AI and Control
Opportunities to tame usage of resources

Sophie Cerf

sophie.cerf@inria.fr

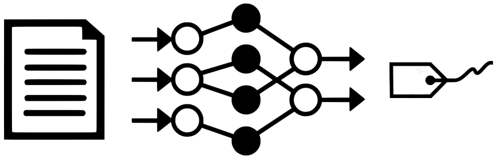
Université de Lille

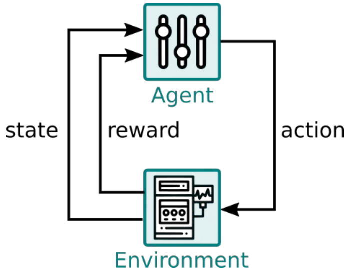
Inria

CRIStAL
Centre de Recherche en Informatique,
Signal et Automatique de Lille

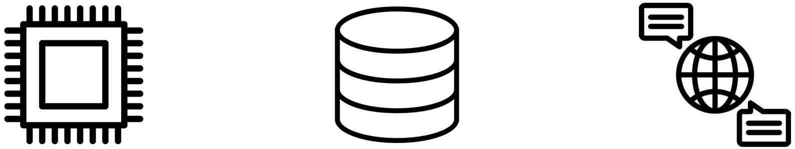
My vision of AI

- Used to make predictions & take actions



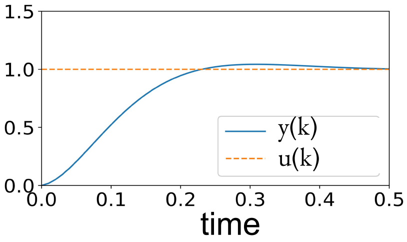


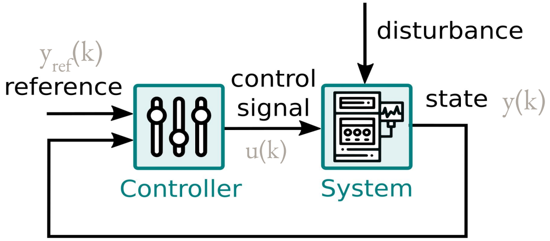
- Limits: data intensive, use of resources



Control Theory

*Field that study and control **dynamical** systems*





- System's model

$$y(k+1) = ay(k) + bu(k)$$

- Controller

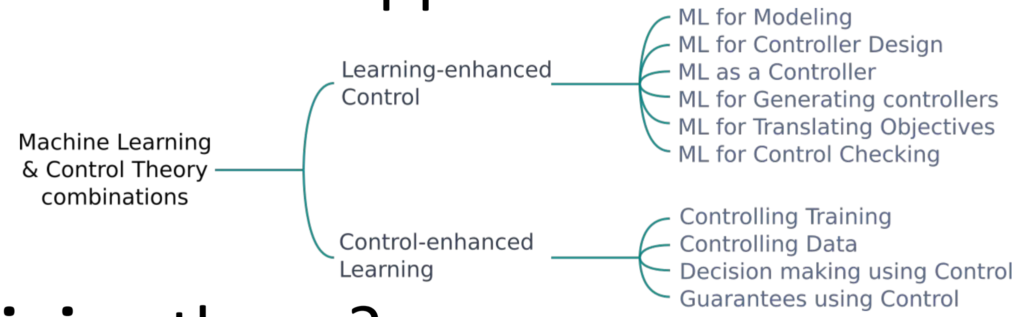
$$u(k) = K[y_{ref}(k)-y(k)]$$

- Objectives among stabilisation, tracking, optimization, etc.
- Some interesting characteristics
  - the Feedback principle
  - Guaranteed behavior
  - Model-based approach

Quentin Guilloteau, Sophie Cerf, Raphaël Bleuse, Bogdan Robu, Eric Rutten. Under Control: A Control Theory Introduction for Computer Scientists. *ACSOS 2024 - 5th IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS 2024)*, Sep 2024, Aahrus, Denmark. pp.1-10. ⟨hal-04666859⟩

AI ⇄ Control

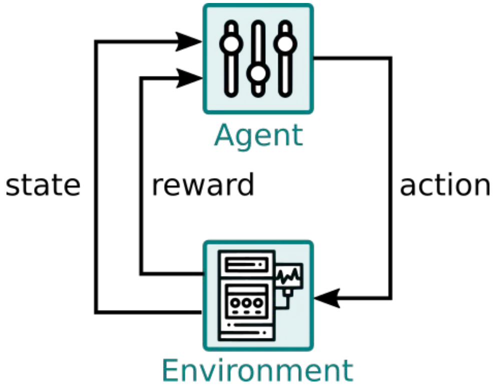
- Data vs. model-based approaches

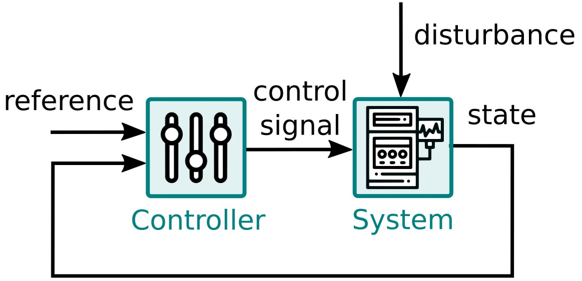

Machine Learning & Control Theory combinations
Learning-enhanced Control
ML for Modeling
ML for Controller Design
ML as a Controller
ML for Generating controllers
ML for Translating Objectives
ML for Control Checking
Control-enhanced Learning
Controlling Training
Controlling Data
Decision making using Control
Guarantees using Control

- Why **combining** them ?
  - Performance
  - Safety
  - Data efficiency
  - Frugality

Sophie Cerf, Eric Rutten. Combining neural networks and control: potentialities, patterns and perspectives. IFAC 2023 - 22nd World Congress of the International Federation of Automatic Control, International Federation of Automatic Control, Jul 2023, Yokohama, Japan. ⟨hal-04060379⟩

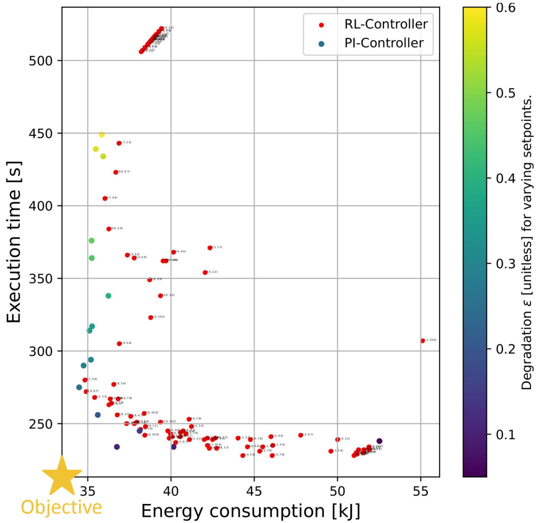
Using Control instead of AI





Using Control instead of AI

Similar performance, with significantly different inference cost



- Control: PI controller
  - 2 parameters

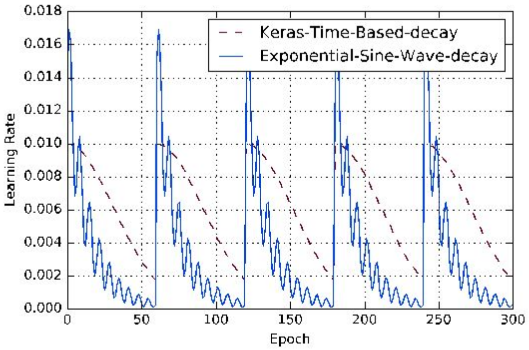
$$\mathbf{u}(t_i) = (K_I \Delta t_i + K_P) \cdot e(t_i) - K_P \cdot e(t_{i-1}) + \mathbf{u}(t_{i-1})$$

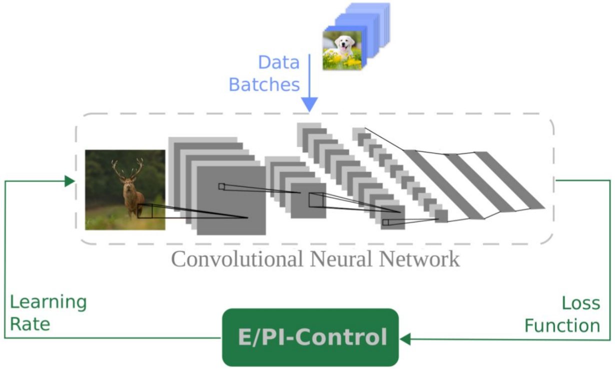
- AI: Reinforcement learning PPO
  - 8100 parameters

Sophie Cerf, Raphaël Bleuse, Valentin Reis, Swann Perarnau, Eric Rutten. Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach. EURO-PAR 2021 - 27th International European Conference on Parallel and Distributed Computing, Aug 2021, Lisbon, Portugal. pp.334-349, ⟨10.1007/978-3-030-85665-6_21⟩. ⟨hal-03259316⟩
A. Raj, S. Perarnau and A. Gokhale, "A Reinforcement Learning Approach for Performance-aware Reduction in Power Consumption of Data Center Compute Nodes," 2023 IEEE International Conference on Cloud Engineering (IC2E), Boston, MA, USA, 2023, pp. 121-130, doi: 10.1109/IC2E59103.2023.00022.

Using Feedback in Training

- Shift from *time-based* adaptation to *feedback*
  - Exploration vs. exploitation trade-off
  - **Training rate** evolution law

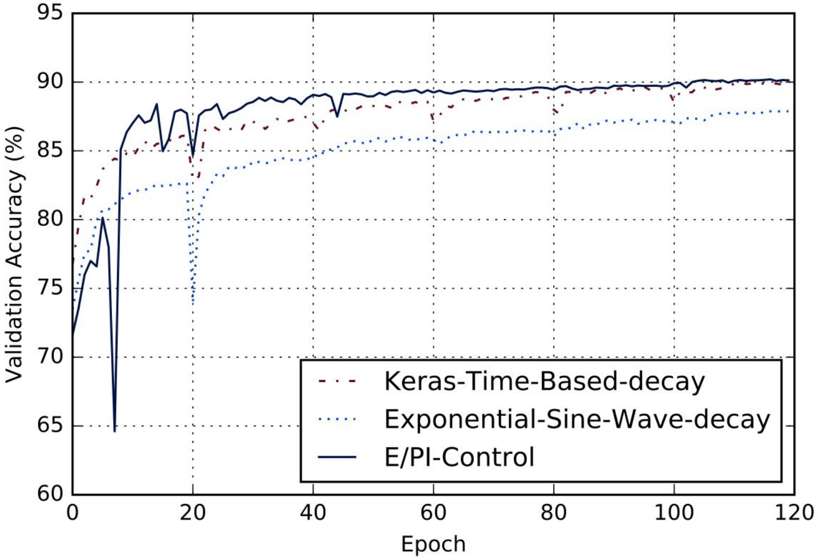


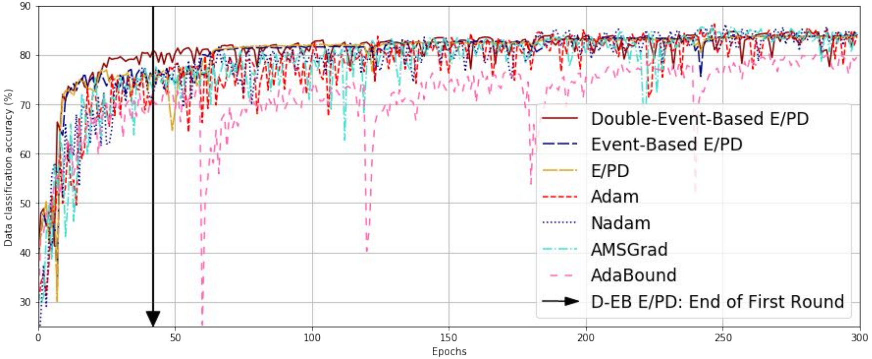


Ghina Dandachi, Sophie Cerf, Yassine Hadjadj-Aoul, Abdelkader Outtagarts, Eric Rutten. A robust control-theory-based exploration strategy in deep reinforcement learning for virtual network embedding. Computer Networks, 2022, 218, pp.1-27. ⟨10.1016/j.comnet.2022.109366⟩. ⟨hal-03792078⟩

Using Feedback in Training

- Faster convergence
  - save up to 67% training time
- Less training
  - switch to the next batch based on the learning speed

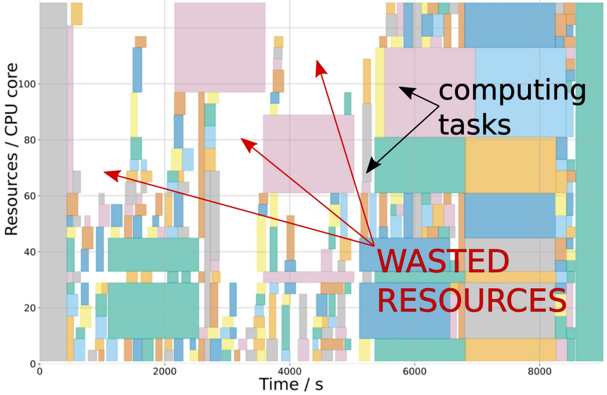




Zilong Zhao, Sophie Cerf, Bogdan Robu, Nicolas Marchand. Event-Based Control for Online Training of Neural Networks. IEEE Control Systems Letters, 2020, 4 (3), pp.773-778. ⟨10.1109/LCSYS.2020.2981984⟩. ⟨hal-02509604⟩

AI as High Performance Computing

- Tasks in shared computing centers
  - scheduling
- Management of **unused** resources
  - inject small tasks to fill the cluster *without impacting the main tasks*
- Open direction
  - federated learning

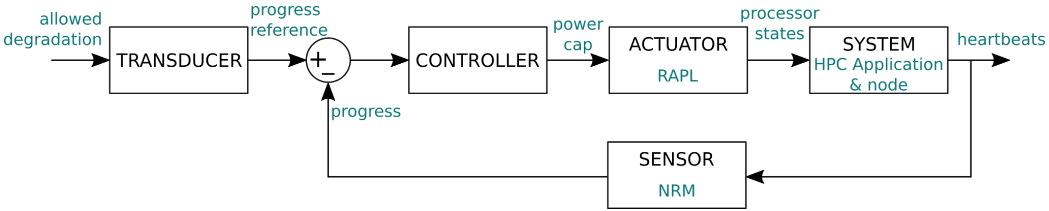


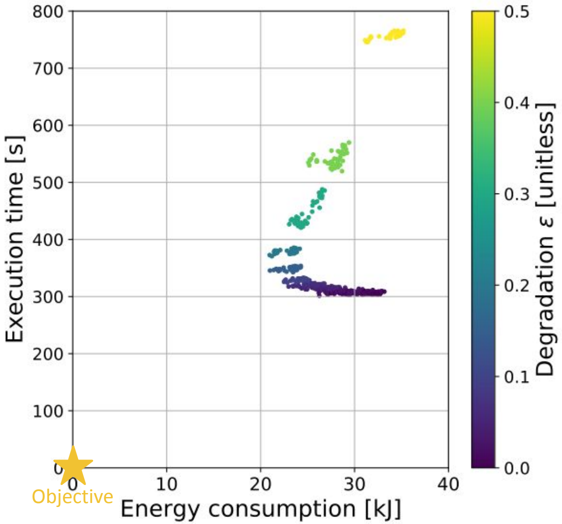
Quentin Guilloteau. Control-based runtime management of HPC systems with support for reproducible experiments. Performance [cs.PF]. Université Grenoble Alpes [2020-..], 2023. English. ⟨NNT : 2023GRALM075⟩. ⟨tel-04389290v2⟩

From Efficiency to Sufficiency

- Reduce CPU **energy** usage in memory-intensive phases

- Acceptable **performance degradation** as a design objective

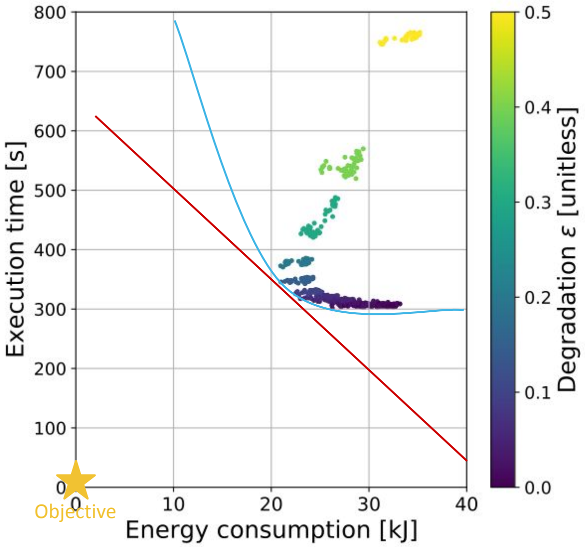




Sophie Cerf, Raphaël Bleuse, Valentin Reis, Swann Perarnau, Eric Rutten. Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach. EURO-PAR 2021 - 27th International European Conference on Parallel and Distributed Computing, Aug 2021, Lisbon, Portugal. pp.334-349, ⟨10.1007/978-3-030-85665-6_21⟩. ⟨hal-03259316⟩

AI shrinkability

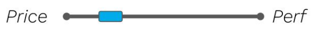
- Can algorithms handle low resource conditions ?
- Which tunable configurations at runtime ?
  - Software: models
  - Hardware: architectures



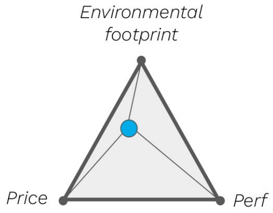
Project ANR RADYAL Resource-Aware DYnamically Adaptable machine Learning

AI within limits

- Environmental footprint as a design objective

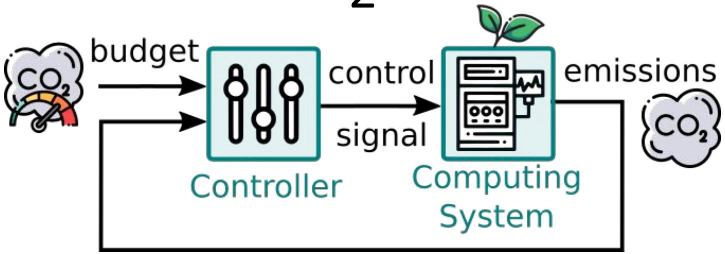


Yesterday



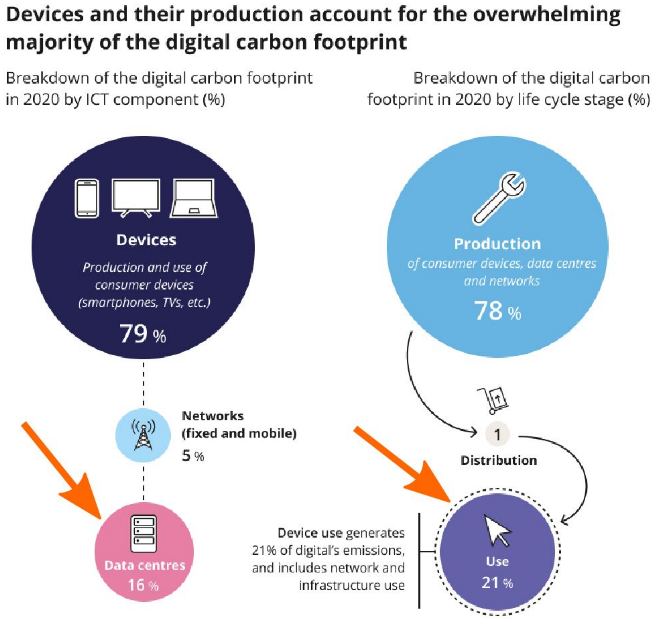
Tomorrow

- Means to ensure a CO_2 budget



Conclusion

- AI and Control: Opportunities to tame usage of resources
- Limitations
  - ○ focus on usage phase, datacenter, climate change
  - ○ Rebound effect



Devices and their production account for the overwhelming majority of the digital carbon footprint

Breakdown of the digital carbon footprint in 2020 by ICT component (%)

Breakdown of the digital carbon footprint in 2020 by life cycle stage (%)

Devices
Production and use of consumer devices (smartphones, TVs, etc.)
79 %
Networks (fixed and mobile)
5 %
Data centres
16 %
Production
of consumer devices, data centres and networks
78 %
1
Distribution
Use
21 %
Device use generates 21% of digital's emissions, and includes network and infrastructure use

ADEME/Arcep

AI and Control
Opportunities to tame usage of resources

Thank You

Sophie Cerf

sophie.cerf@inria.fr

Université de Lille

Inria

CRIStAL
Centre de Recherche en Informatique,
Signal et Automatique de Lille