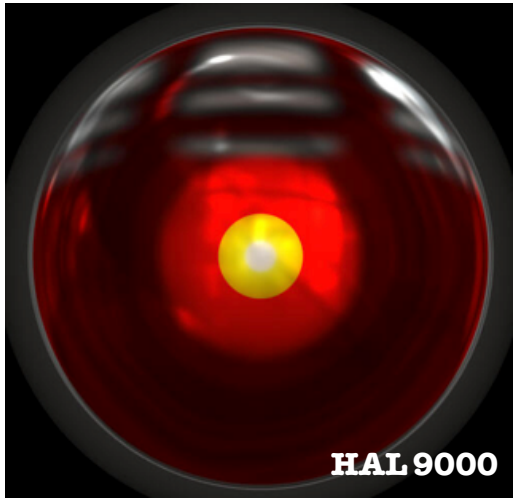
ATS AI/ML infrastructure needs

V. Kain for the ATS ML community forum

IT ML infrastructure workshop, 11-Oct-2023

Key AI/ML use case in ATS

- **Smart and agile accelerator exploitation:** AI for
  - Automatic optimisation and control
  - Preventive/prescriptive maintenance and fault analysis
  - Enhanced diagnostics
  - Advanced data-driven modelling: hysteresis compensation, kicker temperatures with intensity,…
  - Simulations into control room: fast executing surrogates
  - Optimised scheduling for accelerators and beam requests
- **Smart accelerator design:** AI for
  - Speeding up simulations: fully differentiable codes, AI solvers,…
  - Bayesian optimisation
- **AI assistants**
  - co-pilot for code development, finding and digesting documentation,…



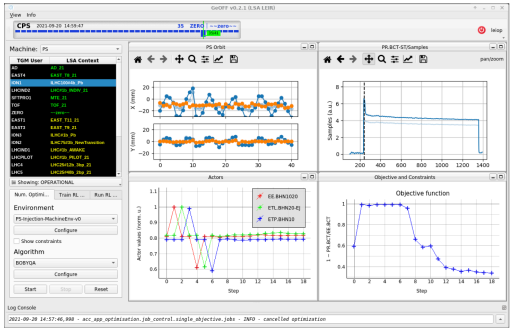
Key AI/ML use case in ATS

- **Smart and agile accelerator exploitation:** AI for — **First operational experience**
  - Automatic optimisation and control
  - Preventive/prescriptive maintenance and fault analysis
  - Enhanced diagnostics
  - Advanced data-driven modelling: hysteresis compensation, kicker temperatures with intensity,…
  - Simulations into control room: fast executing surrogates
  - Optimised scheduling for accelerators and beam requests
- **Smart accelerator design:** AI for — **Not started**
  - Speeding up simulations: fully differentiable codes, AI solvers,…
  - Bayesian optimisation
- **AI assistants** — **Pilot project(s)**
  - co-pilot for code development, finding and digesting documentation,…

Smart and agile accelerator exploitation

= AI/ML into the control room, on the tech net (TN) and in the controls middle ware

- Python in the control room and Python APIs to all databases and equipment communication
  - opensource community spirit: acc-py package index
- Store and share models: "machine learning platform" on K8s with GPUs
- UCAP* as online data processing framework
- Framework for optimisation and RL → GPUs on TN; GPUs on UCAP
- GPUs for offline training → interconnected GPUs to train transformers: VPCs, cloud, ml.cern.ch,…

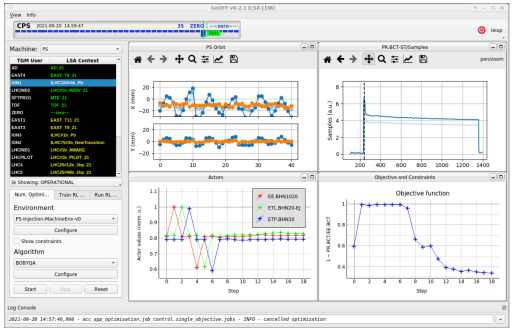


*UCAP=Unified Controls Acquisition and Processing framework

Smart and agile accelerator exploitation

= AI/ML into the control room, into the tech net (TN) and in the accelerator middle ware

- Python in the control room and Python APIs to all databases and equipment communication
  - opensource community spirit: acc-py package index
- Store and share models: "machine learning platform" on **K8s with GPUs**
- UCAP* as online data processing framework
- Framework for optimisation and RL → **GPUs on TN; GPUs on UCAP**
- GPUs for offline training → **interconnected GPUs to train transformers:** VPCs, cloud, ml.cern.ch,…



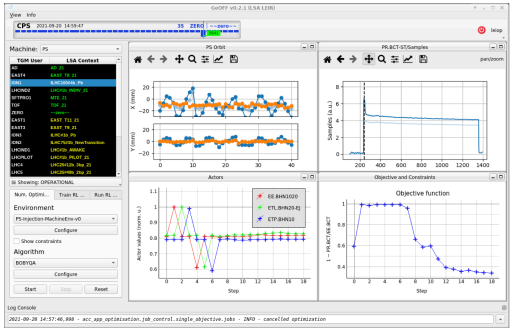
*UCAP=Unified Controls Acquisition and Processing framework

Smart and agile accelerator exploitation

= AI/ML into the control room, into the tech net (TN) and in the accelerator middle ware

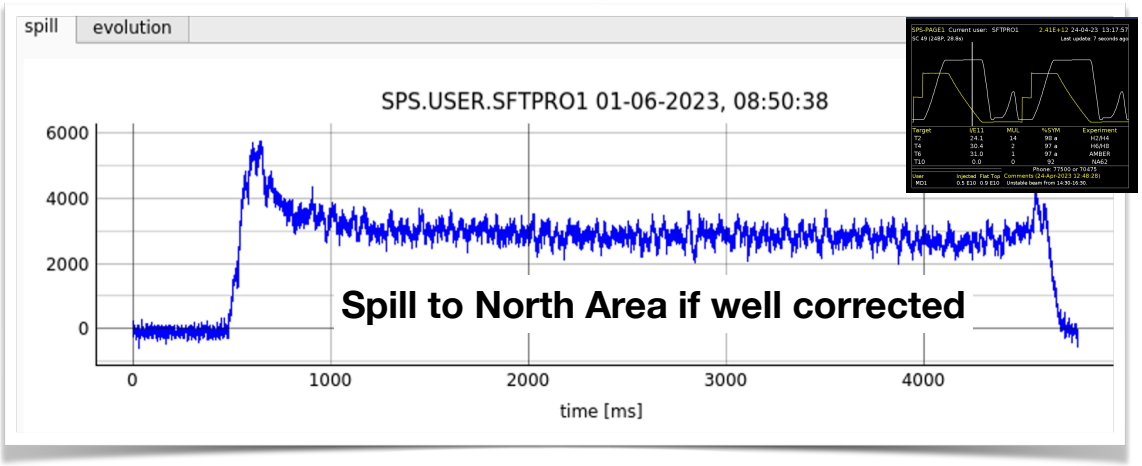
- Python in the control room and Python APIs to all databases and equipment communication
  - opensource community spirit: acc-py package index
- Store and share models: "machine learning platform" on **K8s with GPUs**
- UCAP* as online data processing framework
- Framework for optimisation and RL → **GPUs on TN; GPUs on UCAP**
- GPUs for offline training → **interconnected GPUs to train transformers:** VPCs, cloud, ml.cern.ch, …

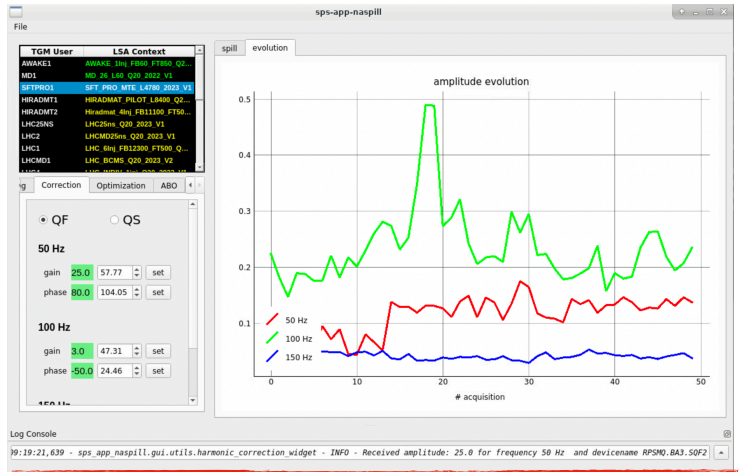
→ currently not even a handful of single GPUs available on TN for ML, "no" interconnected ones



GPUs in the control room - Example

$n \times 50$ Hz control of slow extracted spill to the North Area



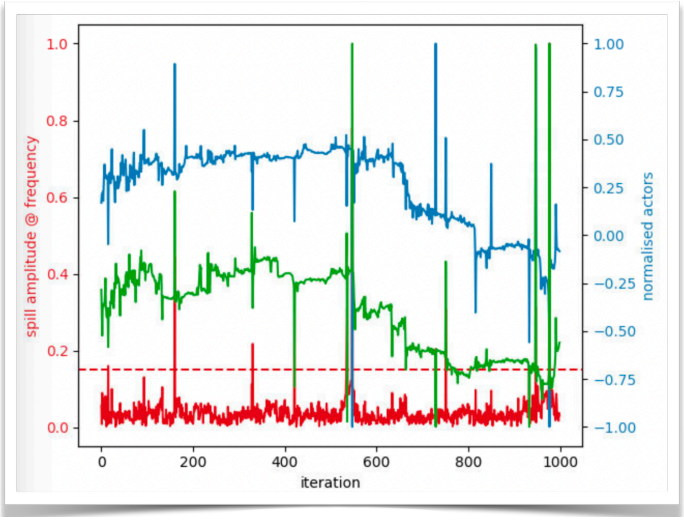


$n \times 50$ **Hz amplitudes over time: drifting**

→ Adaptive Bayesian Optimisation for continuous control

→ add dimension t to model and composite kernel including *SpectralMixtureKernel*: $\sigma^2 \times S(t, t') \times RBF(x, x')$

→ 2 controllers with GPU on UCAP (50 Hz, 100 Hz)



50 **Hz controller 17/8/'23**

On the horizon: proposal for **Efficient Particle Accelerators (EPA) project**

→ Automation for more efficiency, flexibility and reliability.

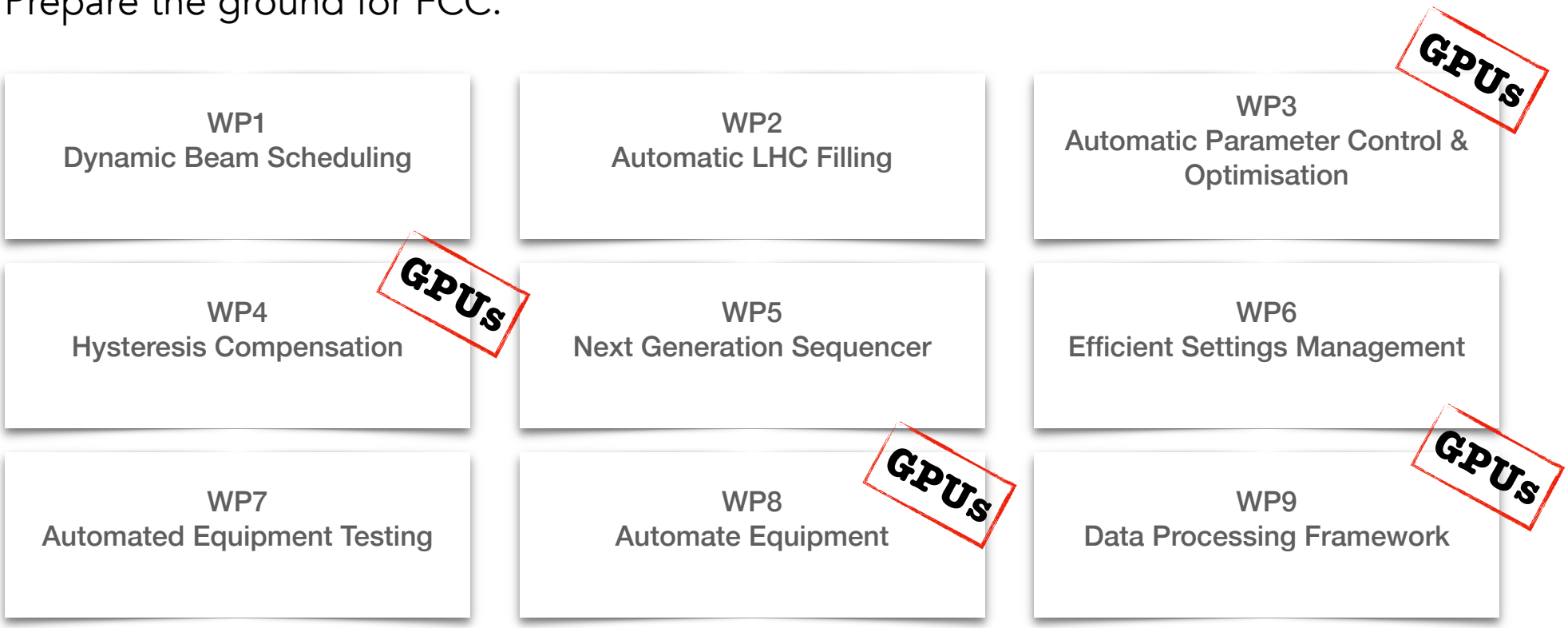
→ Prepare the ground for FCC.

WP1 Dynamic Beam Scheduling	WP2 Automatic LHC Filling	WP3 Automatic Parameter Control & Optimisation
WP4 Hysteresis Compensation	WP5 Next Generation Sequencer	WP6 Efficient Settings Management
WP7 Automated Equipment Testing	WP8 Automate Equipment	WP9 Data Processing Framework

On the horizon: proposal for Efficient Particle Accelerators (EPA) project

→ Automation for more efficiency, flexibility and reliability.

→ Prepare the ground for FCC.



Conclusion

AI/ML techniques finally arriving at CERN's particle accelerators.

- ATS has a lot of obvious use case!
- Various already prepared frameworks should ease development, deployment and maintenance
- The current limitation is availability of GPUs.
  - Collaborating with ATS-IT GPU initiative team, some GPUs should arrive → but orthogonal use case in general (batch system versus interactive setup)
  - Ideally have powerful enough GPU cluster to train (and serve) GPTs.
- Open to cloud solutions, but only useful for offline use cases.