Natural job drainage and power reduction in PIC Tier-1 using HTCondor

J. Flix, K. Fabrega, G. Merino, C. Acosta, J. Casals

WLCG Environmental Sustainability Workshop

CERN

11-13 Dec. 2024

PIC port d'informació científica

IFAE Ciemat RES RED ESPAÑOLA DE SUPERCOMPUTACIÓN España | digital 2026 Unió Europea Fons Europeu Next Generation GOBIERNO DE ESPAÑA Plan de Recuperación, Transformación y Resiliencia Next Generation Catalunya Generalitat de Catalunya Departament de Recerca i Universitats GOBIERNO DE ESPAÑA VICEPRESIDENCIA PRIMERA DEL GOBIERNO MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Introduction

Preliminary studies and ideas to understand **natural job drainage** and **power reduction** in PIC Tier-1, which is using HTCondor

Analyze historical logs from HTCondor to **understand natural job drainage** patterns: when jobs naturally conclude without external intervention

This analysis could reveal particular **patterns** (or a lack thereof) in job drainage, while also providing **insight into expected levels of resource reduction over time**

- It would help us understand how quickly we could scale the farm to adapt to external factors, such as green power availability cycles
- K. Fabrega (last year Physics degree student) involved in these studies

Live demo ⟶

Characteristics of Natural Drainage Cycles

PIC port d'informació científica

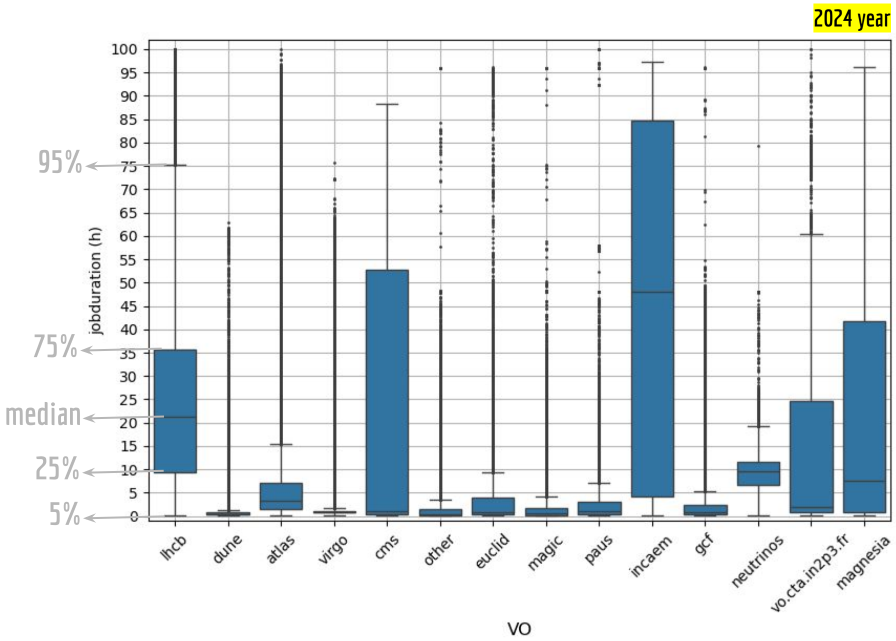
Conduct **extended simulations to observe drainage patterns** under varying load conditions

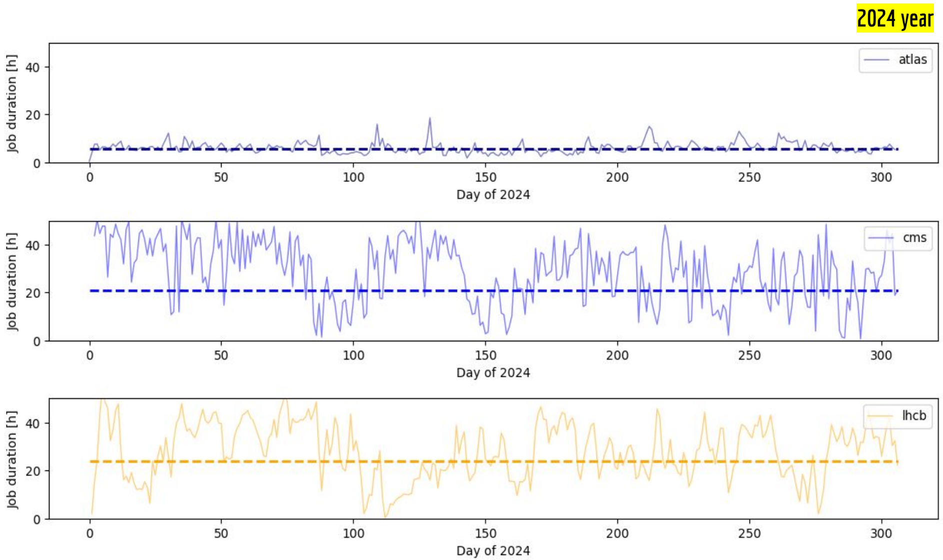
Analyze the **impact of job types** and **VO-specific job durations** on natural drainage cycles, examining how different workloads influence these cycles

These characteristics can be effectively studied by **simulating multiple natural drainage** scenarios over the specified time period

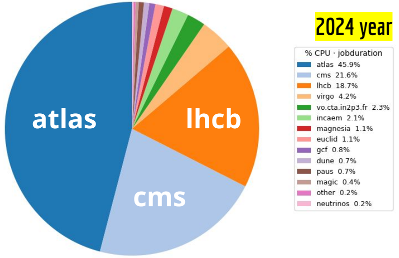
- Using 2024 PIC HTCondor historical data, we performed drainage simulations at hourly intervals

Characteristics of Natural Drainage Cycles



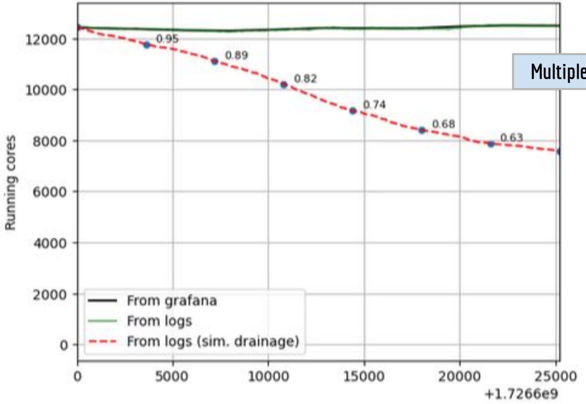


VO-specific job durations variability will impact on the natural drainage scenarios

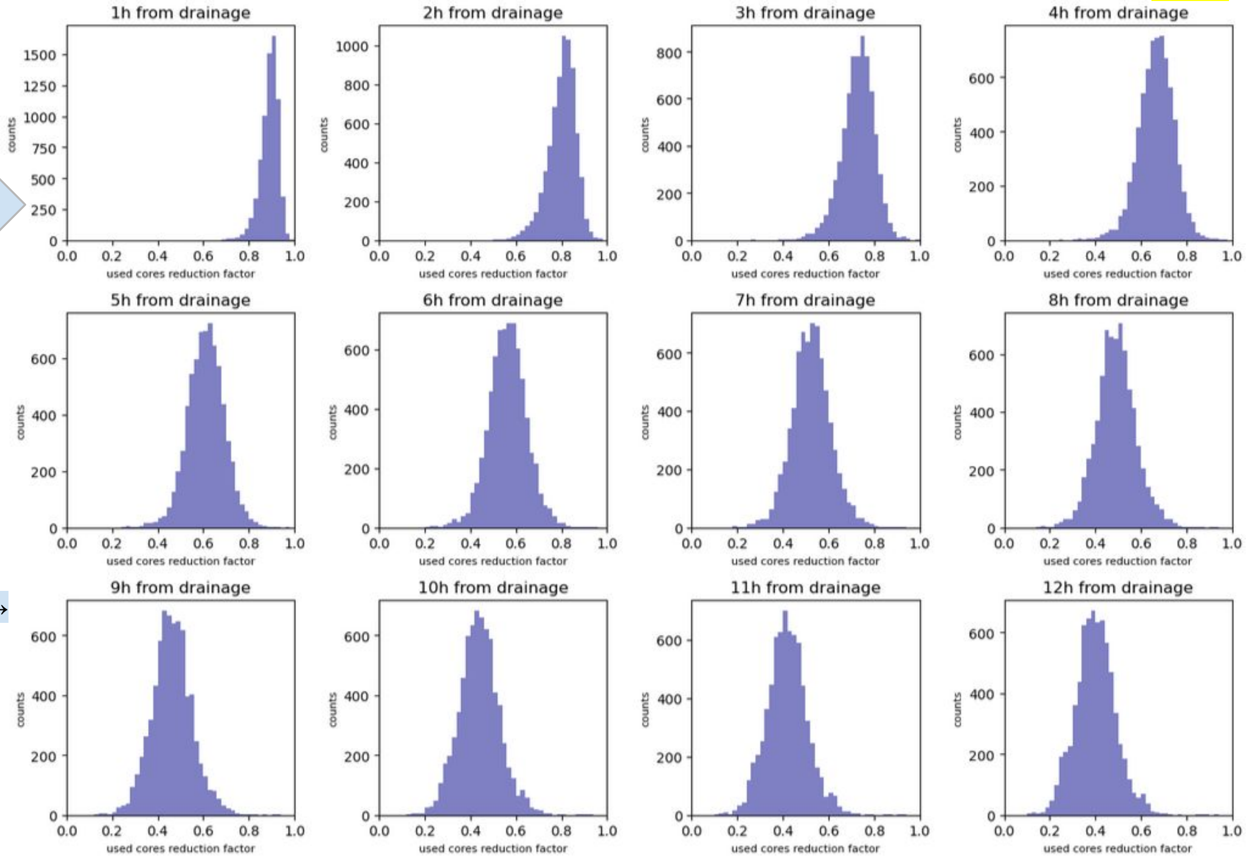


Characteristics of Natural Drainage Cycles

2024 year



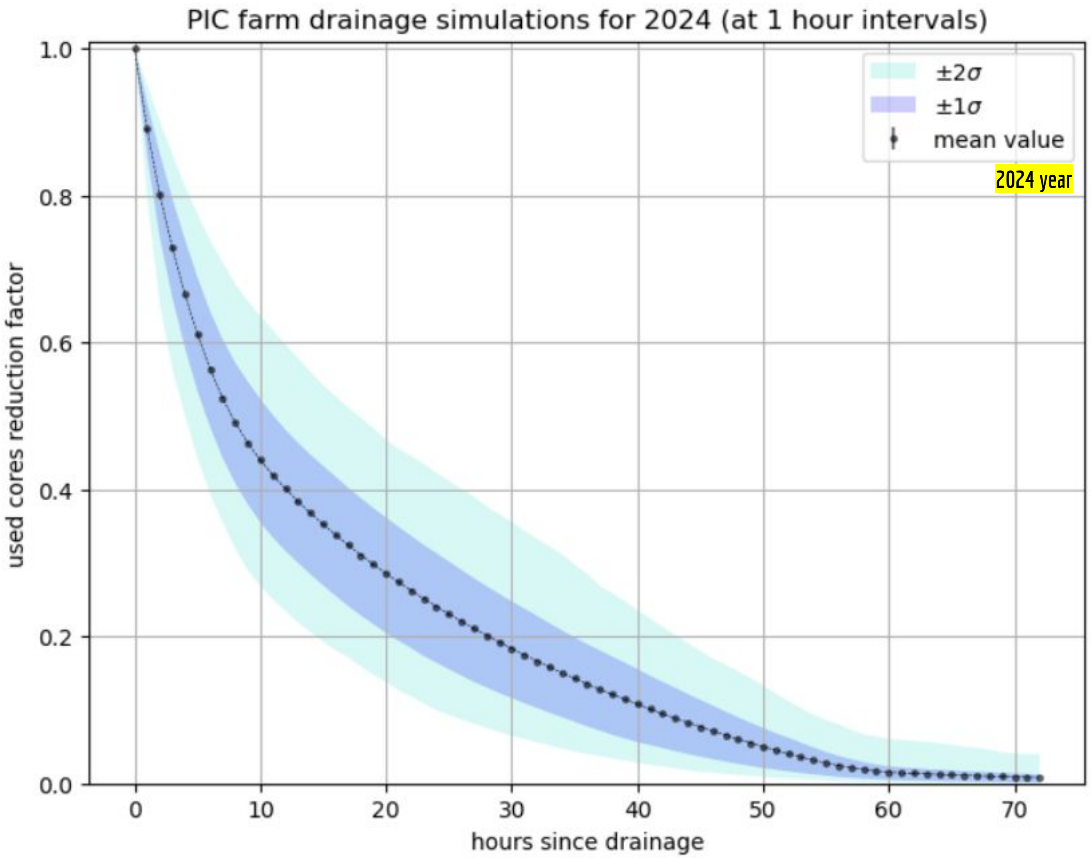
Multiple sim.



Drainage simulations at hourly intervals →

Reduction factors for utilized cores observed at various hours following drainage scenarios

Characteristics of Natural Drainage Cycles



Watts before and after drainage

PIC
port d'informació
científica

Estimate **power consumption** before, during, and after natural drainage events

The PIC farm compute nodes report power consumption through ***IPMItool***, which allows for power monitoring per node on local Grafana portal

Using HTCondor data, we can identify the compute nodes where jobs are being executed. This enables the **calculation of power consumption based on node occupancy**, before, during, and after natural drainage events

For this we need to **characterize the power consumption of a node based on its occupancy** and then apply it to our simulations

- Draining before downtimes or security updates, so we can use that information to characterize Watts vs. occupancy with real jobs

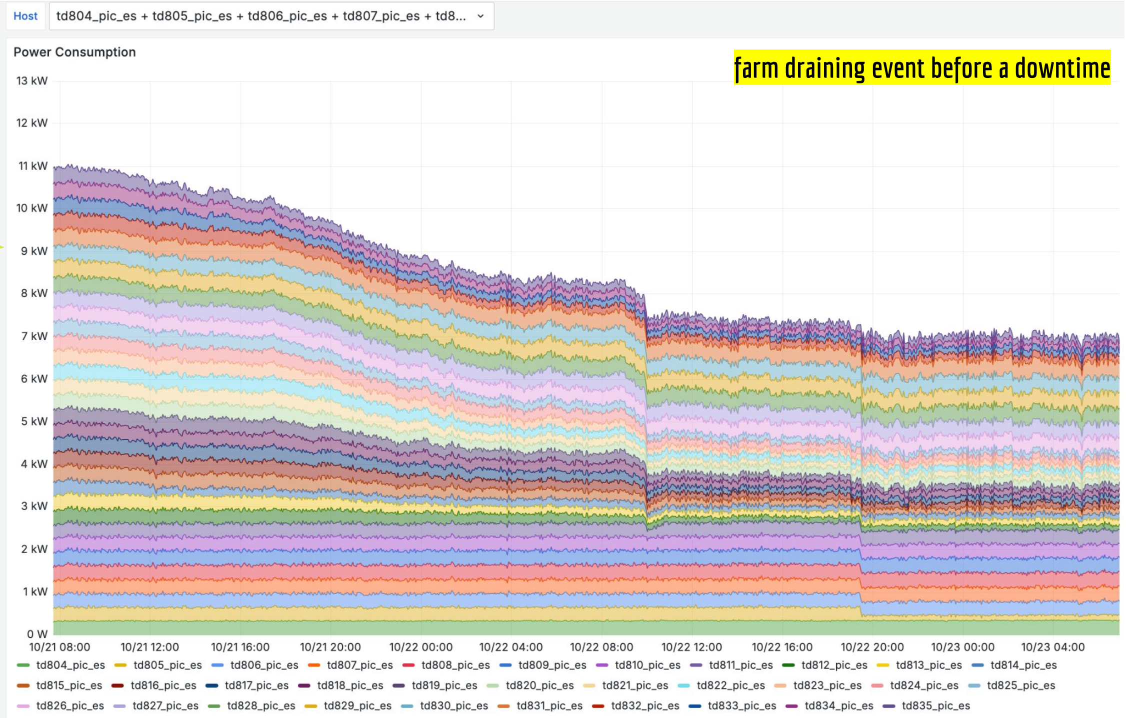
Watts before and after drainage

PIC port d'informació científica

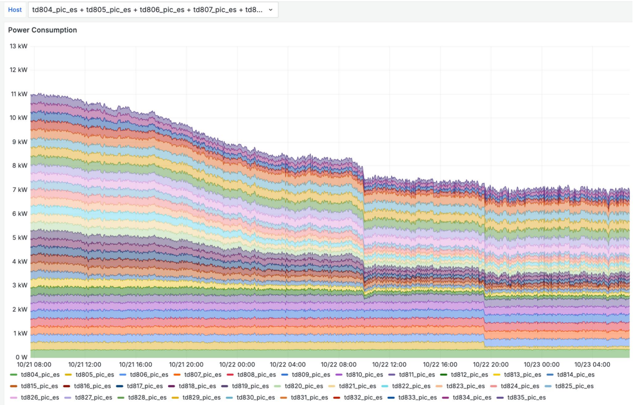
HTCondor Worker Nodes

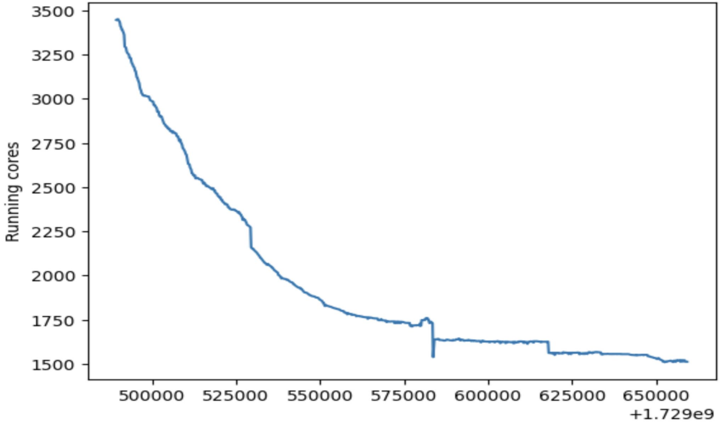
CPU Type	Number of Nodes	Numbre of Slots	HS06
E5-2640v3 Alma9 (1.2554)	33	792	12130.1769
E5-2680v4 Alma9 (1.2996)	53	1696	26890.2835
EPYC-7452 (1.0077)	32	3980	48929.8812
EPYC-7502 (1.0739)	44	5632	73788.0985
EPYC-7662 (0.8278)	2	256	2585.3849
gpu01 (1.1985)	1	49	716.4633
gpu02-gpu03 (1.0597)	2	24	310.2801
gpu05 (1.6258)	1	48	952.0684
tdm002 (1.0419)	1	48	610.1366
TOTAL	**169**	**12525**	**166912.7734**

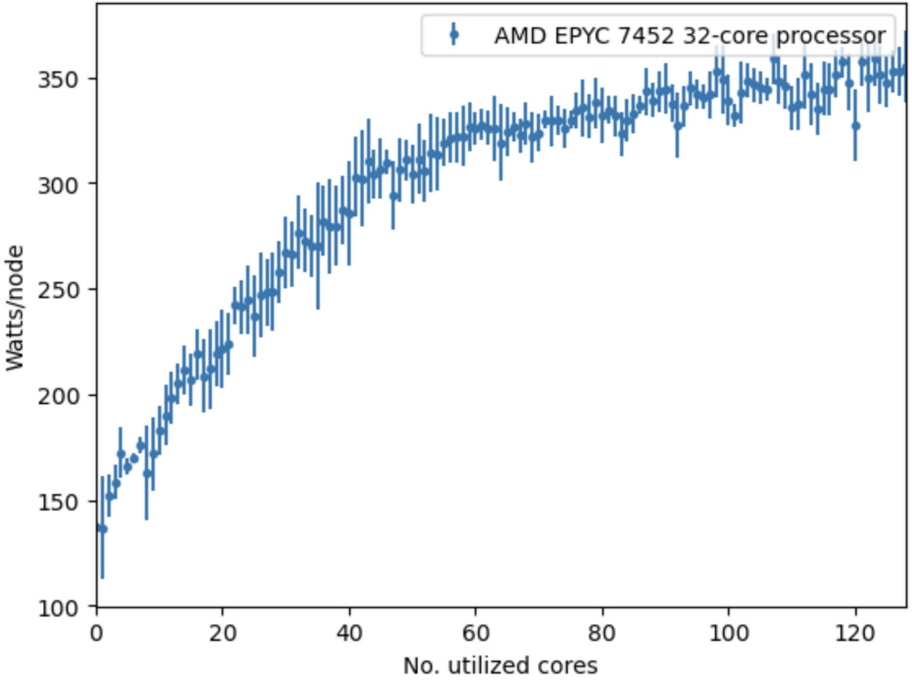
"AMD EPYC 7452 32-core processor" dual-CPU with hyperthreading enabled → 128 cores per node



Watts before and after drainage

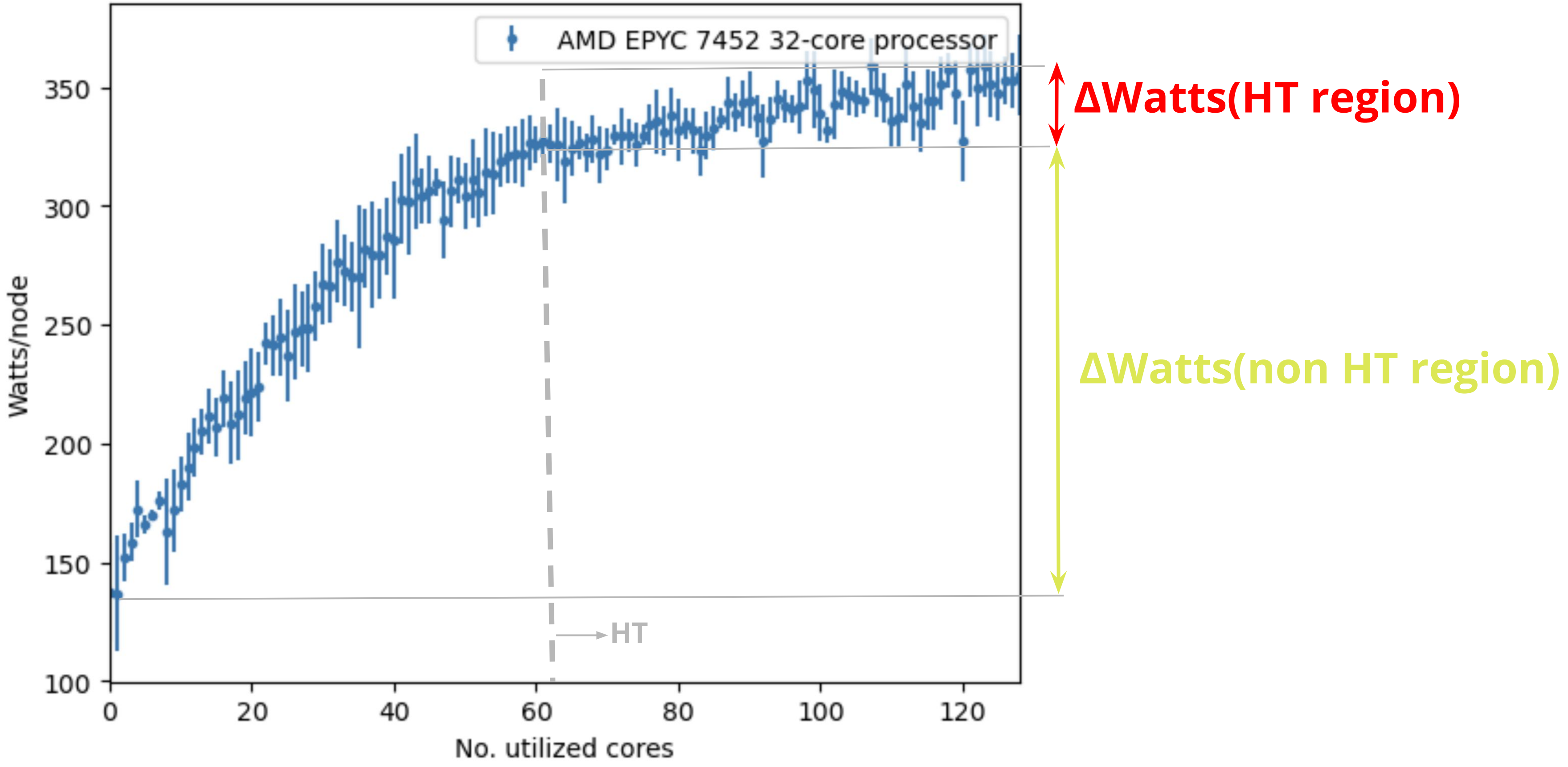




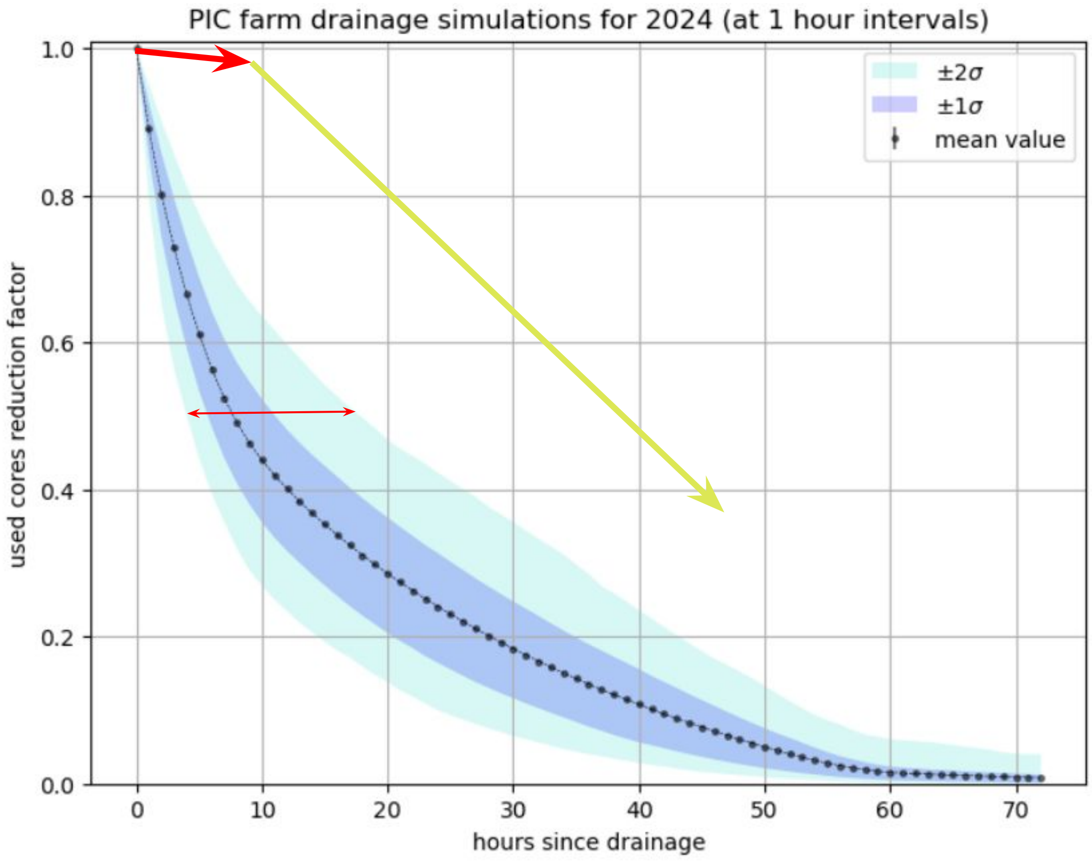


More statistics can be added including
other PIC farm draining events

Watts before and after drainage



Watts before and after drainage



Since natural draining occurs randomly across compute nodes, **a significant reduction in power consumption is not expected immediately after draining begins**

Noticeable **reductions are likely only when node occupations fall below the hyper-threading (HT) regions**
→ this would limit capability to sites to modulate farm utilization to save energy or adapt to clean energy cycles

We will model this reduction with the IPMI information

Next steps

Analyze in detail the **impact of job types** and **VO-specific job durations** on natural drainage: how different workloads contribute to natural drainage cycles

Estimate **power consumption** before and after drainage (*ipmitool*), for all of the compute nodes available at PIC (and for those that have been retired)

- We can use natural drainages that actually happened in the past at PIC, prior to downtimes, to model Watts vs. occupancy for all CPU types

Next steps

PIC
port d'informació científica

Analyze in detail the **impact of job types** and **VO-specific job durations** on natural drainage: how different workloads contribute to natural drainage cycles

Estimate **power consumption** before and after drainage (*ipmitool*), for all of the compute nodes available at PIC (and for those that have been retired)

- We can use natural drainages that actually happened in the past at PIC, prior to downtimes, to model Watts vs. occupancy for all CPU types

Develop **machine learning models** for predictive power scaling

Evaluate **potential carbon emission reductions** in these drainage scenarios

Design a **feedback loop to HTCondor for real-time power modulation** based on green energy cycles

Preliminary conclusions

PIC
port d'informació
científica

We are **modeling natural job drainage** and **power reduction** in the PIC Tier-1 system using historical HTCondor data combined with compute node power consumption records

Our goal is to **develop machine learning models for predictive power scaling**. If successful, this approach could identify scenarios where compute nodes can be quickly drained from the hyper-threading (HT) region, enabling more efficient adaptation to external factors

The objective is not to terminate jobs prematurely, so **'preemptible' jobs could facilitate more efficient drainage processes**. This would allow for better modulation of farm utilization while maintaining operational efficiency, and maybe this is something WLCG experiments should consider at some point

Acknowledgements

The authors of this work express their gratitude to the PIC and CIEMAT teams for their support in these studies and for deploying novel cache services for the CMS experiment in the Spanish region. This project is partially financed by the Spanish Ministry of Science and Innovation (MINECO) through grants FPA2016-80994-C2-1-R, PID2019-110942RB-C22 and BES-2017-082665, which include FEDER funds from the European Union. It has also been supported by the Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21, the Catalan government under contract 2021 SGR 00574, and the Red Española de Supercomputación (RES) through the grant DATA-2020-1-0039.

CosmoHub has been developed by PIC (maintained by IFAE and CIEMAT) in collaboration with ICE-CSIC. It received funding from the Spanish government (grant EQC2021-007479-P funded by MCIN/AEI/10.13039/501100011033), the EU NextGeneration/PRTR (PRTR-C17.I1), and the Generalitat de Catalunya.

Thanks!
Questions?