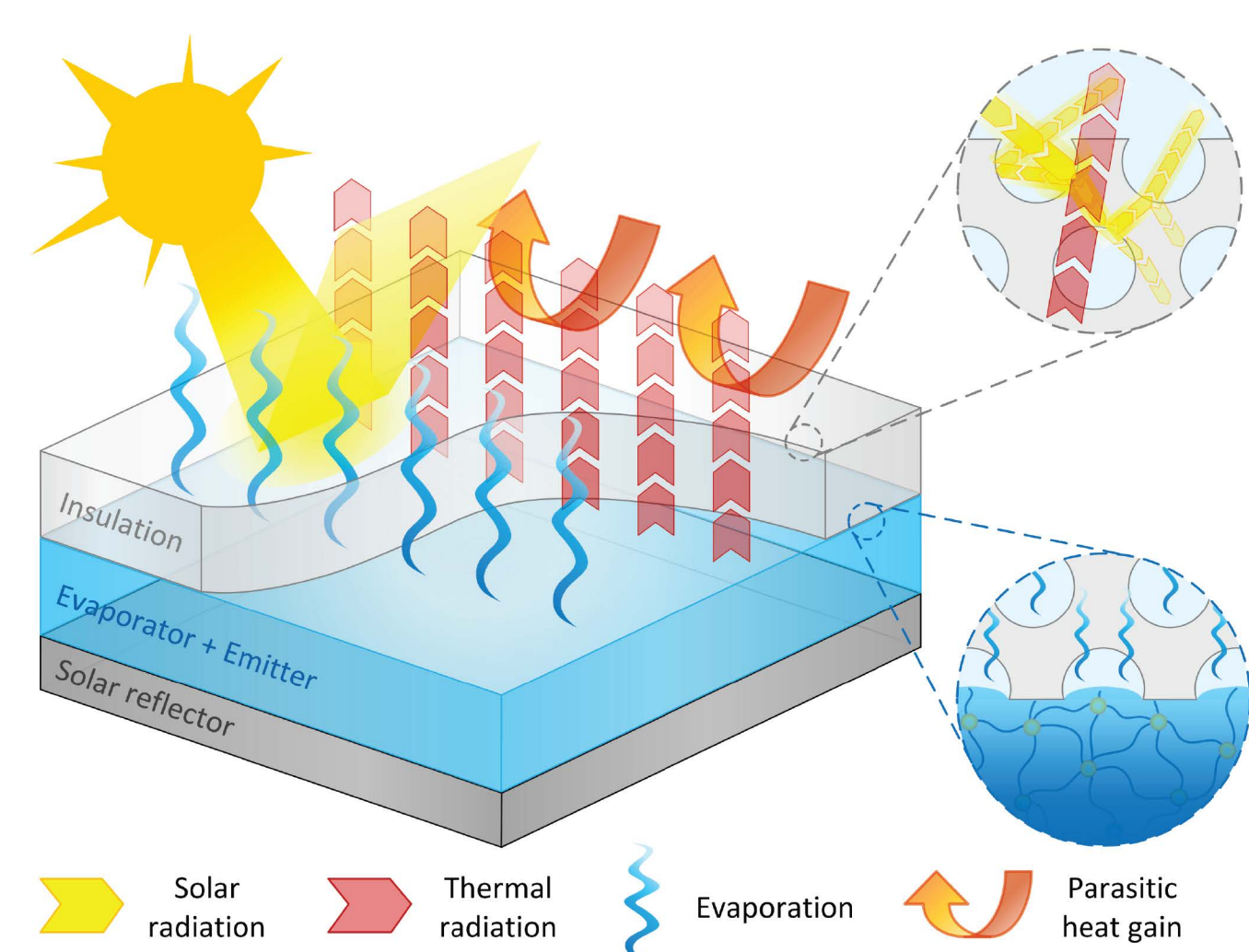


Passive Evaporative Cooling



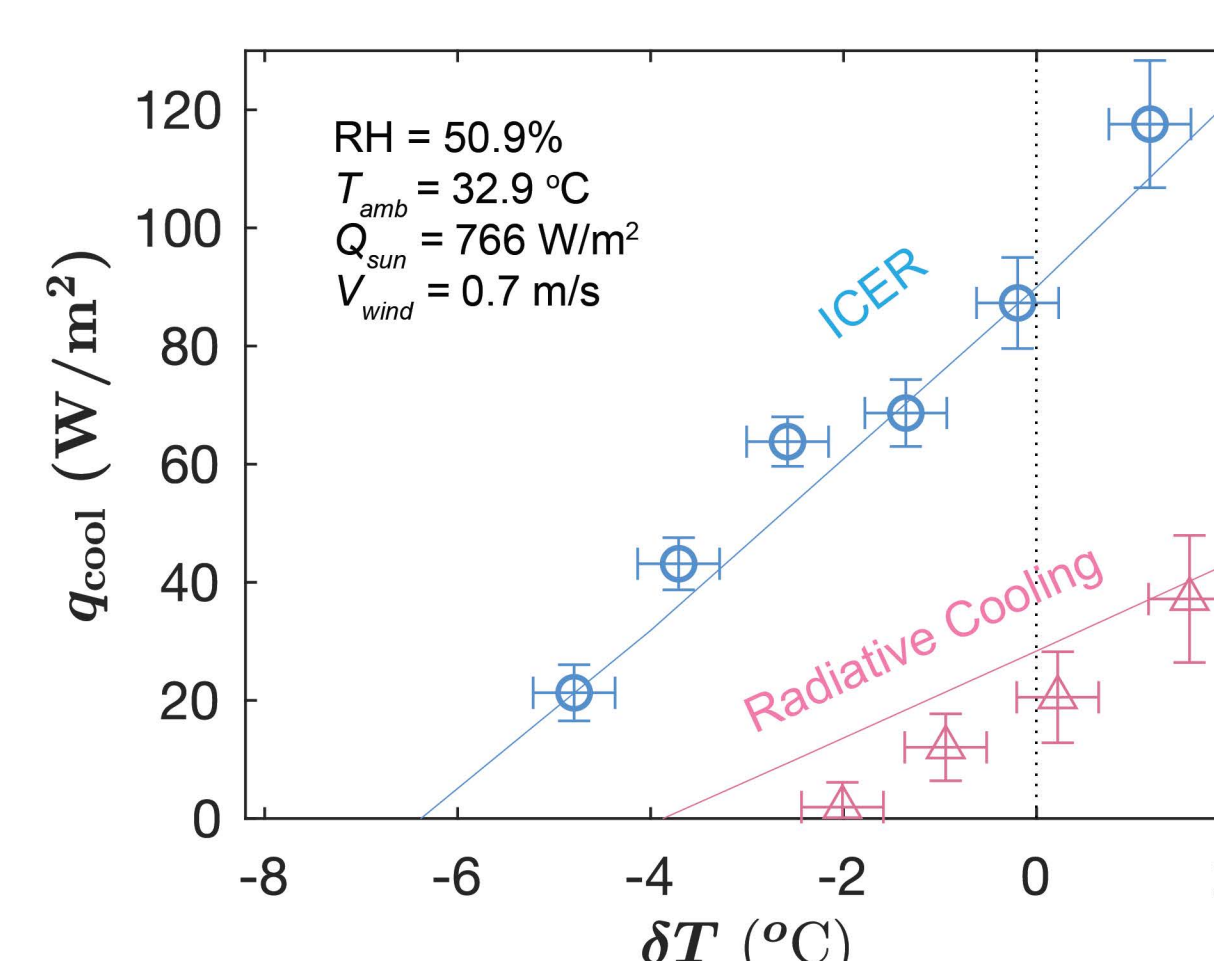
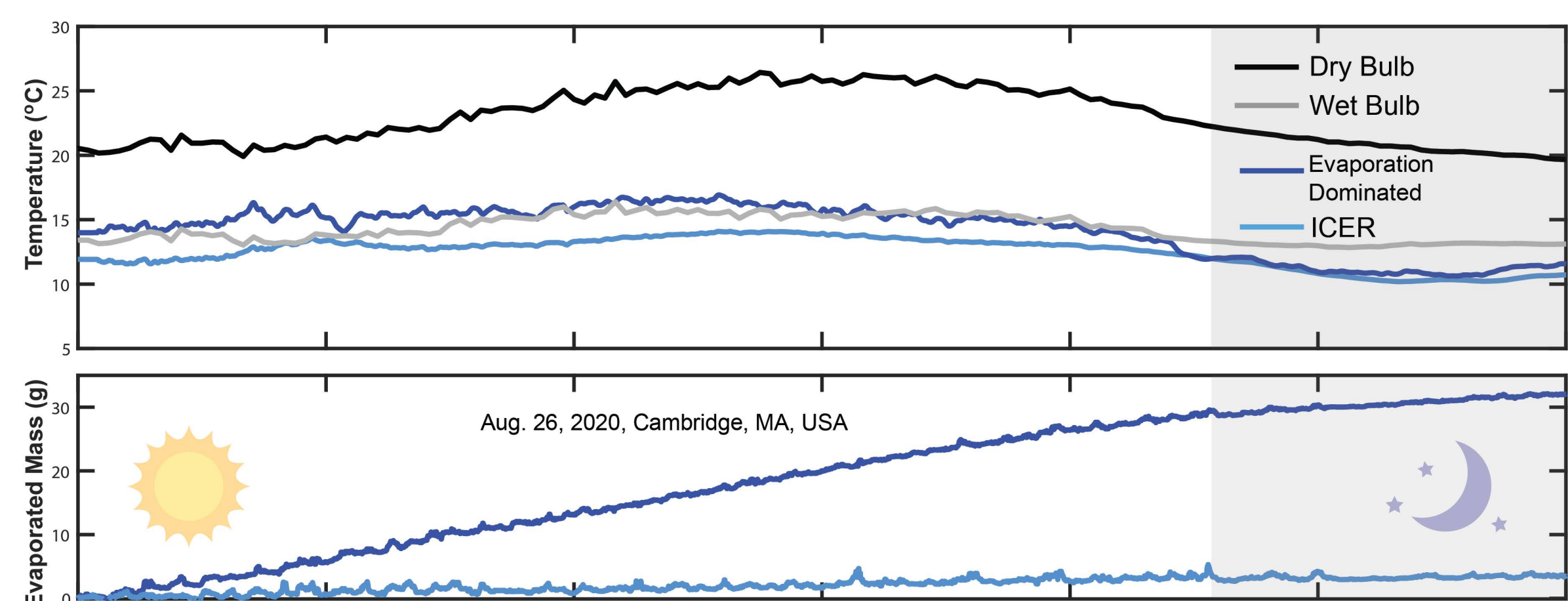
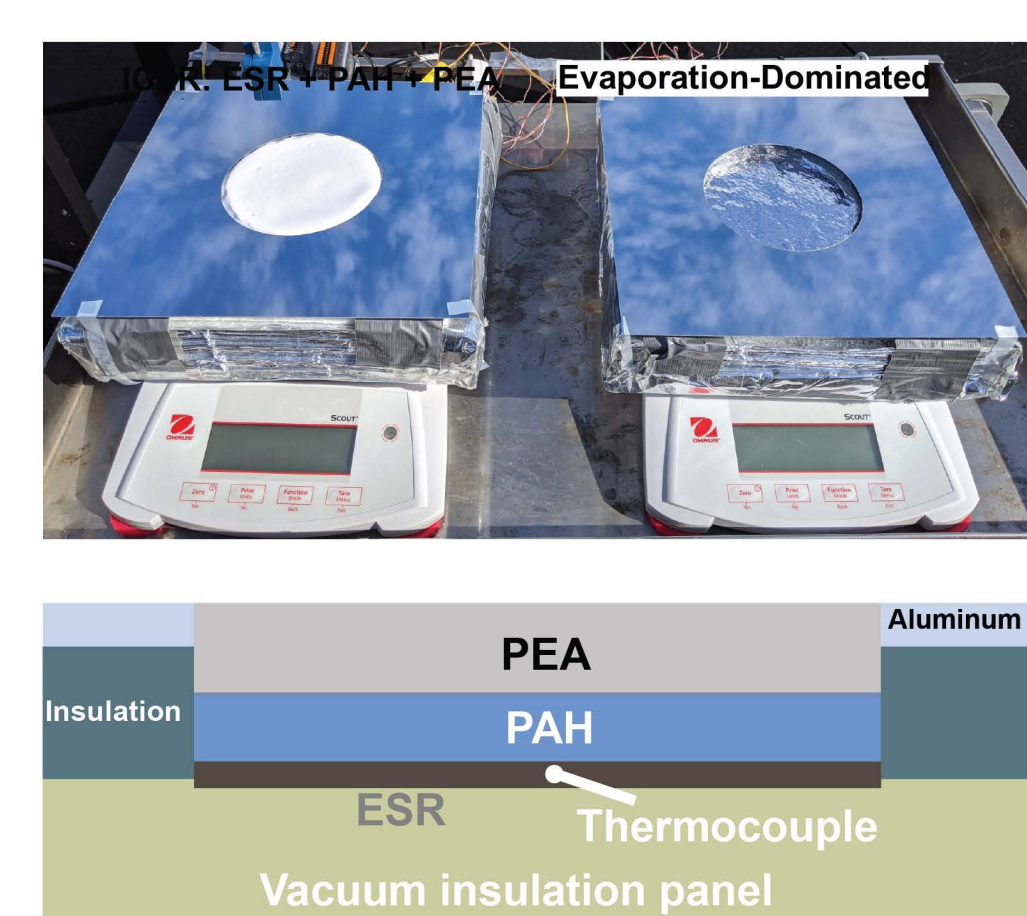
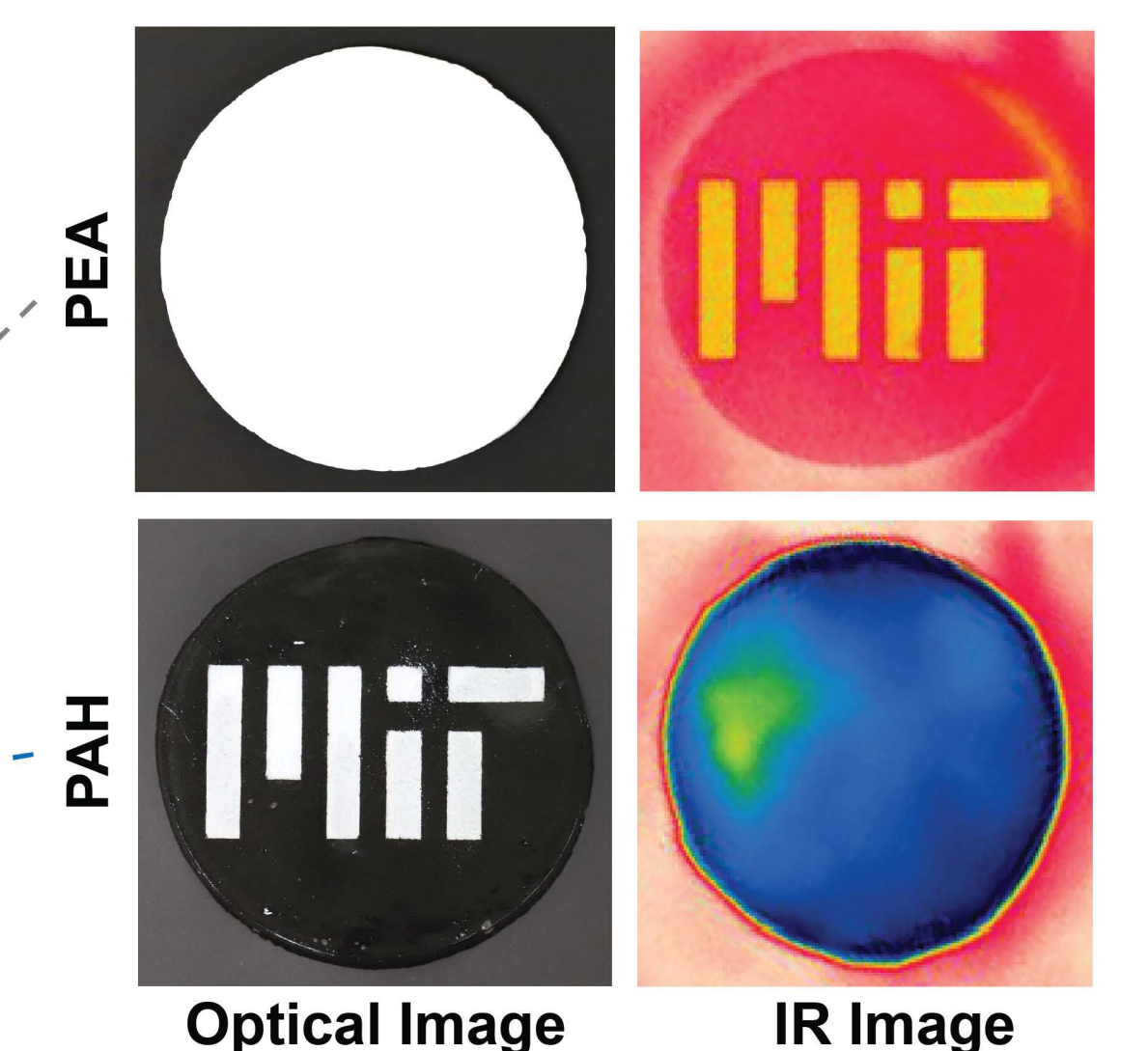
Challenges: Water consumption, environmental heating, RH

Previous work: Insulated Cooling with Evaporation and Radiation (ICER)



- Insulation layer (top)**
Low thermal conductance, high vapor permeance
High solar reflectance, high IR-transmittance
- Evaporation layer (middle)**
High water capacity, high IR-emittance
- Solar reflection layer (bottom)**
High solar reflectance (3M ESR film)

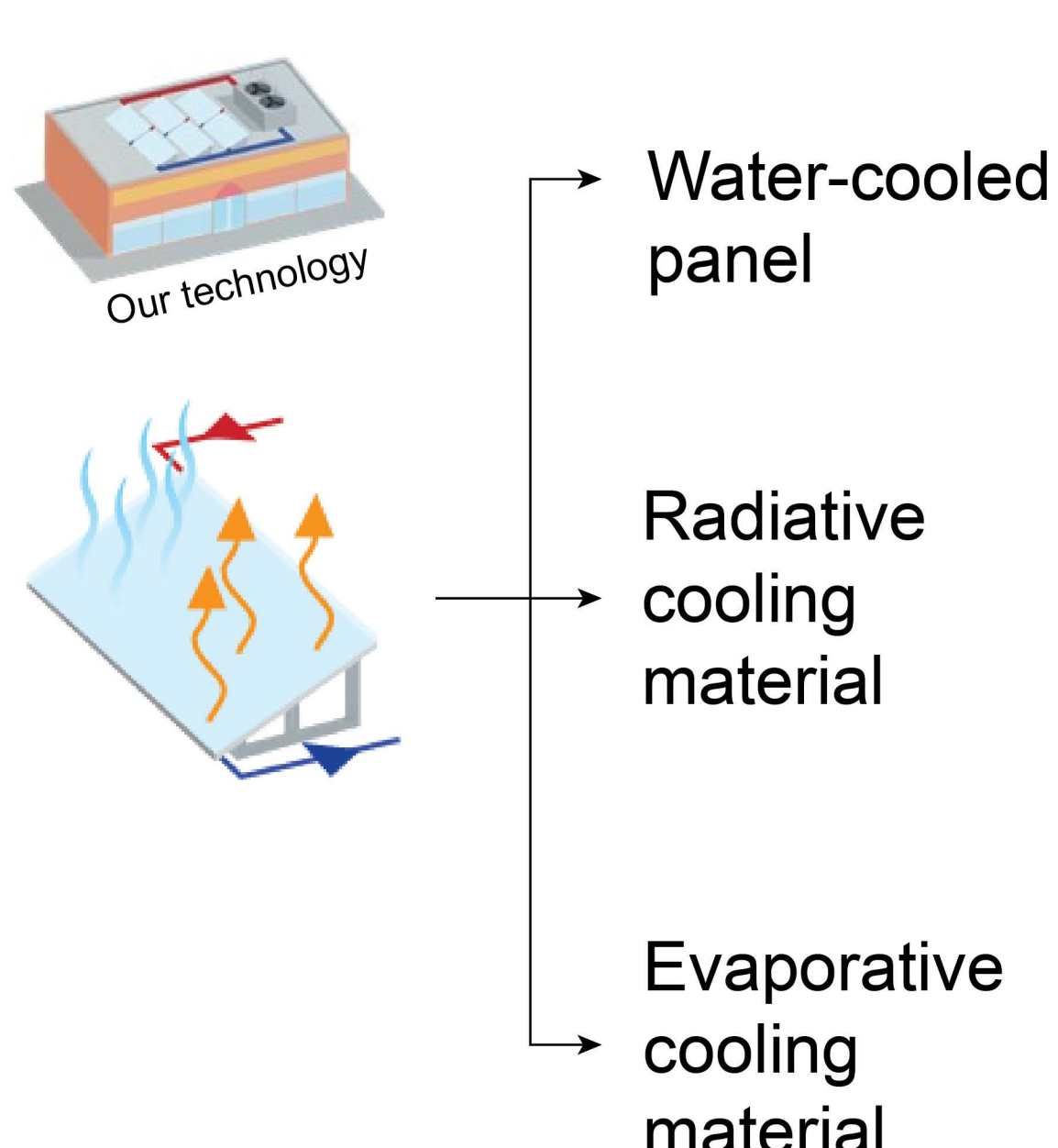
Layer	Material	Visible spectrum	IR spectrum
Insulation	Polyethylene aerogel (PEA)	Opaque	Transparent
Evaporation	Polyacrylamide hydrogel (PAH)	Transparent	Opaque



- -10°C below ambient temperature
- **300%** enhancement in daytime cooling power over radiative cooling
- High cooling power despite **high air density and high RH**

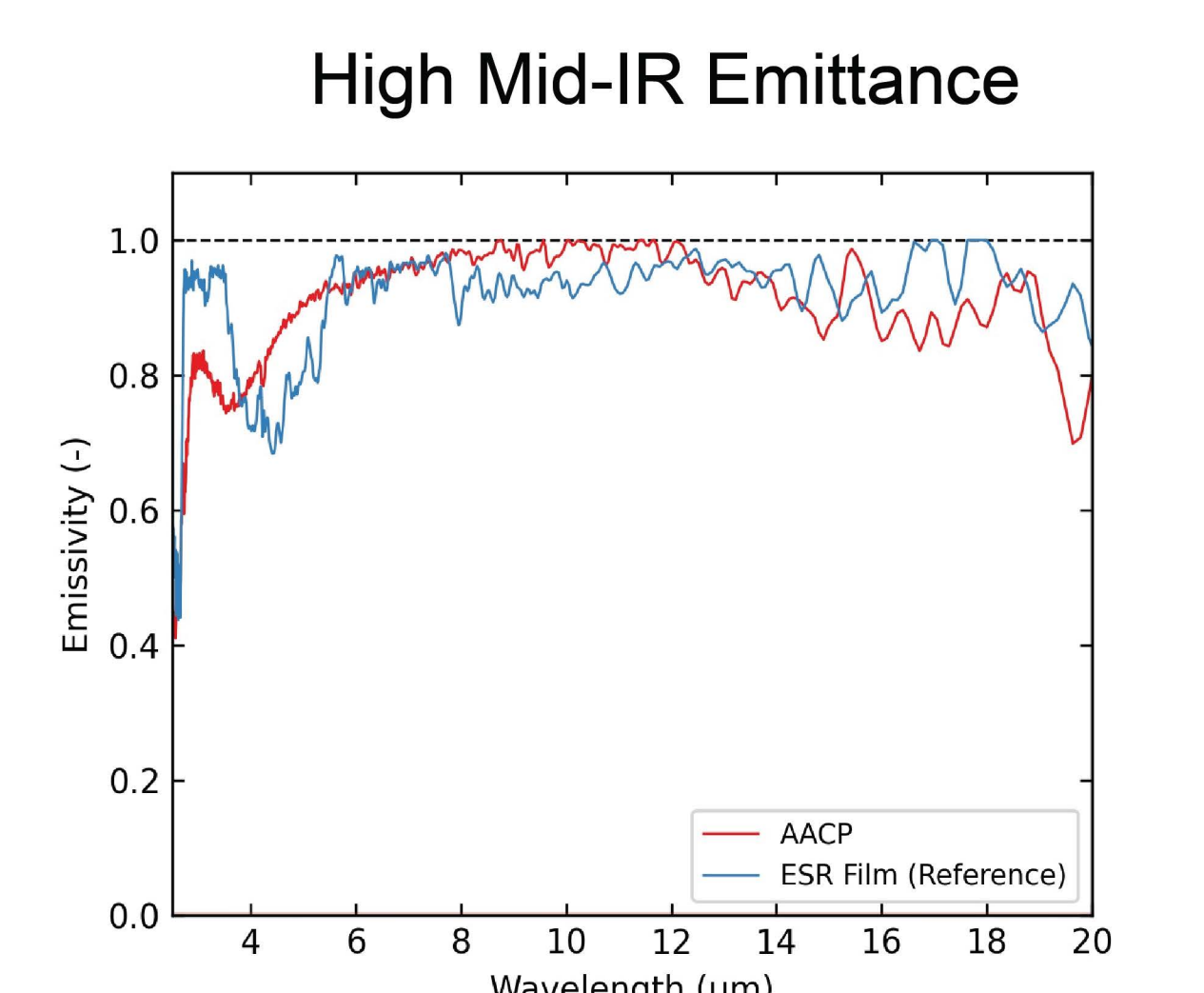
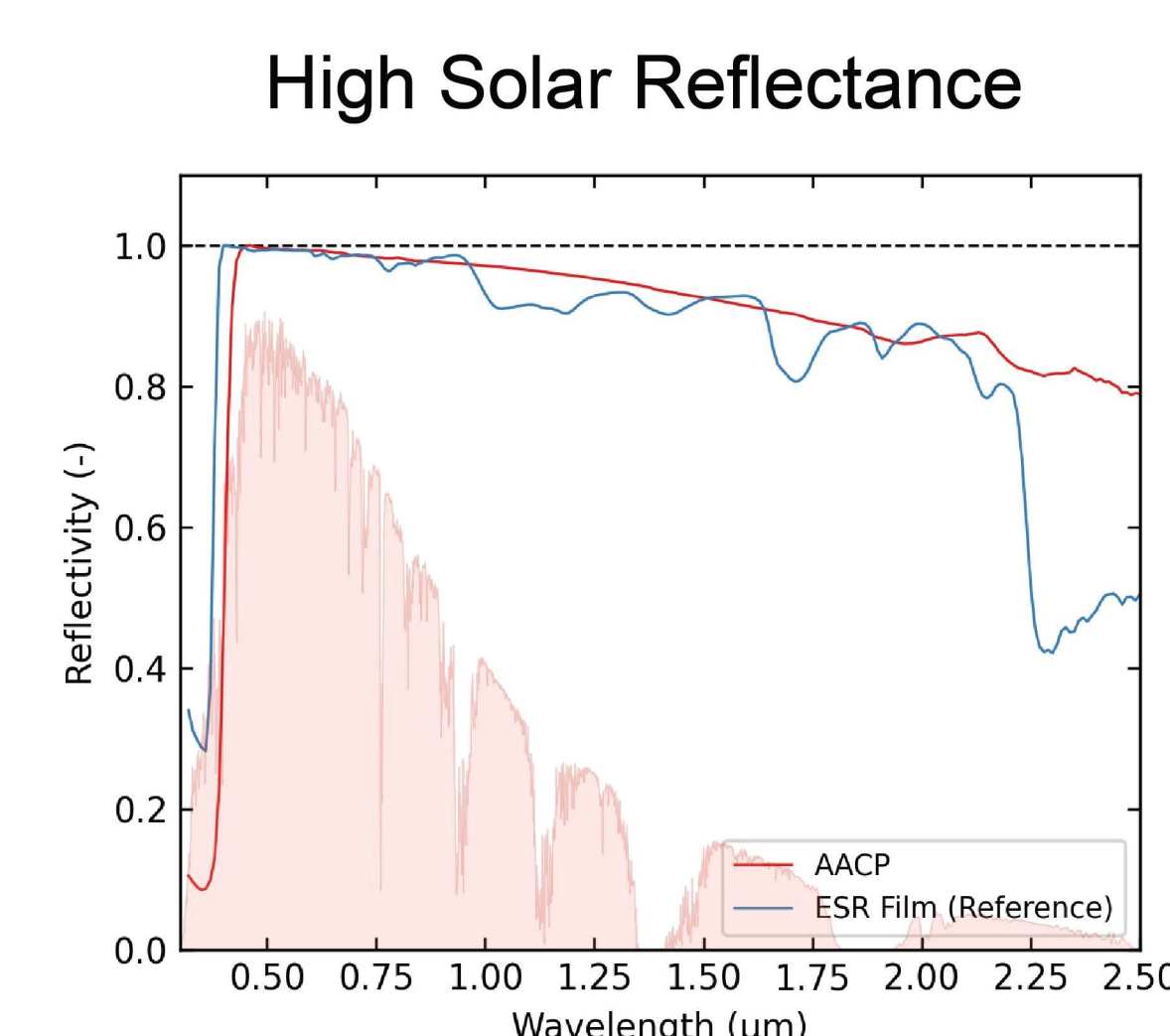
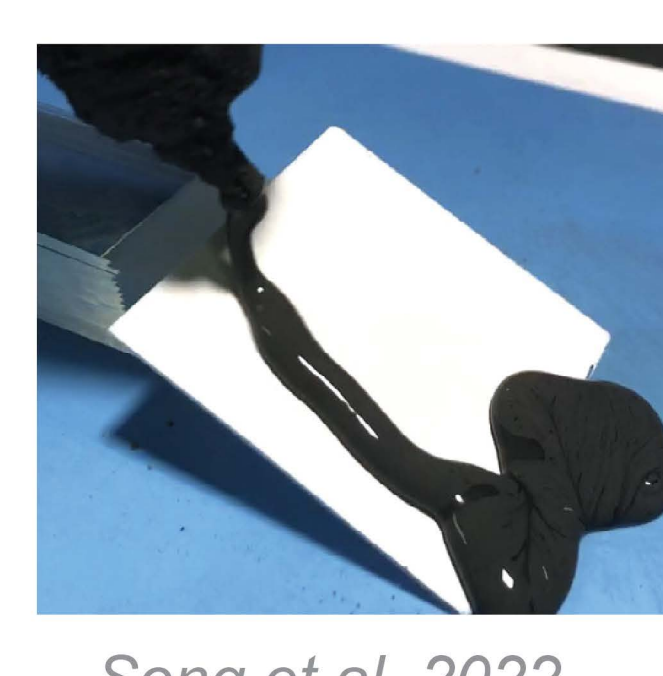
Published in Cell Reports Physical Science 3, 101068

Ongoing work: Cold Water Generation with Evaporative Radiative Cooling Panels



Suggested radiative cooling material

- **Anti-aging cooling paint (AACP)**
 - High solar reflectivity (TiO₂ Nanoparticles light scattering)
 - High emittance in the LWIR
 - Reduced NIR absorbance
 - Increased UV resistance
 - Hydrophobic coating (anti fouling)



Material synthesized at ETA-Lab