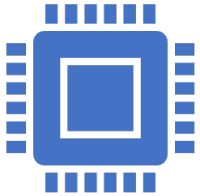
AI in operations at SURF

Damian Podareanu

ai@surf.nl

What do we understand at SURF by AI in operations

Use predictive tasks and analyse operational data from various services to gain valuable insights, bring energy savings, automate processes, and optimise and inform decisions.



Data collection methods are enabled at user, system, and infrastructure levels.

We want to help system operators and users.

Overview of Predictive Tasks

Anomaly Detection: Identifying abnormal patterns or behaviours in the system to detect potential issues or security threats.

Capacity Planning: Predicting future resource requirements and optimising system capacity to ensure smooth operations.

Utilisation Prediction: Forecasting resource utilisation trends to optimise resource allocation and improve system efficiency.

Security Flagging: Analysing system data to identify potential security vulnerabilities and suspicious activities.

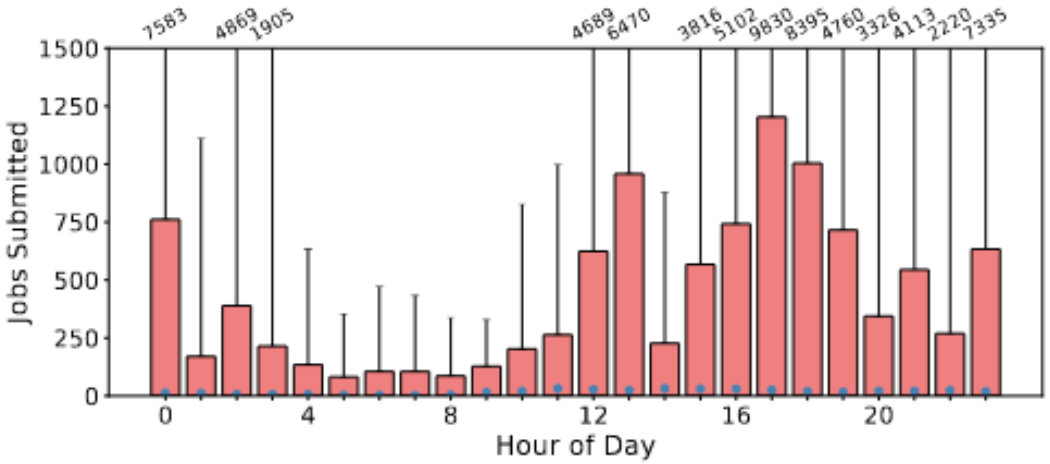
Language Automation for Support: Automating customer support interactions using natural language processing and machine learning techniques.

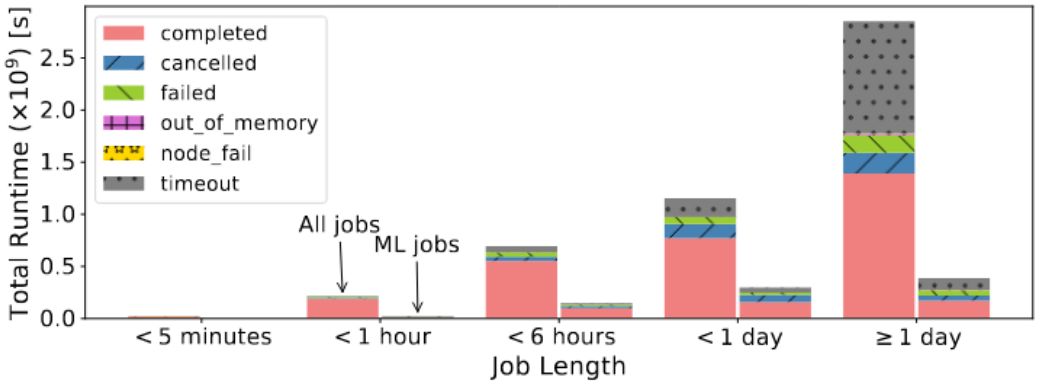
Hardware Tuning for Energy Savings: Analysing system data to optimise hardware settings and configurations for energy efficiency.

Infrastructure Insights: Gaining a holistic view of the system infrastructure for identifying areas of improvement.

Advanced Reporting: Generating comprehensive reports and visualisations to present key system metrics and insights.

Use case: Compute Utilisation prediction



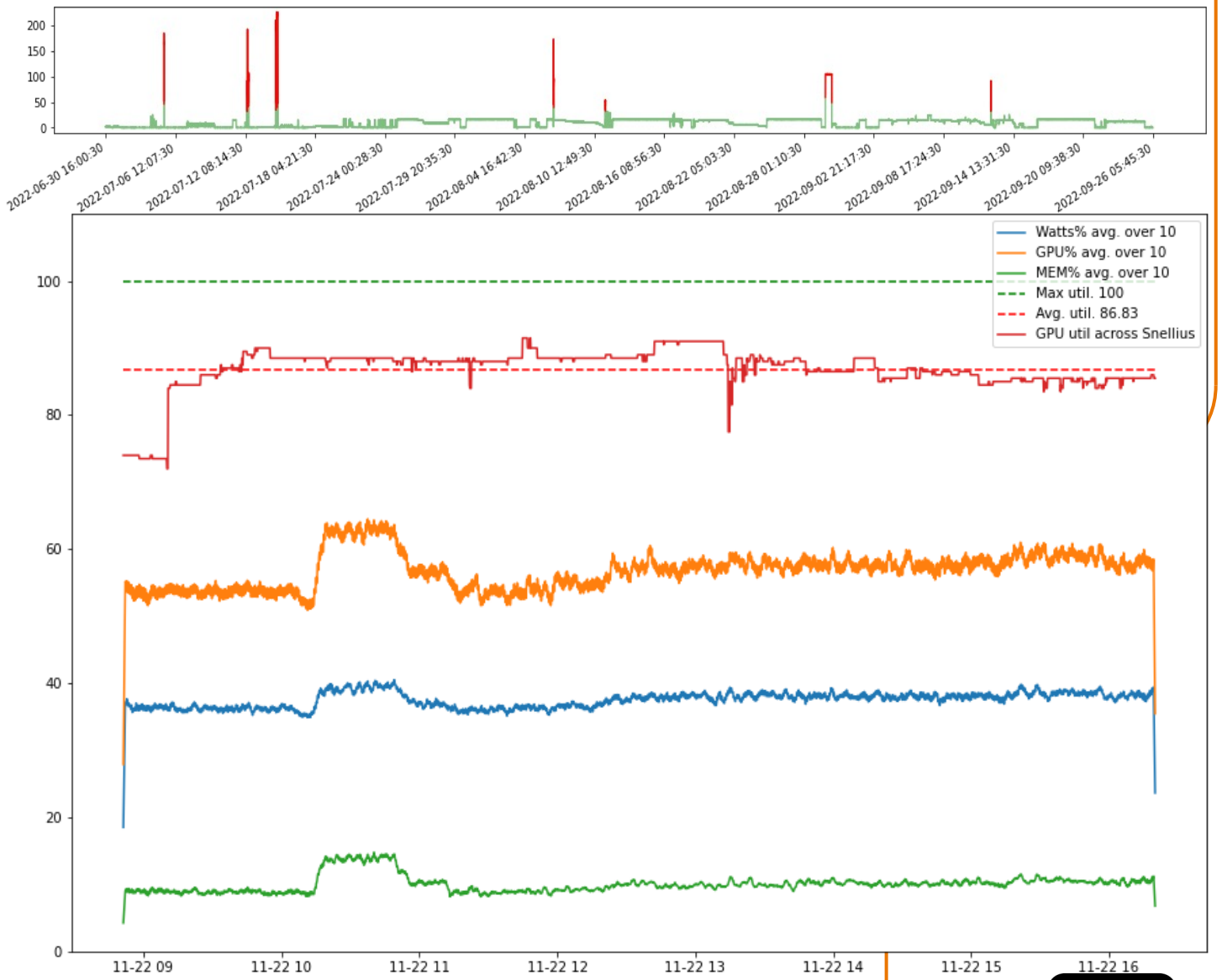


- Good users, bad users
- Beyond seasonal trends
- Predicted load, power consumption
- Efficient idle node shutdown
  - Combine SLURM + Prometheus
  - Less often used nodes/queues can be turned off more often
  - Save power by turning off nodes

https://zenodo.org/record/3878143 https://arxiv.org/pdf/2107.11832.pdf

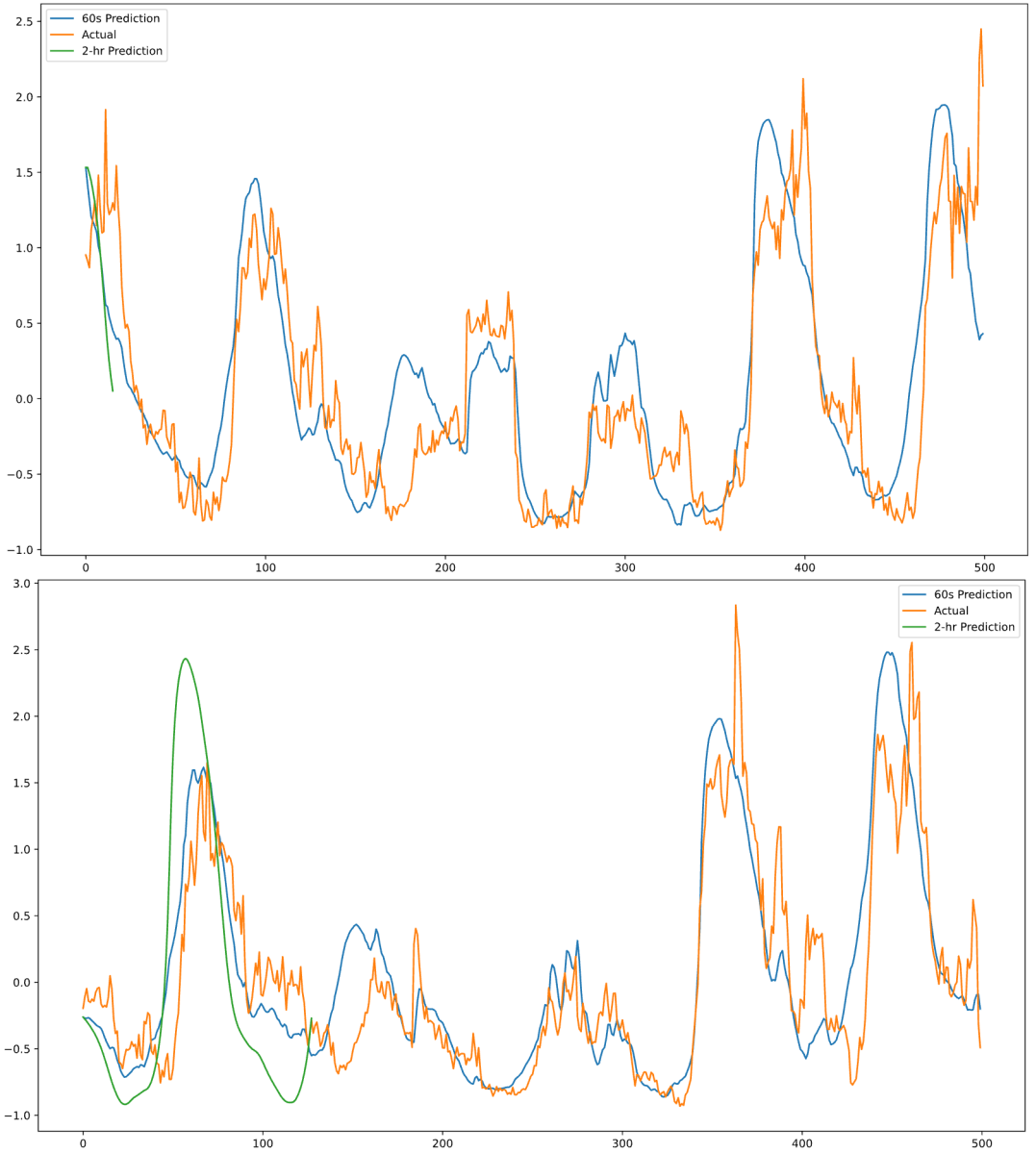
Use case: Platform insights

- Detect anomalies
- Identify cross-system effects
- Detect inefficient library versions
  - Combine XALT + Prometheus
  - Two versions of the same library have different CPU utilization → look into it
  - Automated user feedback

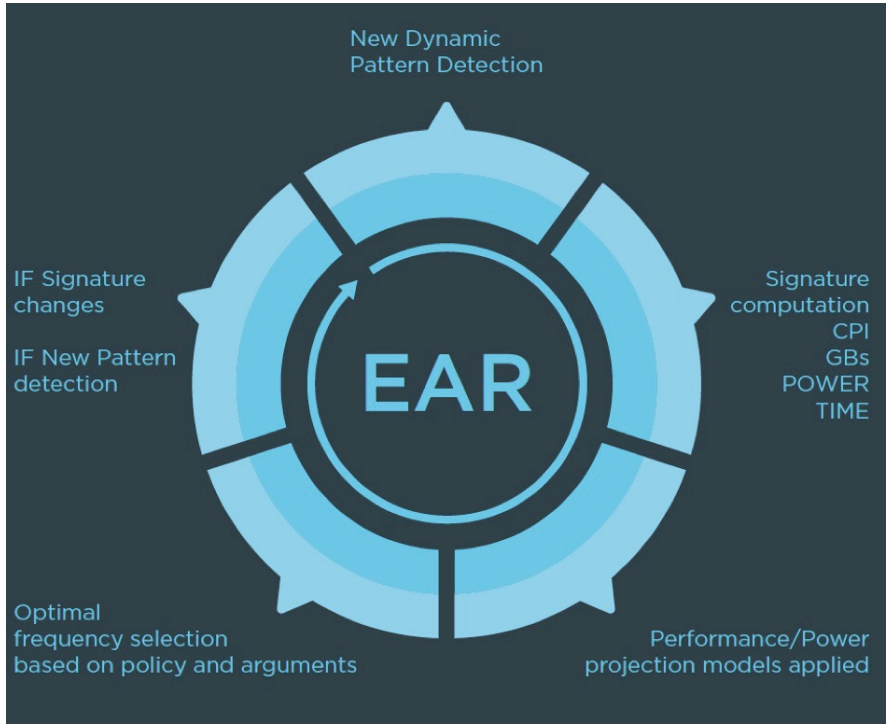


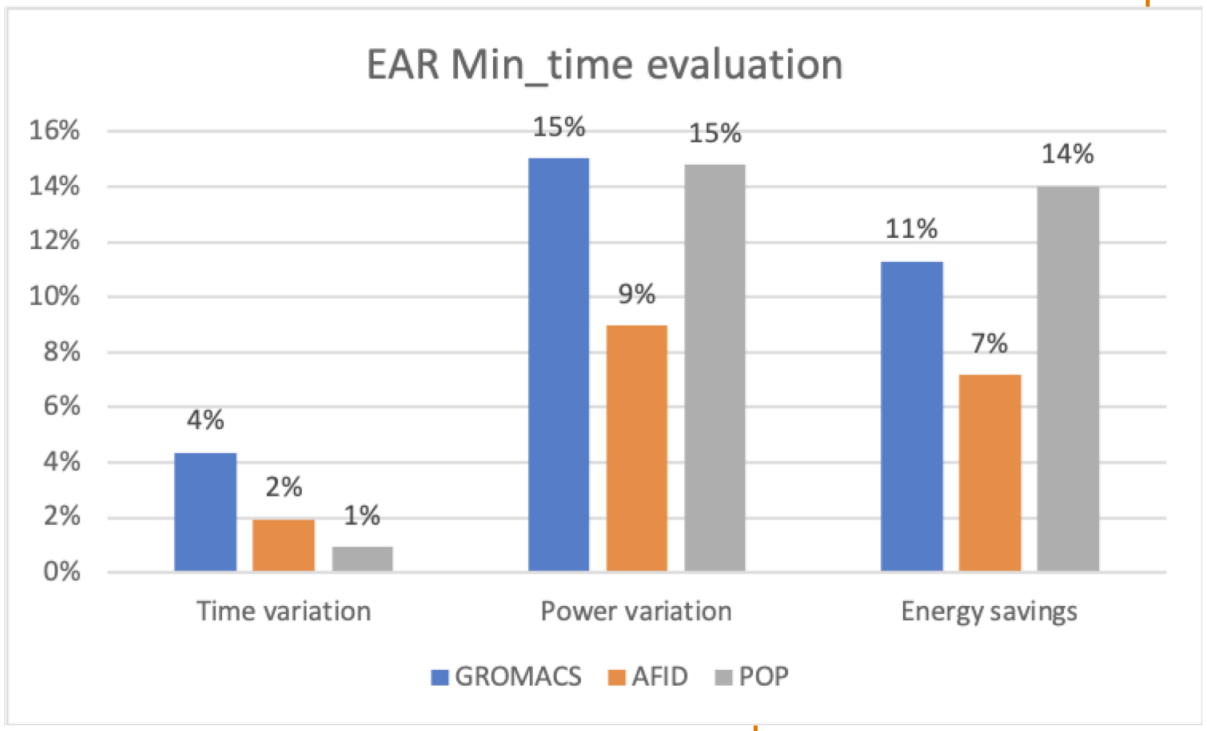
Use case: Network Utilisation prediction

- Dynamic nature of network traffic, which can exhibit significant fluctuations and unpredictability
- Modern networks' increasing complexity and scale make capturing and analysing all relevant data challenging.
- Factors such as diverse network topologies, varying traffic patterns, and anomalies further complicate the prediction process.
- More historical data or adequate monitoring infrastructure needed to ensure the accuracy and reliability of network utilisation predictions.
- Addressing these issues requires advanced data analytics techniques, robust modelling approaches, and a comprehensive understanding of network dynamics.



Use case: Energy aware runtime (EAR)



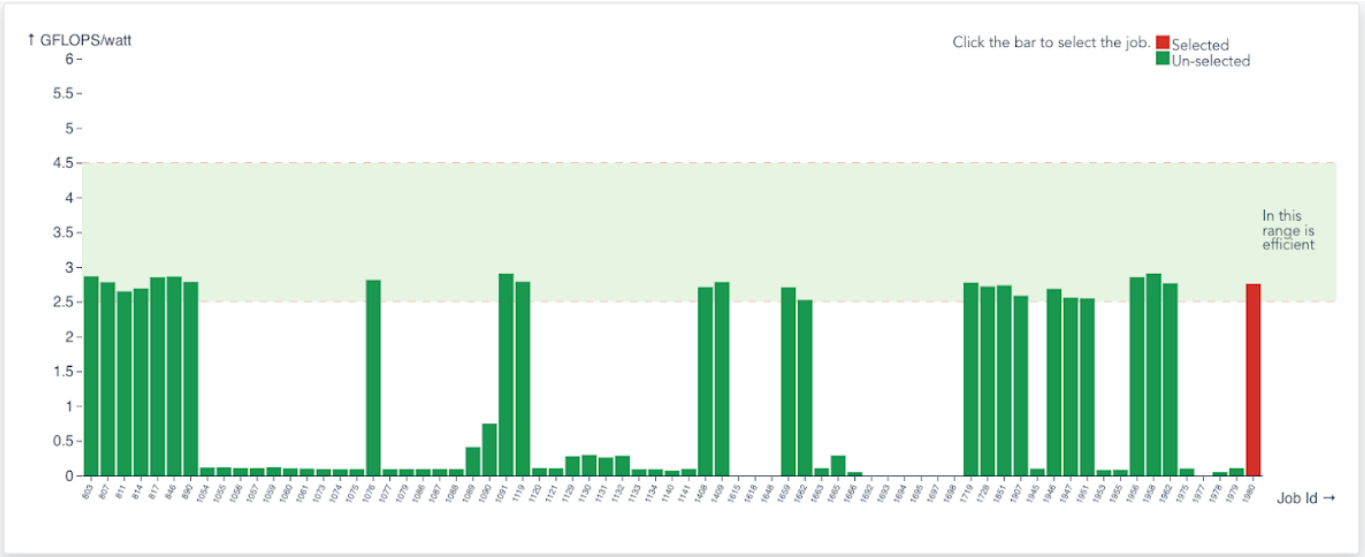


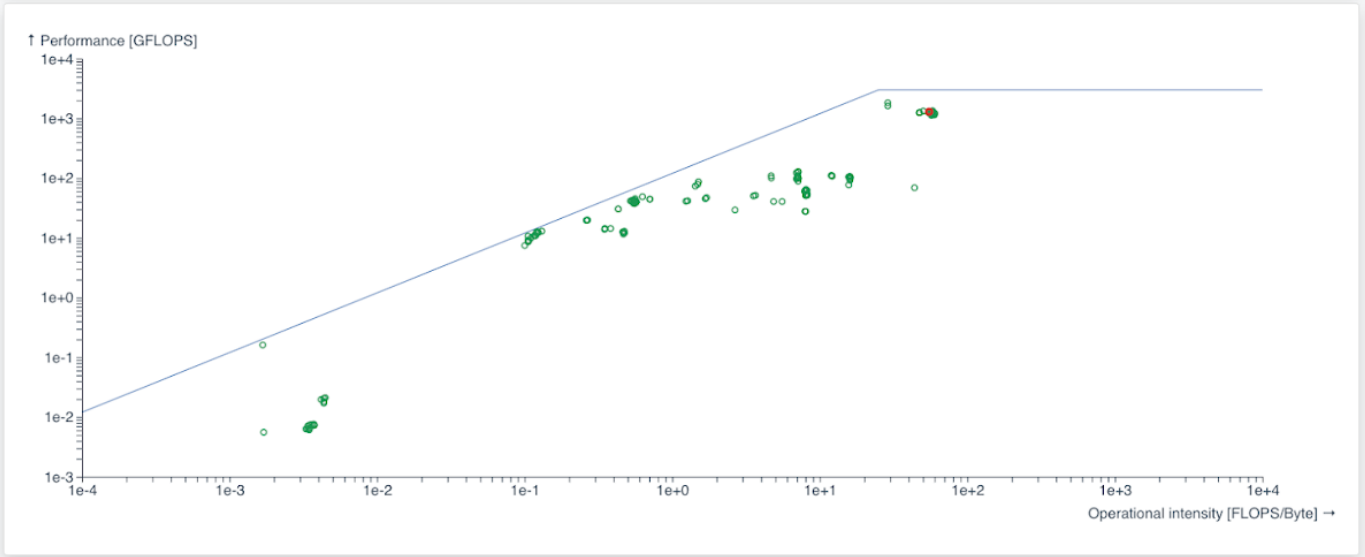
Relevant application use cases inspired by SABS

- Experiments executed in Lenovo SD530 system
- Skylake 6148 @2.4GHz 20c nodes with EDR network
- Default frequency=2.0GHz
- GROMACS 640 processes. 16 nodes
- AFID 600 processes. 15 nodes
- POP 400 processes. 10 nodes

Average energy savings of 10%

Use case: Job insights





Recommendations

Under construction.

Energy consumption

For job with id: 1980

House holds:

0.000035

Electric car distance:

0.474511 km

Solar panel area:

0.000565 m^2

Query job information

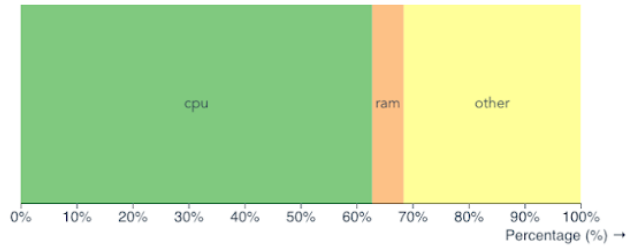
User name

sagard

Job ids

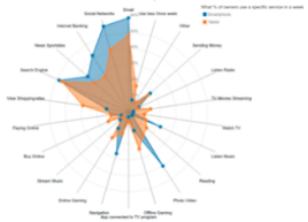
Remove username

Submit

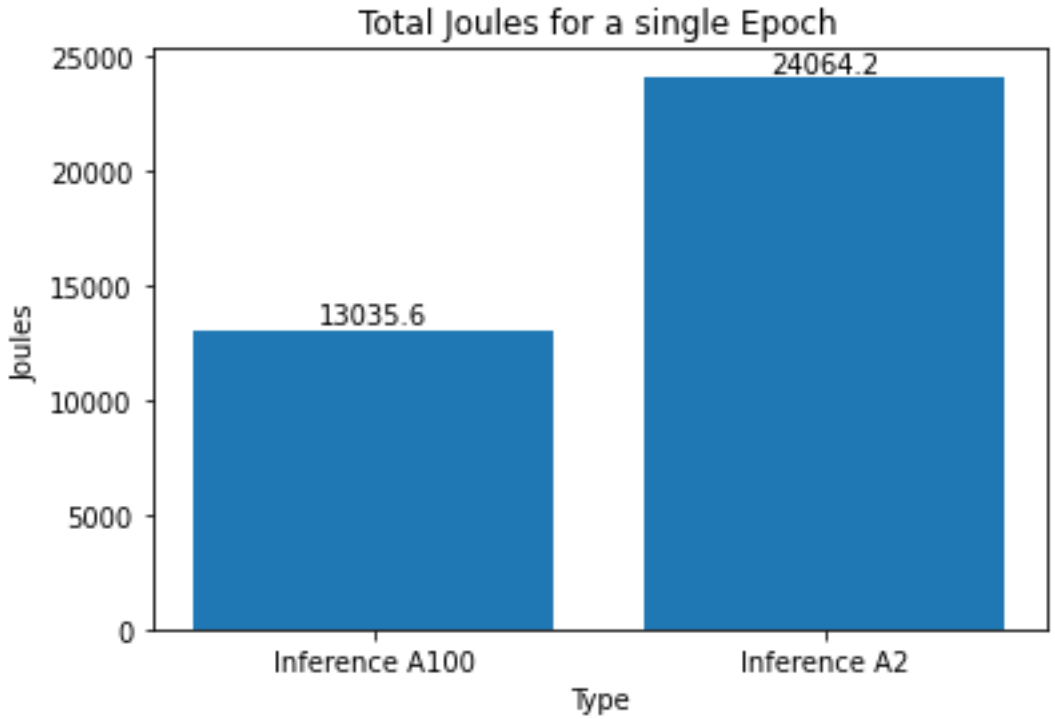


Application characteristics

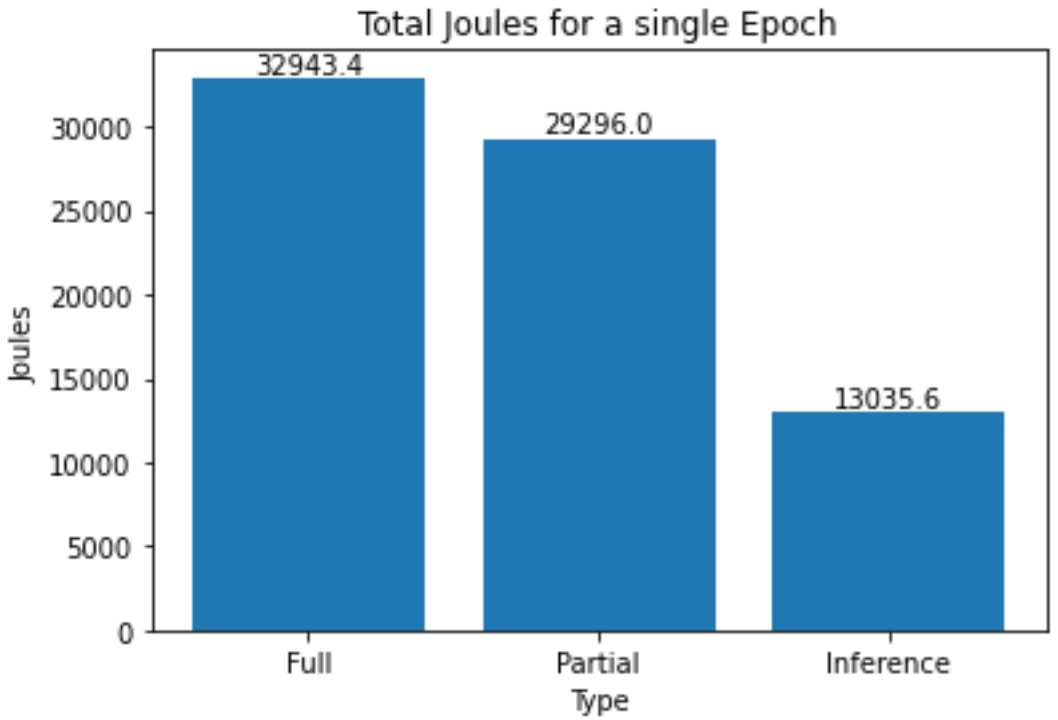
Under construction, placeholder image:



Use case: Energy-aware job insights (EAR)



Language Model efficiency with different configurations on A100 GPU



Language Model efficiency on different GPUs

Compare CPU/GPU energy efficiency

Inform users in their energy usage

Get “free” power savings

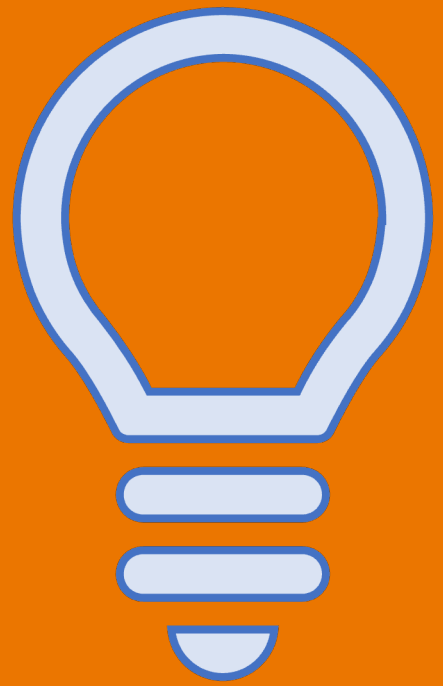
Simple comparison with different hardware

Use case: Helpdesk automation

- Ingest anonymised ticket database
- Give helpdesk support lines additional information by summarising and linking relevant existent information
- Create an interactive Q&A
- Reduces the number of support requests
- Can aid in knowledge transfer
- Difficult privacy concerns

Challenges and Limitations

- Data privacy, scalability, and accuracy of predictions
- Energy efficiency is increasingly important
- Data is already collected on the infrastructure → Should be explored
- Any small optimisation can save a lot of power when working on a large scale.



ai@surf.nl

Driving innovation together!