A Deep Reinforcement Learning Framework for Locking Optimization in Simulated Optical Cavities

16th March 2026

Andrea Svizzeretto[1,2]
andrea.svizzeretto@dottorandi.unipg.it

Mateusz Bawaj[1,2]
mateusz.bawaj@unipg.it

[1]Dipartimento di Fisica e Geologia, Università di Perugia, I-06123 Perugia, Italy; [2]INFN, Sezione di Perugia, I-06123 Perugia, Italy

PAN
POLSKA AKADEMIA NAUK

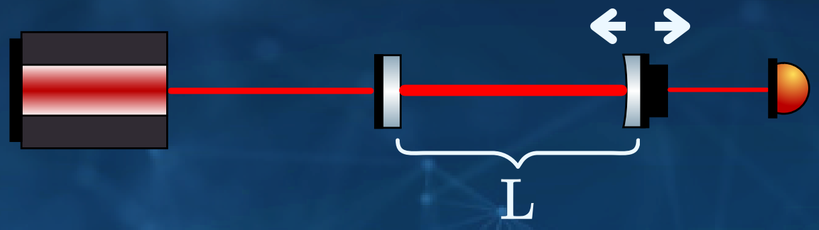
A.D. 1308
unipg
UNIVERSITÀ DEGLI STUDI
DI PERUGIA

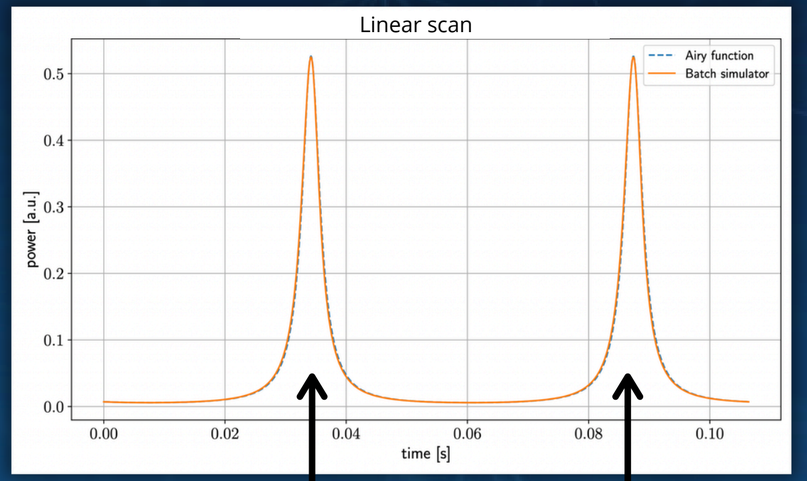
INFN
Sez. di Perugia

VIRGO

EINSTEIN
TELESCOPE

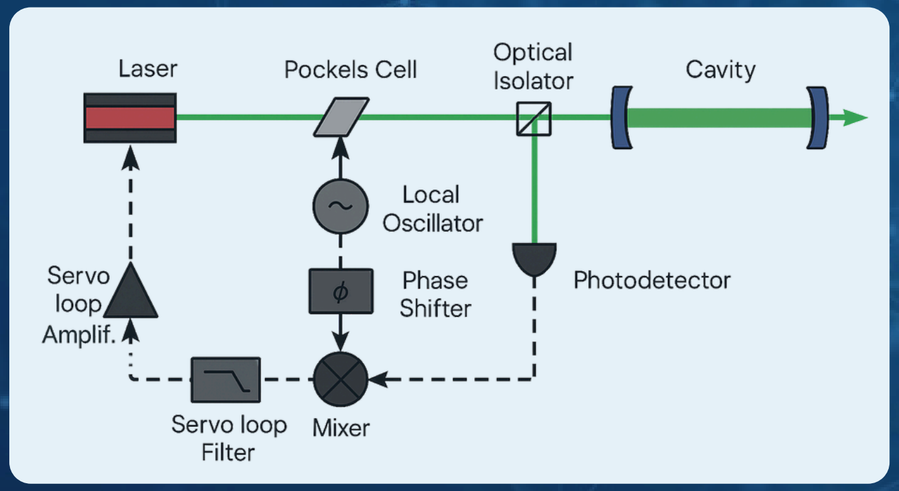
Lock Acquisition





- By shifting the mirror position or changing the laser frequency ⟶ Linear scan
- One peak for each resonance in the optical power.
- Locking means putting the cavity in resonance.

Resonances $L = m\frac{\lambda}{2}$

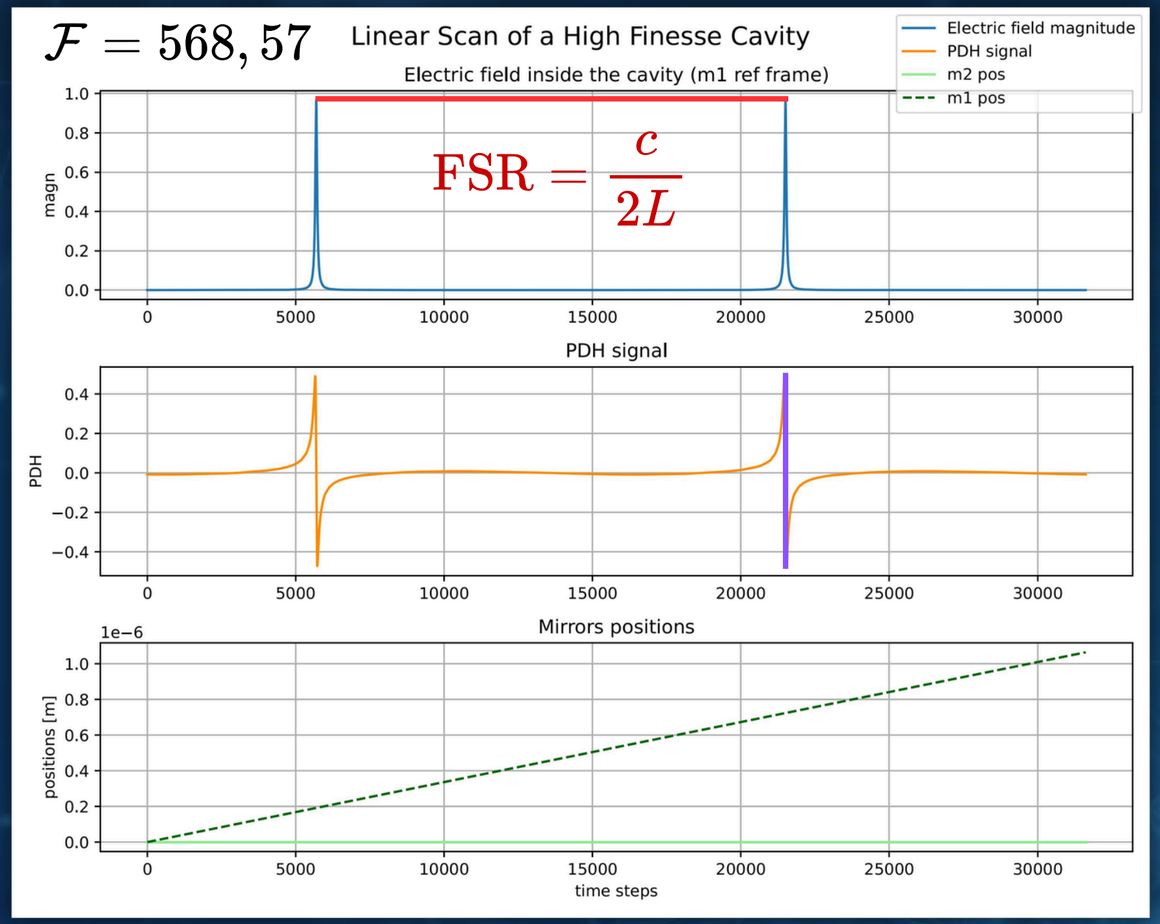


Pound-Drever-Hall (PDH) Technique [1]

- Method for obtaining an error signal useful for acquiring lock condition.
- Zero crossing of the PDH signal for each resonance, so that it's possible to understand, based on its sign, in which side of the resonance the cavity is.
- It works only in the linear regime of the signal ⟶ Very narrow

[1] *"An Introduction to Pound-Drever-Hall laser frequency stabilisation" Eric D. Black, 2001*

High Finesse Cavities



- Finesse is an important characteristic of optical cavities.

$$\mathcal{F} = \frac{\pi}{2}\sqrt{F} \qquad F = \frac{4r_a r_b}{(1 - r_a r_b)^2}$$

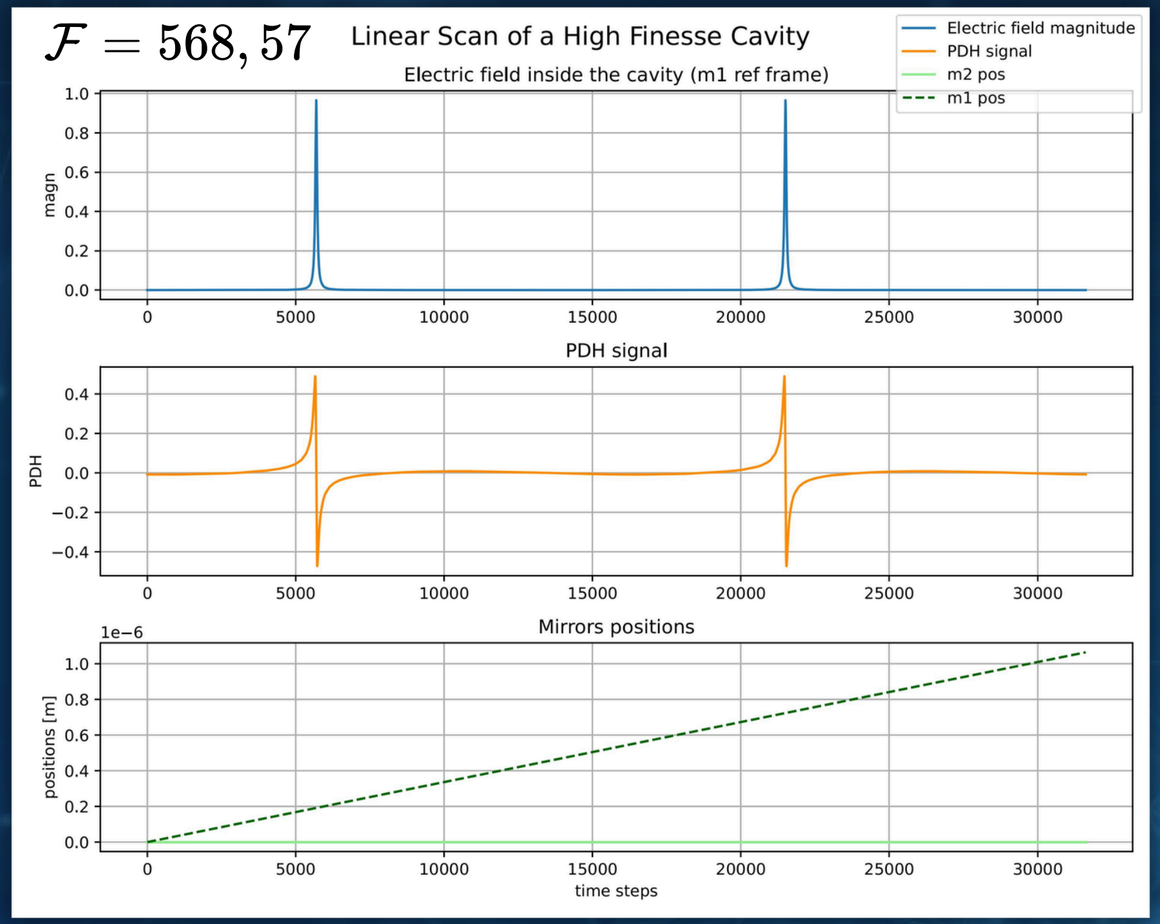
- For high finesse cavities linear region shrinks a lot. PDH technique can be used only here.

$$\text{linewidth} = \frac{\text{FSR}}{\mathcal{F}}$$

- It corresponds to a small percentage of one Free Spectral Range

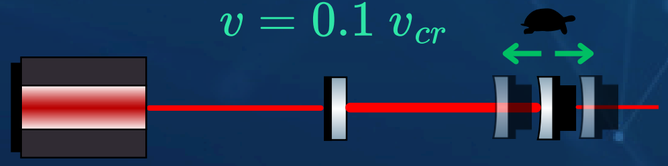
- High Finesse cavities are the ones used within GW detectors to enhance the sensitivity.
- Lock acquisition efficiency becomes crucial.

Non Linear Dynamics



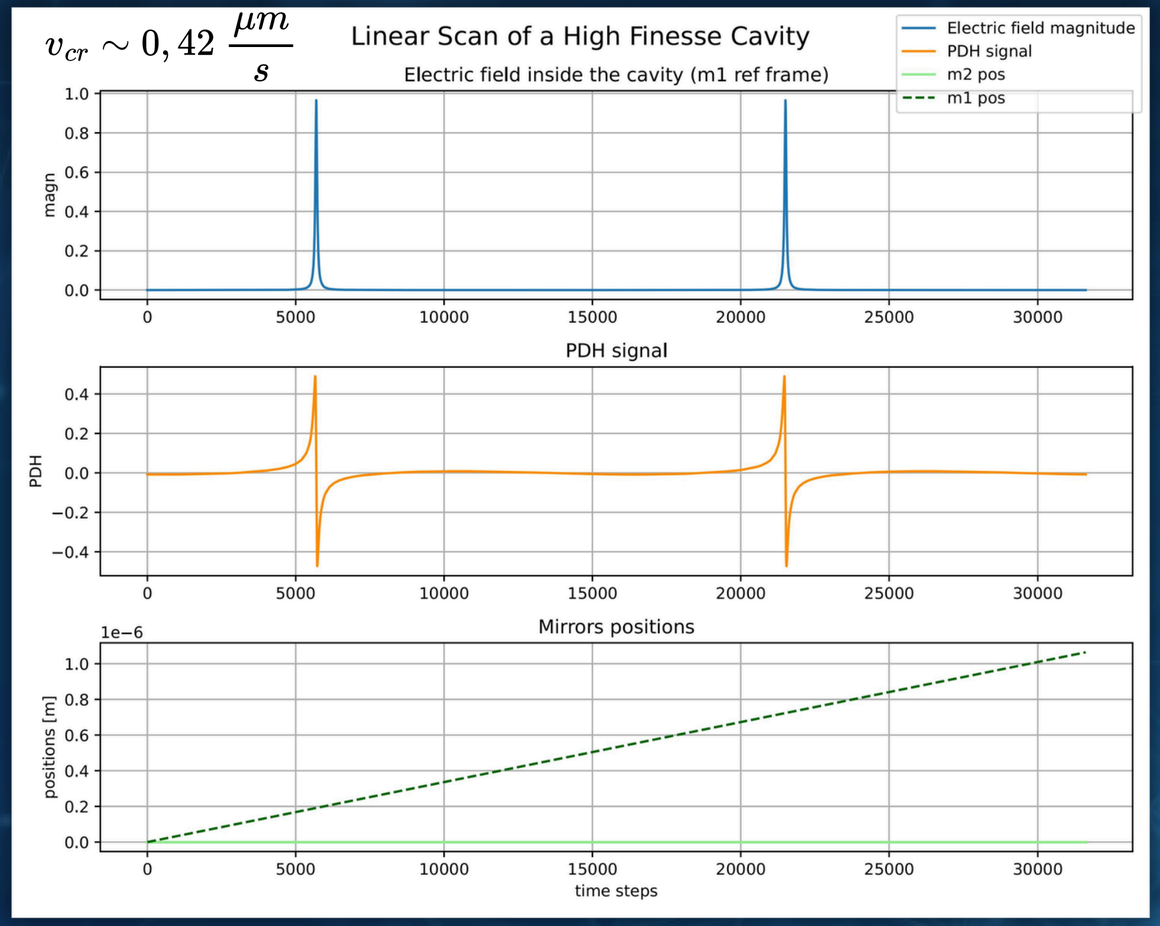
$$v \geq v_{\text{cr}} \approx \frac{\lambda \pi c}{4 L \mathcal{F}^2}$$

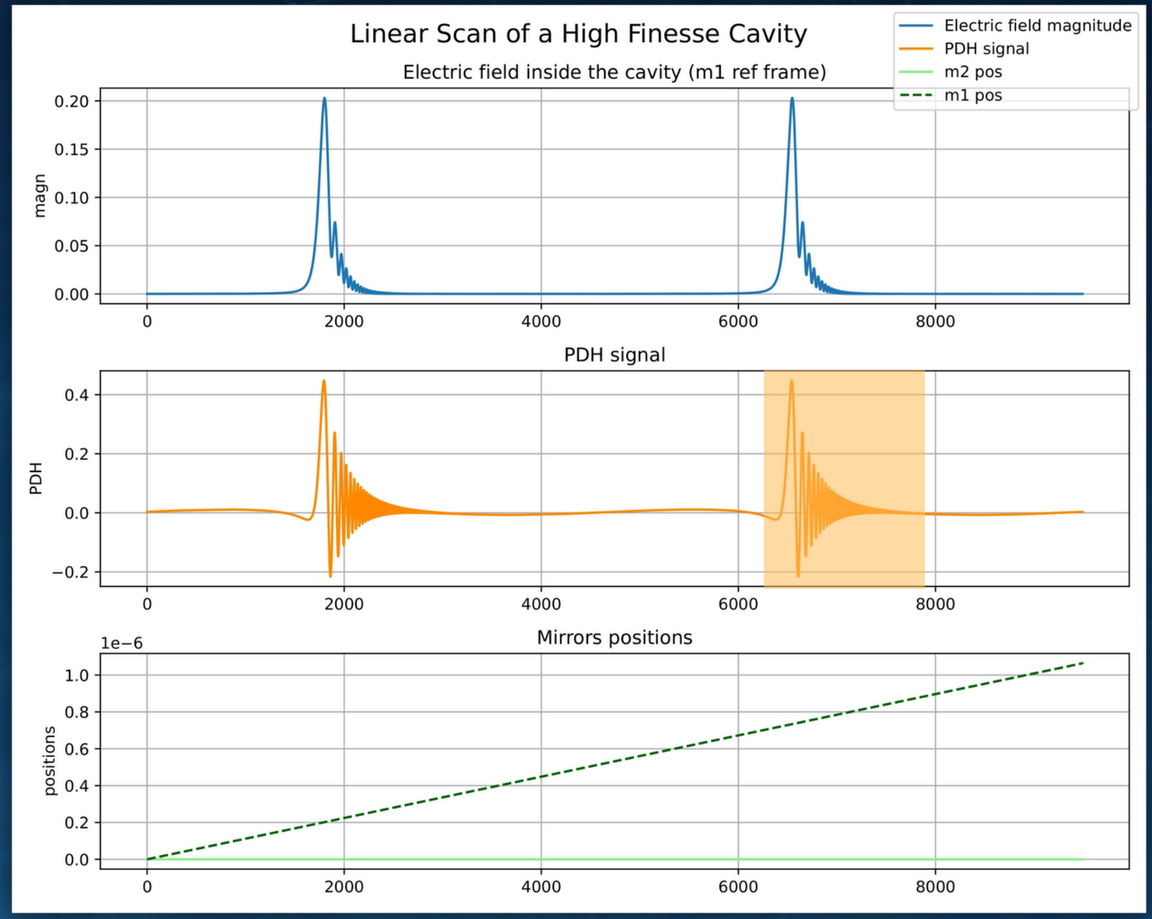
$v = 0.1\, v_{cr}$



Non Linear Dynamics

$v_{cr} \sim 0,42 \ \frac{\mu m}{s}$

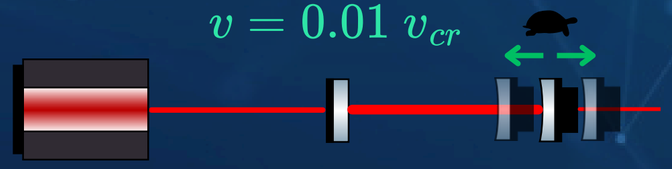




$$v \geq v_{cr} \approx \frac{\lambda \pi c}{4 L \mathcal{F}^2}$$

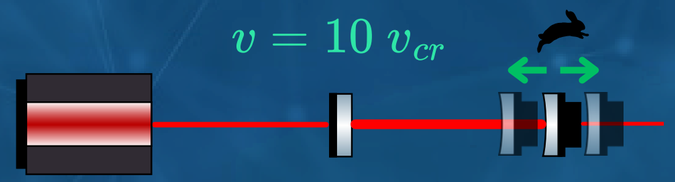
Ring-down Effect

$v = 0.01 \ v_{cr}$



We can use RL to extract info from Non Linear Region, which corresponds to the ~30% of one FSR.

$v = 10 \ v_{cr}$



Why Reinforcement Learning?

Highly suitable in several control tasks and for highly dimensional parameter spaces.

Classic feedback controllers	Deep Reinforcement Learning Models
• Assume linear, time-invariant system responses. • Often require manual tuning and do not adapt to changing dynamics.	• Learn a control policy by interacting with the environment. • Optimizing performance directly based on a reward function, even in the absence of a precise model.

Why Reinforcement Learning?

Highly suitable in several control tasks and for highly dimensional parameter spaces.

Classic feedback controllers (e.g. PID)	Deep Reinforcement Learning Models
• Assume linear, time-invariant system responses. • Often require manual tuning and do not adapt to changing dynamics.	• Learn a control policy by interacting with the environment. • Optimizing performance directly based on a reward function, even in the absence of a precise model.

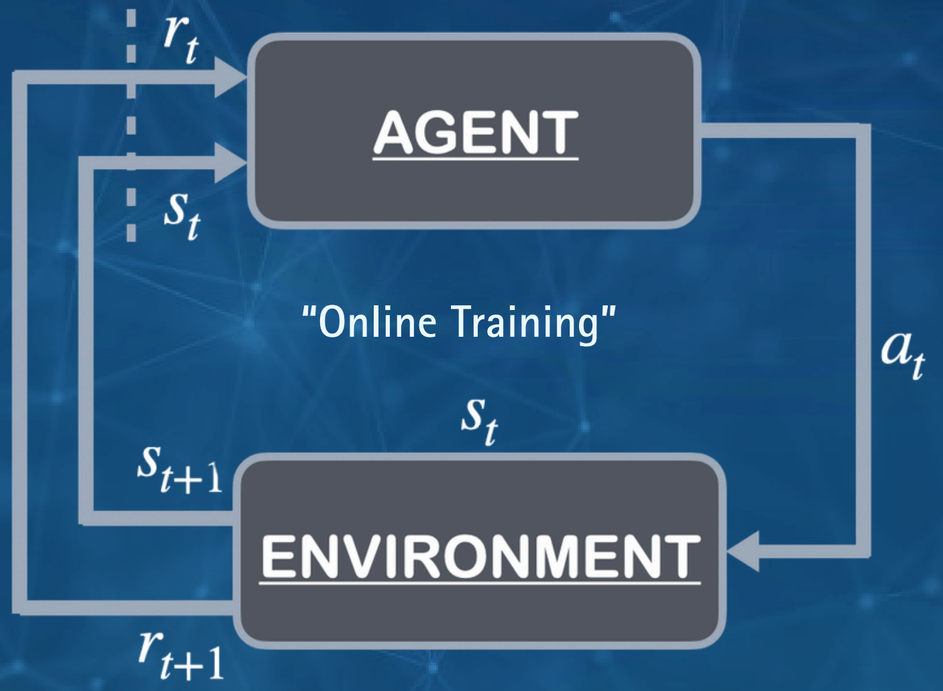
Markov Decision Processes (MDP)

s_t = Current state of the environment

a_t = Action chosen by the agent

r_t = Reward earned by the agent based on the goodness of state

- Recent works ([2], [3], [4], [5]) with several Machine Learning and Reinforcement Learning applications for controlling and alignigning tasks



[2] *"A Deep Learning Technique to Control the Non-linear Dynamics of a Gravitational-wave Interferometer" P. Ma, G. Vajente 2023*

[3] *"First demonstration of neural sensing and control in a kilometer-scale gravitational wave observatory" N. Mukund et al. 2023*

[4] *"Interferobot: aligning an optical interferometer by a reinforcement learning agent" D. Sorokin et al. 2021*

[5] *"Automated alignment of an optical cavity using machine learning" J. Qin et al. 2025*

[6] *"Improving cosmological* reach of a gravitational wave observatory using Deep Loop Shaping*" J. Buchli et al. 2025*

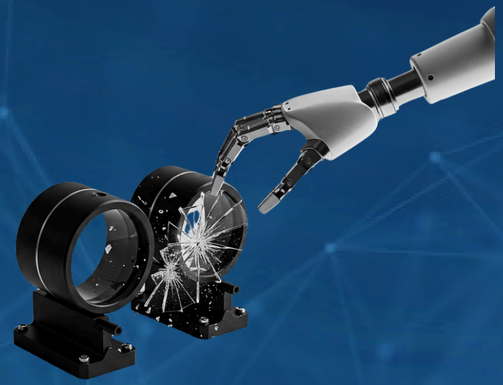
Training on Real World?

Let's train the model directly on the real set up! Not so fast....

RL needs a lot of tries to learn the optimal policy through trial and error

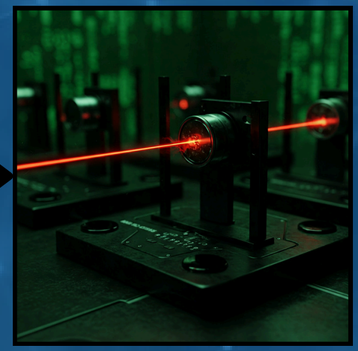
Exploration stage

High risk of hardware damage when training on real set up



Time and money commitment

Better to train first on simulated environment



Fast time-domain Simulator

We developed a fast time domain simulator based on the equation of fields from [6].
This equation describes the dynamics of the electric field within the cavity.

Inputs

Mirror Positions $x^{\mathrm{a}}(t)$, $x^{\mathrm{b}}(t)$

Calculation Frequency
(Just a the beginning)

Input Electric Field $E_{\mathrm{in}}(t)$

$$E(t) = t_{\mathrm{a}} \sum_{n=0}^{N-1} (r_{\mathrm{a}} r_{\mathrm{b}})^n e^{-2ik S_{\mathrm{n}}(t)} e^{-2ik x_{\mathrm{a}}(t)} E_{\mathrm{in}}(t - 2nT) + \\ + (r_{\mathrm{a}} r_{\mathrm{b}})^N e^{-2ik S_N(t)} E(t - 2NT)$$

$$S_{\mathrm{n}}(t) = \sum_{p=0}^{n-1} d(t - 2pT) \qquad d(t) = x_{\mathrm{b}}(t) - x_{\mathrm{a}}(t)$$

N.B: The cavity is stationary and reactive, once you send an input the length will change immediately and it will remain in that state until you will send another shift.

$\sim 3\ \mu s$

Outputs

$E(t)$

$\varepsilon_{\mathrm{PDH}}(t)$

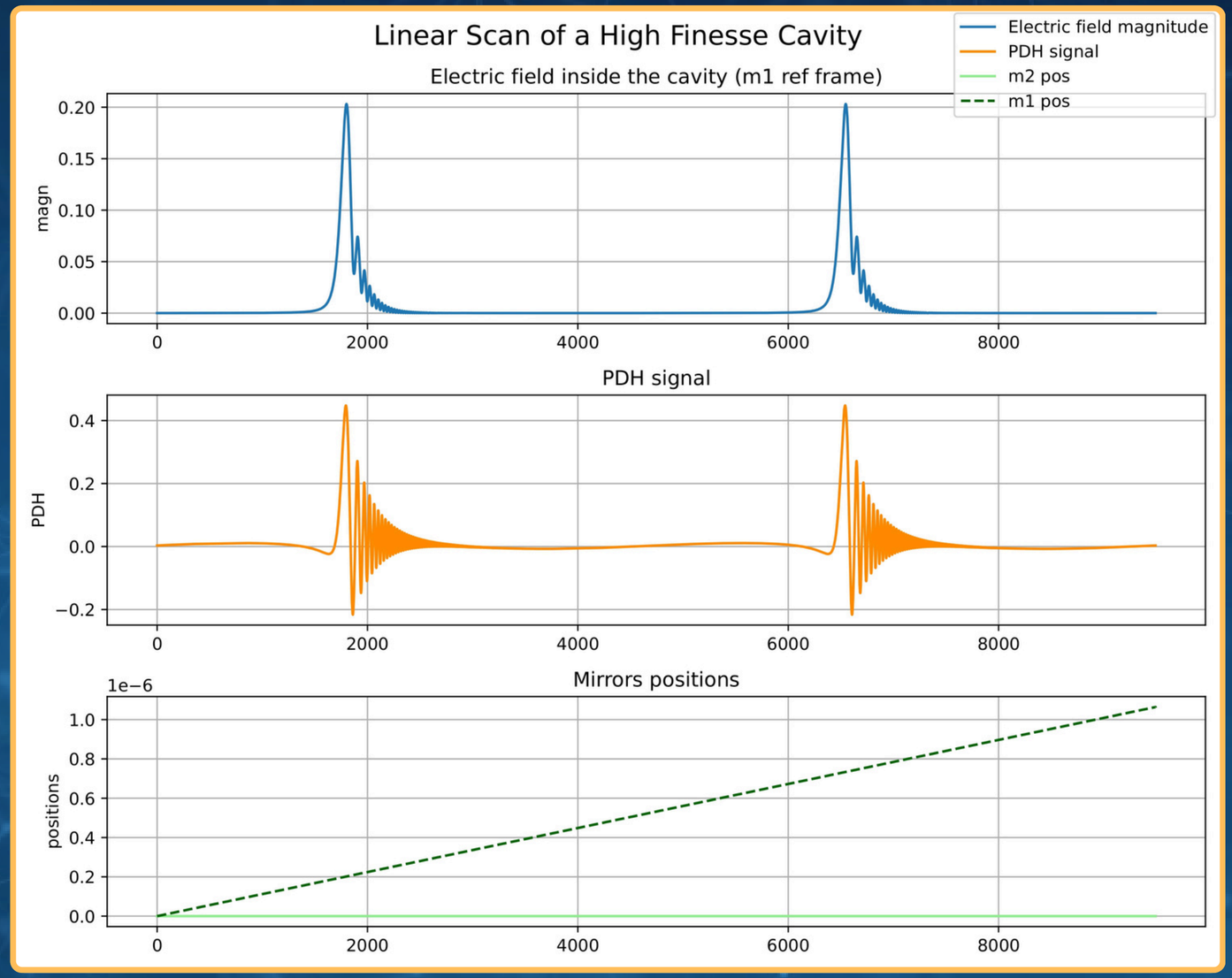
$E_{\mathrm{ref}}(t)$

[8] *"Dynamics of Laser Interferometric Gravitational Wave Detectors" M. Rakhmanov, Phd Thesis, 2000*

Fast time-domain Simulator

We are able of reproducing ring down for various mirror velocity in wide range of cavities.

We are simulating only the longitudinal degree of freedom.



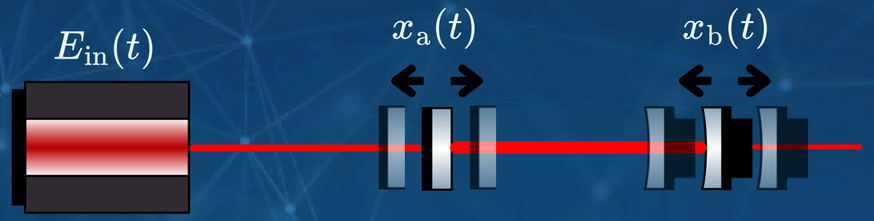
[9] *"Continuous control with deep reinforcement learning" T. P. Lillicrap et al. 2019*
[10] *"Gymnasium: A Standard Interface for Reinforcement Learning Environments" M. Towers et al. 2024*

Agent

DDPG
Deep Deterministic Policy Gradient [9]

Simulated Environment

Implemented with Gymnasium [10]



Reinforcement Learning Framework

[7] "Continuous control with deep reinforcement learning" T. P. Lillicrap et al. 2019
[8] "Gymnasium: A Standard Interface for Reinforcement Learning Environments" M. Towers et al. 2024

Agent

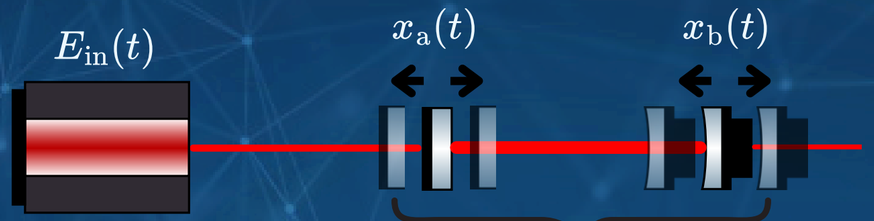
Action

$$a_t = \{x_a(t), x_b(t)\}$$

DDPG
Deep Deterministic Policy Gradient [9]

Simulated Environment

Implemented with Gymnasium [10]



$$\mathrm{L} = \mathrm{L} + d(t) = \mathrm{L} + (x_b(t) - x_a(t))$$

Reinforcement Learning Framework

[7] "Continuous control with deep reinforcement learning" T. P. Lillicrap et al. 2019
[8] "Gymnasium: A Standard Interface for Reinforcement Learning Environments" M. Towers et al. 2024

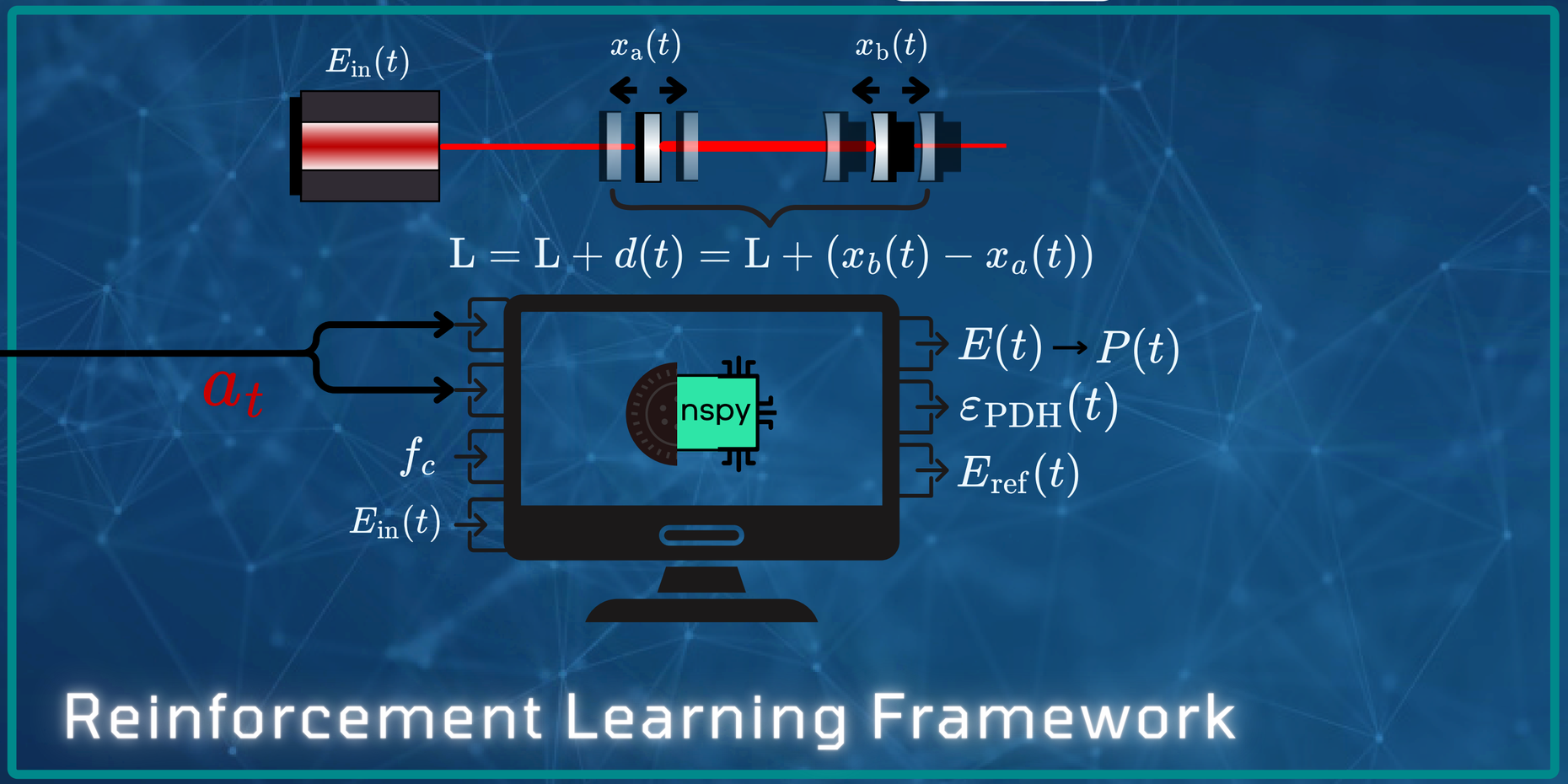
Agent

Action

$a_t = \{x_a(t), x_b(t)\}$

DDPG
Deep Deterministic Policy Gradient [9]

Simulated Environment Implemented with Gymnasium [10]



[7] "Continuous control with deep reinforcement learning" T. P. Lillicrap et al. 2019
[8] "Gymnasium: A Standard Interface for Reinforcement Learning Environments" M. Towers et al. 2024

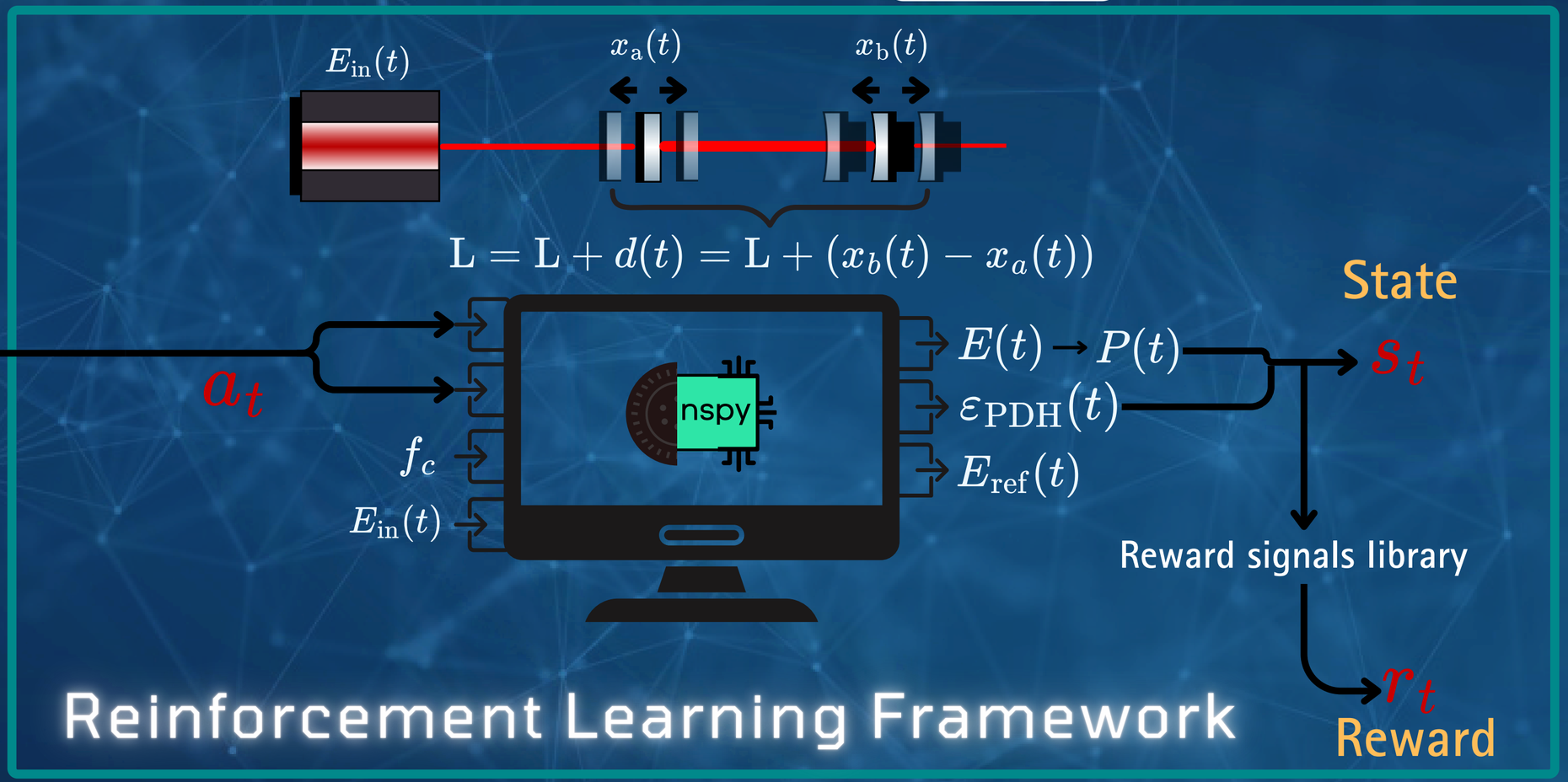
Agent

Action

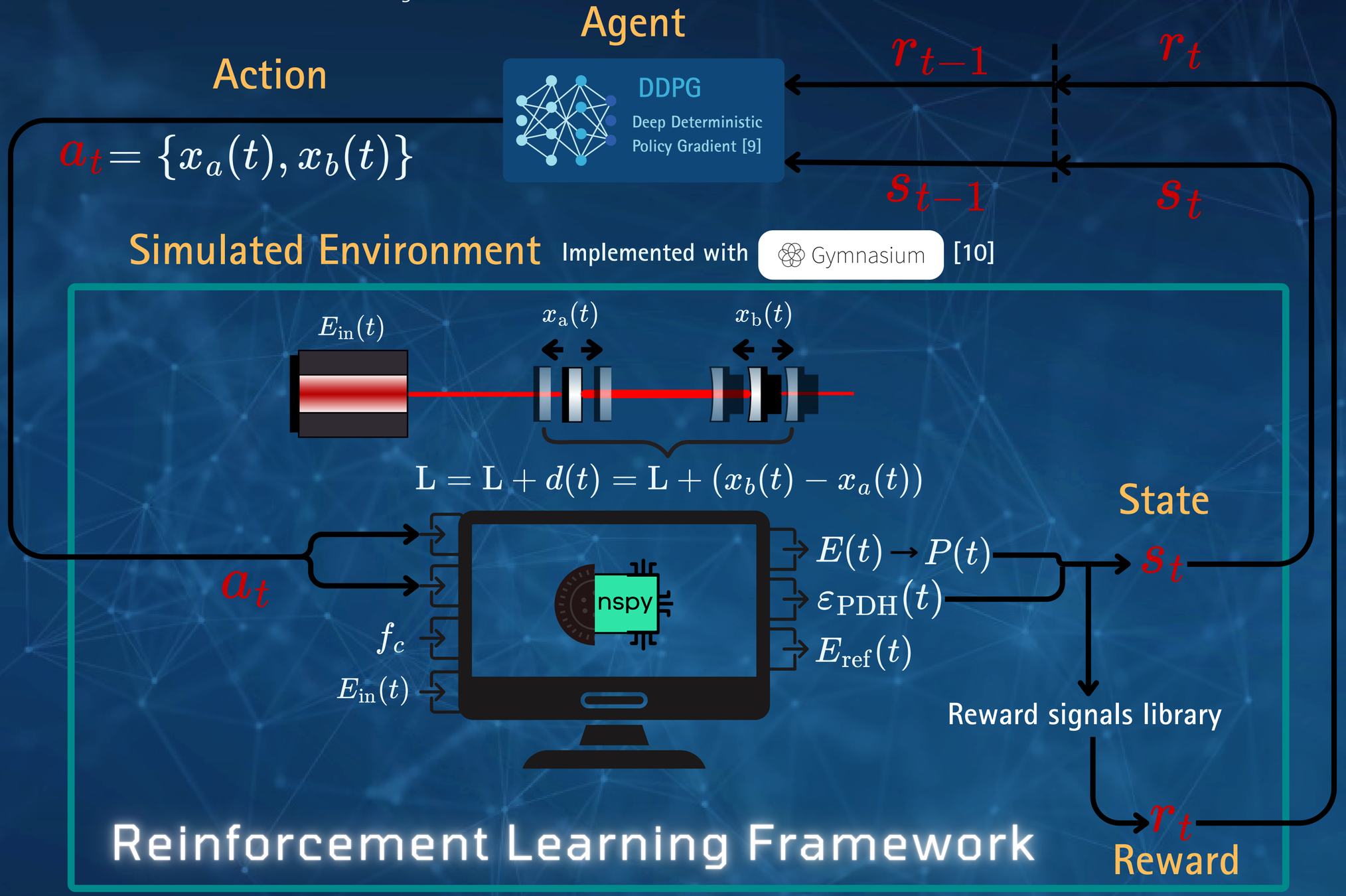
$a_t = \{x_a(t), x_b(t)\}$

DDPG
Deep Deterministic Policy Gradient [9]

Simulated Environment Implemented with Gymnasium [10]



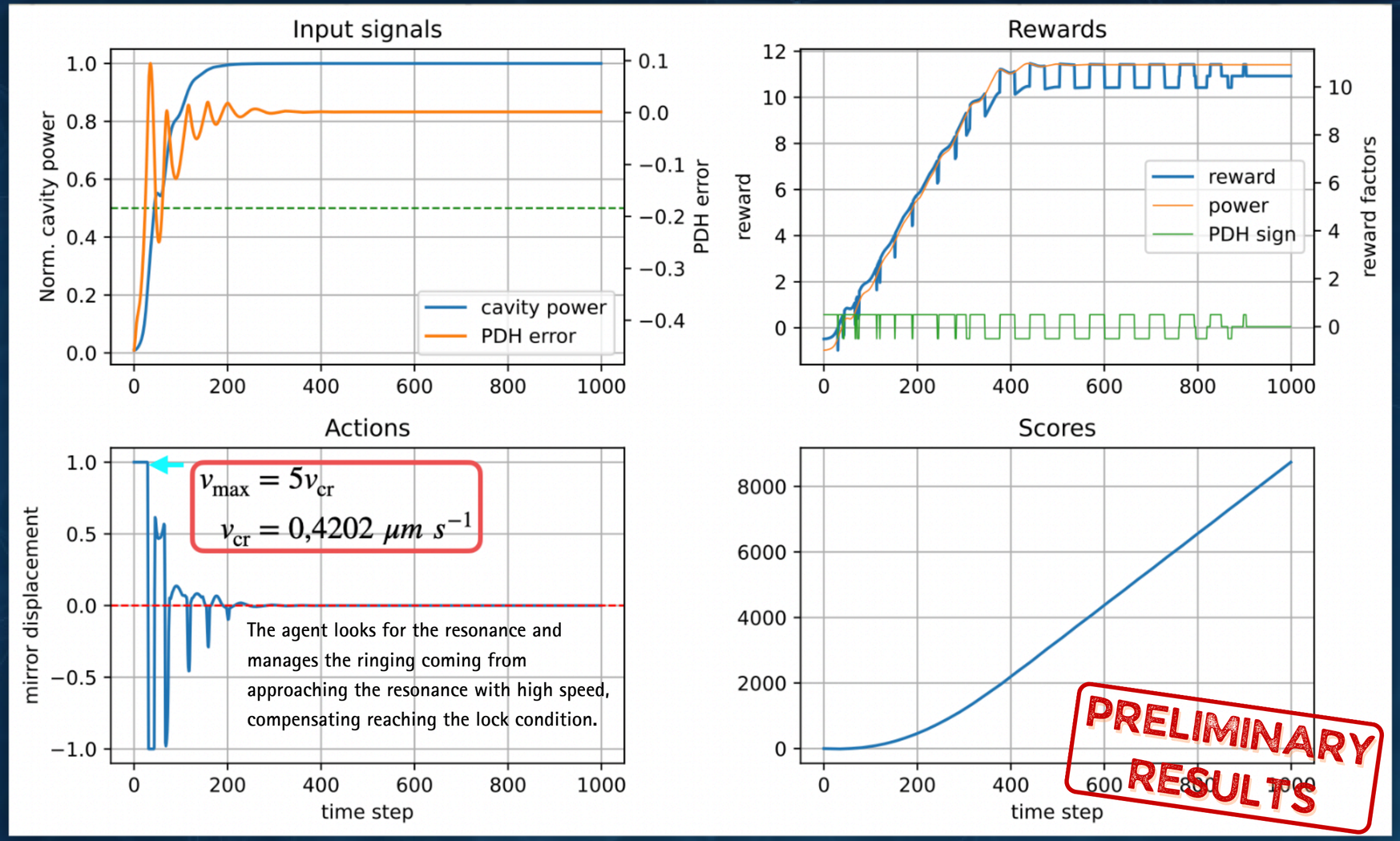
[7] "Continuous control with deep reinforcement learning" T. P. Lillicrap et al. 2019
[8] "Gymnasium: A Standard Interface for Reinforcement Learning Environments" M. Towers et al. 2024



Training and Test

Test of a DDPG model trained for 120000 time-steps.

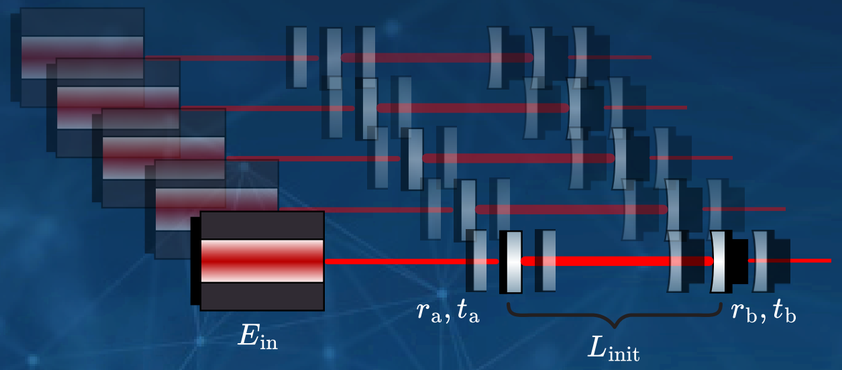
- Only for the output mirror.
- ~ 200 time-steps to properly lock the cavity.
- The agent approaches the resonance with maximum speed allowed by the environment



HPC Training Sessions

Different models

- DDPG
- TD3
- RecurrentPPO

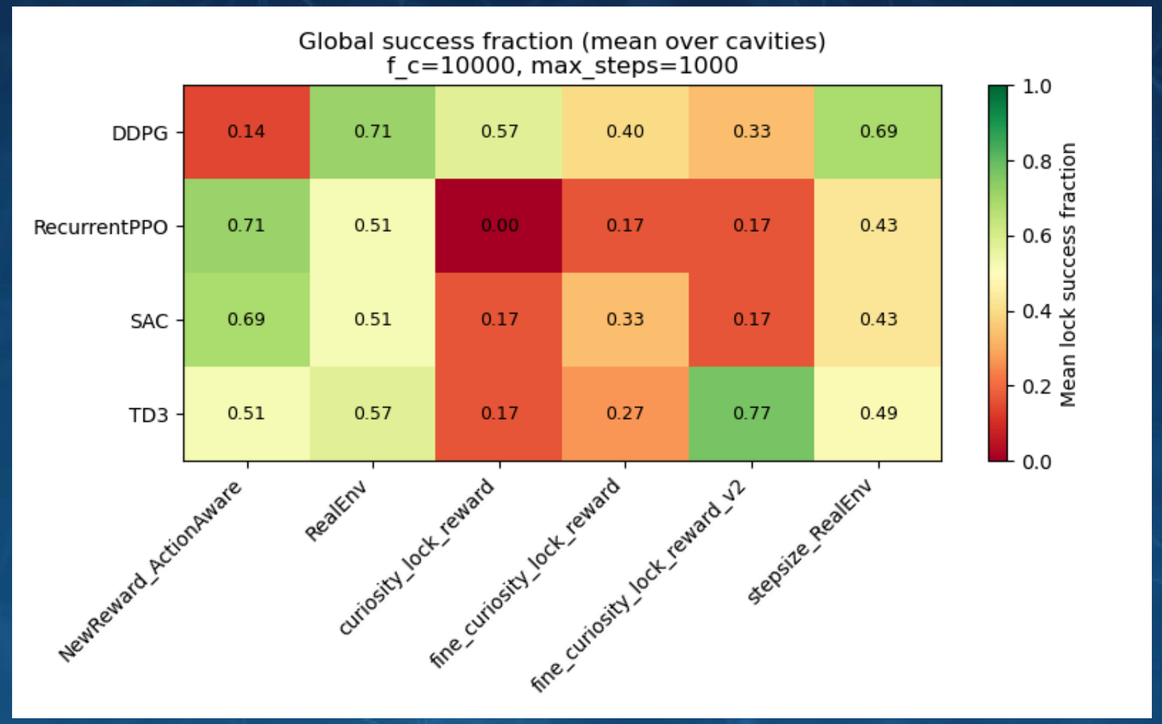


Different cavities

- ARM
- Microcavities
- ...

Domain Randomization

$\langle L_{init}, E_{in}, \text{signal noise}\ldots \rangle$



HPC Infrastructures

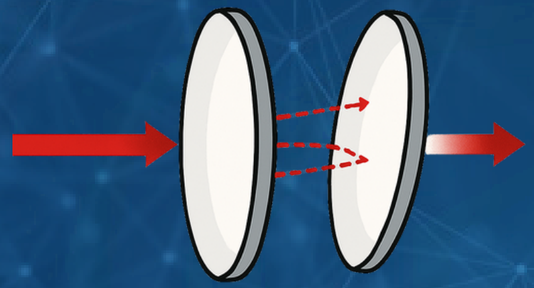
KEY BENEFIT

Run heavy training sessions to find the best model–reward configurations for future SimToReal transfer.

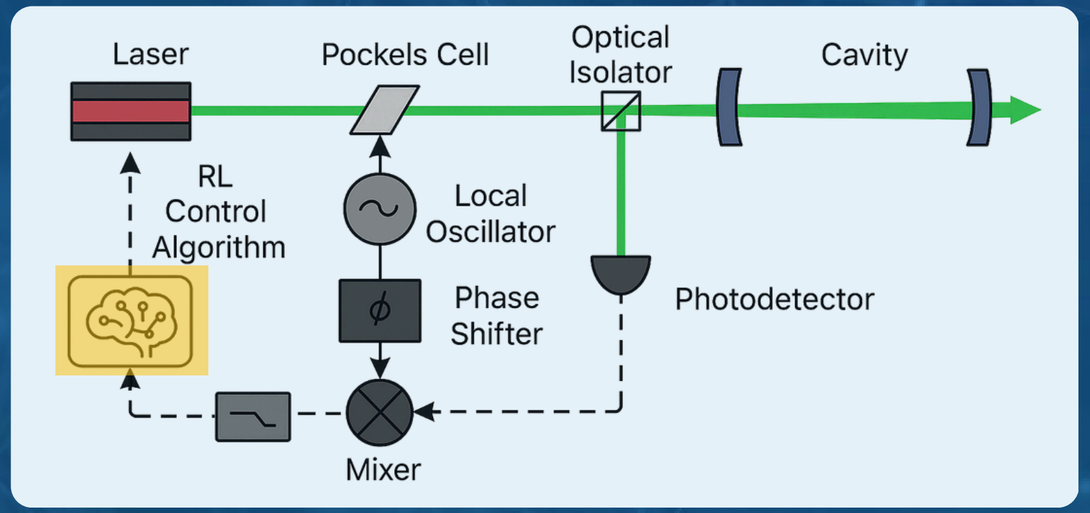
Future Steps

- Face the Sim2Real transfer problem [11], trying to decrease the reality gap.

1. Finite force
2. Angular displacements
3. Delayed interaction



- Integrate the AI framework inside a real control loop and actually control an optical cavity.



[11] *"Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey" Wenshuai Zhao et al. 2021*

Thanks for your attention!

References

[1] *“An Introduction to Pound-Drever-Hall laser frequency stabilisation” Eric D. Black, 2001*

[2] *"A Deep Learning Technique to Control the Non-linear Dynamics of a Gravitational-wave Interferometer" P. Ma, G. Vajente 2023*

[3] *"First demonstration of neural sensing and control in a kilometer-scale gravitational wave observatory" N. Mukund et al. 2023*

[4] *"Interferobot: aligning an optical interferometer by a reinforcement learning agent" D. Sorokin et al. 2021*

[5] *"Automated alignment of an optical cavity using machine learning" J. Qin et al. 2025*

[6] "Improving cosmological reach of a gravitational wave observatory using Deep Loop Shaping" J. Buchli et al. 2025

[7] *"Dynamics of Laser Interferometric Gravitational Wave Detectors" M. Rakhmanov, Phd Thesis, 2000*

[8] *"Continuous control with deep reinforcement learning" T. P. Lillicrap et al. 2019*

[9] *“Gymnasium: A Standard Interface for Reinforcement Learning Environments” M. Towers et al. 2024*

[10] *"Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey" Wenshuai Zhao et al. 2021*

“Advanced Virgo Plus: Future Perspectives” Acernese et al. 2023

"New algorithm for the Guided Lock technique for a high-Finesse optical cavity" D. Bersanetti et al. 2019

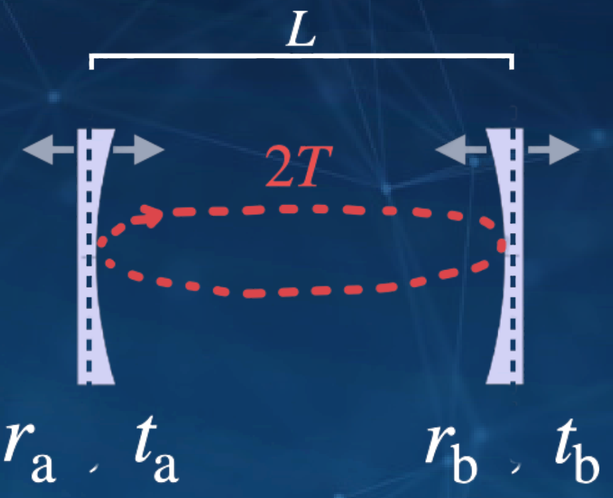
BACKUP

Backup

After one round-trip the electric field is described [Rakhmanov 2000]:

$$E(t) = t_a E_{\text{in}}(t) + r_a r_b e^{-2\imath k\, d(t)}\, E(t - 2T)$$

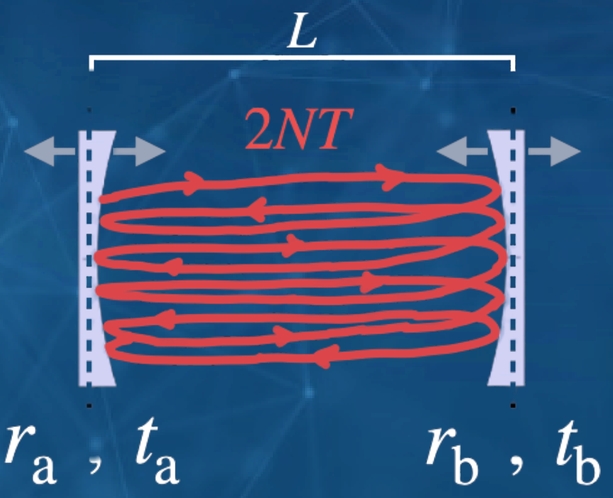
$$d(t) = L + \xi(t) = L + x_b(t - T) - x_a(t)$$



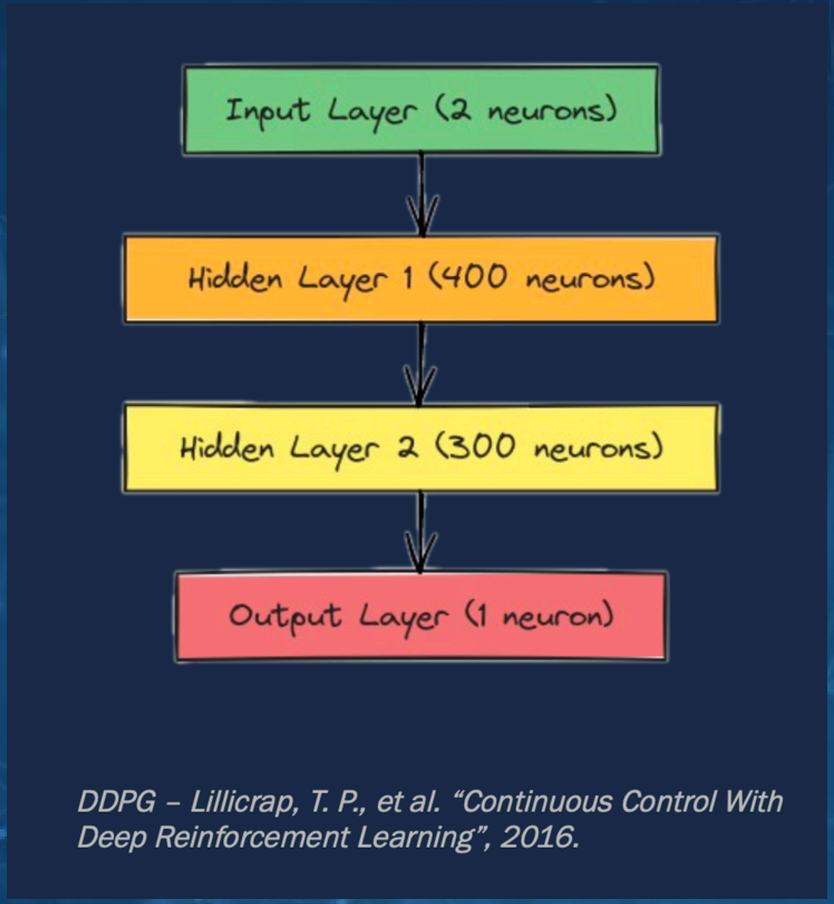
After N round-trip:

$$E(t) = t_a \sum_{n=0}^{N-1} (r_a r_b)^n e^{-2ik S_n(t)}\, E_{\text{in}}(t - 2nT) + (r_a r_b)^N e^{-2ikS_N(t)} E(t - 2NT)$$

$$S_n(t) = \sum_{p=0}^{n-1} d(t - 2pT)$$



Backup



DDPG – Lillicrap, T. P., et al. "Continuous Control With Deep Reinforcement Learning", 2016.

Backup

DSP for Super Attenuator control 3.125 µs delay,

Real Time DAQ 100 µs.

DDPG actor inference in 1.20 ± 0.87 ms.

DAQ a maximum rate of 200 Hz.

Backup

NewReward_ActionAware

$$[r_{\text{power}} = P - \log(1-P) - 1][r_{\text{action}} = 0.8\,\text{sign}(-E\,a)][r = r_{\text{power}} + \frac{1}{2} r_{\text{action}}]$$

RealEnv

$$[r = -|P-1| - \log(|1-P|)]$$

curiosity_lock_reward

$$[g(P) = \frac{1}{1 + e^{-k(P - p_0)}}]$$

$$[r_{\text{power}} = \tanh(3P)][r_{\text{pdh}} = -0.5\,E^2][r_{\text{act}} = -(0.01 + 0.20\,g(P))\,a^2][r_{\text{lock}} = 1 - e^{-0.3\,t_{\text{lock}}}]$$

$$[\Delta P = P_t - P_{t-1}][\Delta E = E_t - E_{t-1}][\text{novelty} = \sqrt{\Delta P^2 + w_E\,\Delta E^2}][r_{\text{int}} = \beta \tanh(\alpha\,\text{novelty})(1 - g(P))]$$

$$[r = r_{\text{power}} + r_{\text{pdh}} + r_{\text{act}} + r_{\text{lock}} + r_{\text{int}}]$$

stepsize_RealEnv

$$[r_{\text{power}} = -|P-1| - \log(|1-P|)] \qquad [\text{err} = (1-P) + E^2][r_{\text{step}} = -0.05 \frac{|a|}{\text{err} + 0.05}]$$

$$[r = r_{\text{power}} + r_{\text{step}}]$$

Backup

fine_curiosity_lock_reward_v2

$$[g_c(P,E) = \exp\left(-\frac{(1-P)^2}{2\sigma_P^2}\right)\exp\left(-\frac{E^2}{2\sigma_E^2}\right)]$$

$$[r_{\text{power}} = \tanh(3P)] \qquad [r_{\text{peak}} = 0.9\exp\left(-\frac{(1-P)^2}{2\sigma_P^2}\right)]$$

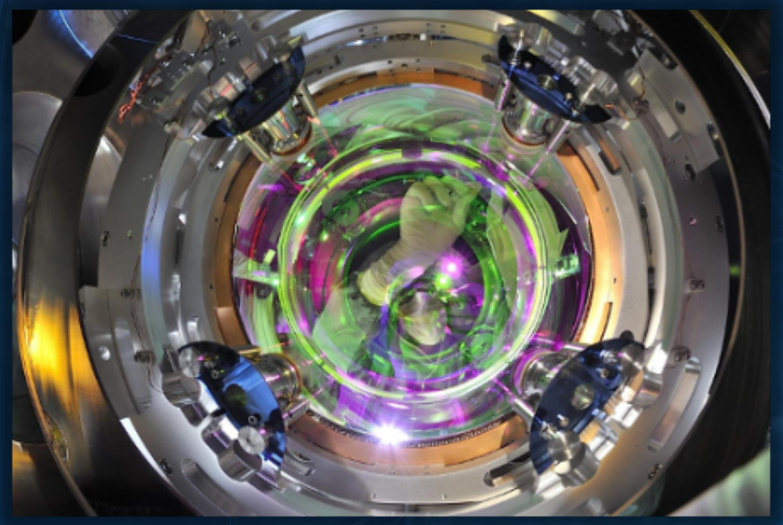
$$[r_{\text{pdh}} = -(0.2 + 5g(P))E^2] \quad [r_{\text{act}} = -(0.01 + 0.25g_c)a^2]$$

$$[r_{\text{align}} = 0.15\ \tanh(3(-Ea))\,g(P)] \quad [r_{\text{int}} = \beta\tanh(\alpha\,\text{novelty})(1-g(P))]$$

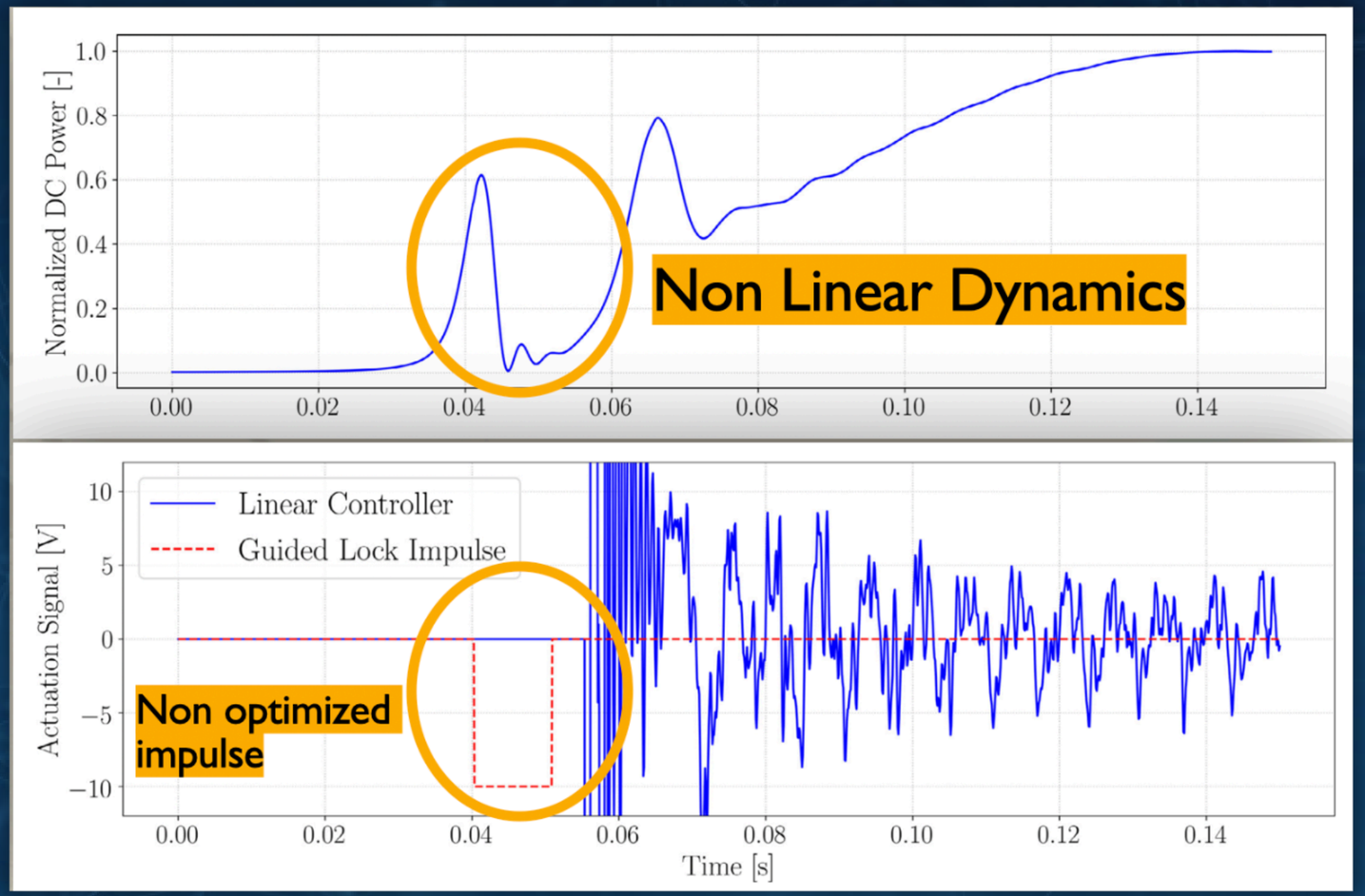
$$[r_{\text{lock}} = 1 - e^{-0.3\,t_{\text{lock}}}]$$

$$[r = r_{\text{power}} + r_{\text{peak}} + r_{\text{pdh}} + r_{\text{act}} + r_{\text{align}} + r_{\text{lock}} + r_{\text{int}}]$$

Modified Guided Lock

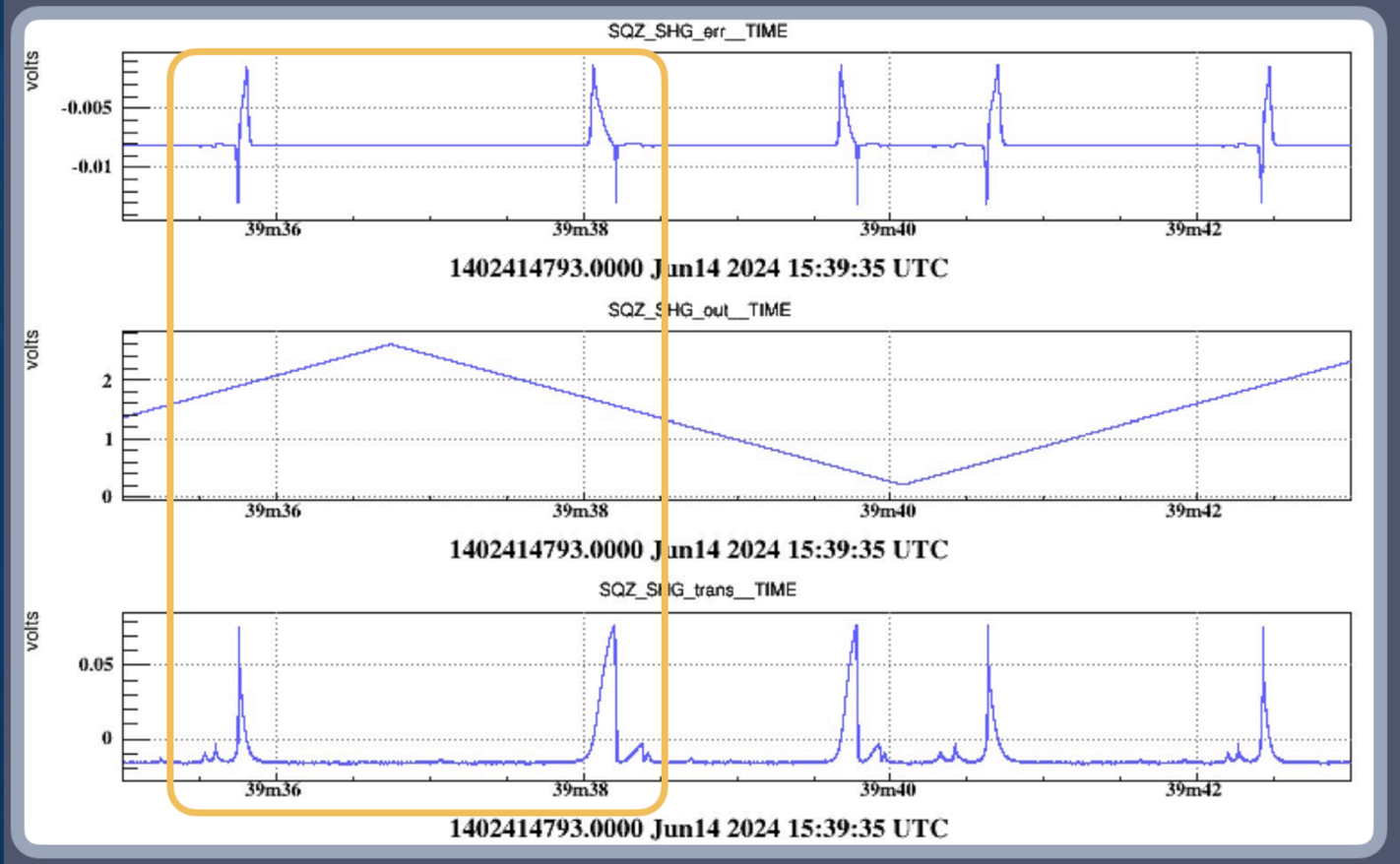


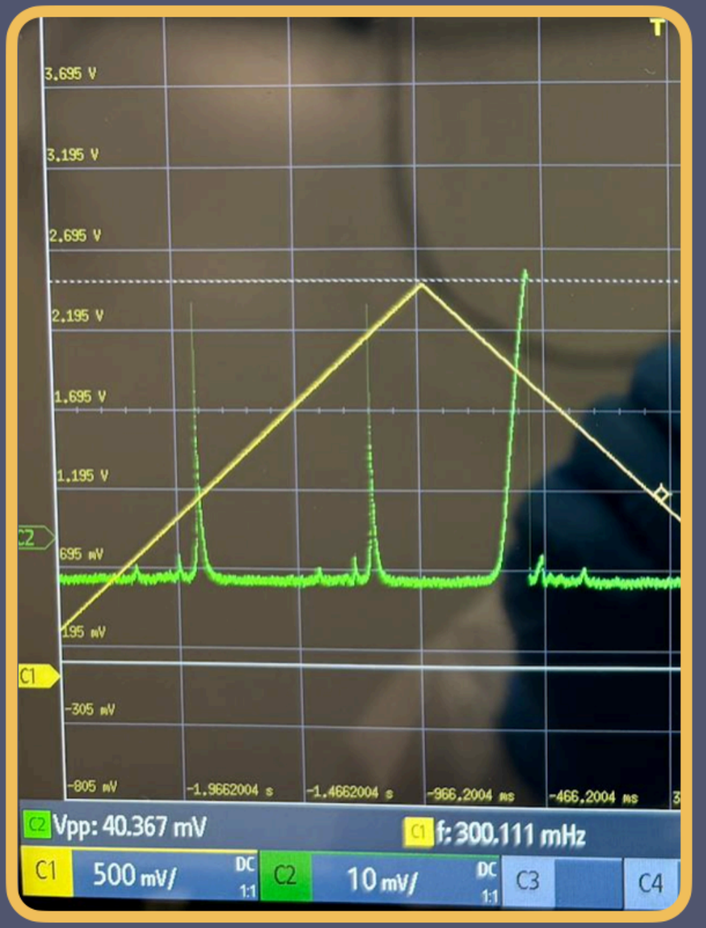
Currently Ring Down effect is managed by actuating with maximum force available.



[7] *"New algorithm for the Guided Lock technique for a high-Finesse optical cavity" D. Bersanetti et al. 2019*

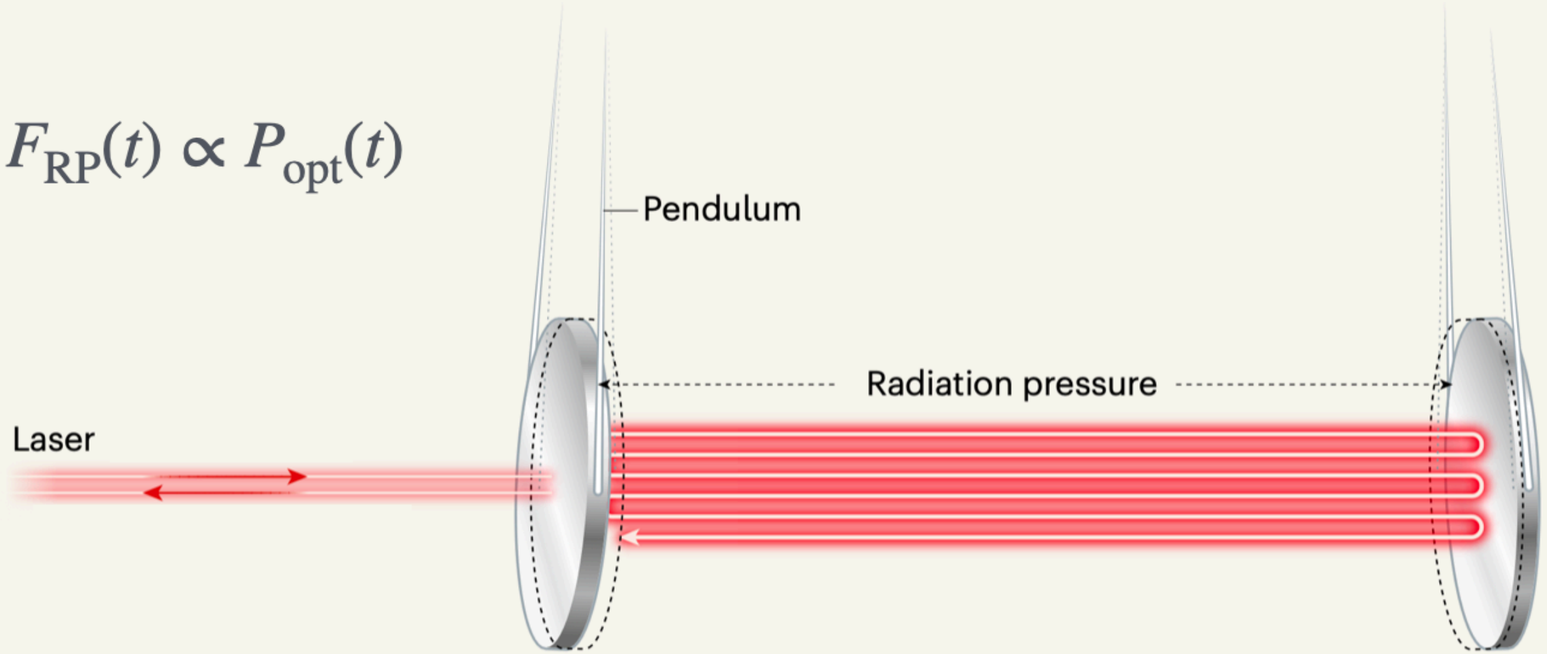
Backup


SQZ_SHG_err__TIME
volts
-0.005
-0.01
39m36
39m38
39m40
39m42
1402414793.0000 Jun14 2024 15:39:35 UTC
SQZ_SHG_out__TIME
volts
2
1
0
39m36
39m38
39m40
39m42
1402414793.0000 Jun14 2024 15:39:35 UTC
SQZ_SHG_trans__TIME
volts
0.05
0
39m36
39m38
39m40
39m42
1402414793.0000 Jun14 2024 15:39:35 UTC


3.695 V
3.195 V
2.695 V
2.195 V
1.695 V
1.195 V
C2
695 mV
195 mV
C1
-305 mV
-805 mV
-1.9662004 s
-1.4662004 s
-966.2004 ms
-466.2004 ms
C2 Vpp: 40.367 mV
C1 f: 300.111 mHz
C1 500 mV/ DC 1:1
C2 10 mV/ DC 1:1
C3
C4

Backup

$$F_{\mathrm{RP}}(t) \propto P_{\mathrm{opt}}(t)$$



In resonance condition, $F_{\mathrm{RP}}(t)$ will be at its maximum.