ErUM data footprint WG

Jun 1, 2023

G. Duckeck
N. Eich
J. Erdmann
S. Neuhaus
M.Schumacher

Disclaimer

- 8 ErUM communities:
  - KAT, KET, KfB, KFN, KFS, KFSI, KhuK, RDS
- WG members: 4 x KET, 1 x KhuK
  - Somewhat biased view and experience
- Focus on computing / digital technologies related footprint
  - No accelerator, detector, lab, instruments, infrastructure, travel related activities
-

Where does quantitative knowledge exist?

~ok

- Large organized computing efforts beyond institute level:
  - WLCG (KET&KHuK, LHC experiments, Belle2): Simulation, Reconstruction and Analysis of data from large collider experiments
    - Monthly CPU time monitoring per site and experiment (EGI/WLCG)
      - Data public and can be used to estimate CO2 footprint
  - ATLAS experiment:
    - detailed per job info on power consumption and local CO2 footprint
      - (see RW talk on Wed)
  - Large theory projects on Gauss/Prace HPC systems (e.g. Lattice QCD, Cosmology simulations, ...)
    - Formal project proposals with detailed CPU allocation & usage record
      - Data can be used to estimate CO2 footprint (how to find??)

Where does quantitative knowledge exist?

~partial

- Groups with large processing activities at big centers
  - XFEL@Desy (O(50 PB data) and large processing capacities),
  - Theory groups at computing centers (eg MPP@MPCDF)
  - smaller KET experiments
  - Other ErUM communities (KAT, RDS, KFN, KfB, ...)
- Online farms of big experiments (similar scale as WLCG)
- For above cases accounting data from batch systems should be available
  - Mostly no public/standardized access
- Storage, networking and services
  - No systematic records with public access
    - Individual info by data centers on power share
      - WLCG Cern data center: disk ~21%, services 17%, network 5% (Chep talk)
    - No info at level of individual data-set transfer&storage
      - Eg what is the power/CO2 to transfer a 1 TB dataset from A to B and store it on HDD for 6 month

Where is it lacking?

- Research groups and institutes
  - Desktop clusters
  - Basement servers
  - Local GPU nodes for ML
  - ...
- Manufacturing&disposal footprint
  - Some studies available (eg Dell) but large span/uncertainties
  - will probably vary a lot given the local energy mix at the factory

Opportunities for savings?

- HW Technology progress
- SW/algorithm optimization
- Different architectures (GPU, ARM, …)
- ML optimization
- Dynamic power provision
- Lifetime extension
- Minimize losses

- **All discussed in other WGs in more detail**

-

Raise awareness

- Provide users with info for effect of jobs
  - Footprint should be transparent
  - Though steering choices on world-wide Grid must be restricted
- Provide developers/experts with incentive for savings
  - sustainability benchmarks for code&tools
  - ML competitions: not just score but also footprint

    …
- Provide managers/convenors with info on footprint for planned reconstruction/simulation campaign
  - and keep record of (accumulated) footprint for datasets
- Include footprint estimate in publications/thesis/...
- ‘Education process’ at all levels to take footprint into account

Suggestions

- Contact all ErUM communities for feedback on ErUM data related footprint in their domain
- Develop sustainability benchmarks for reco/multi-threading/ML
- Untracked institute clusters
  - Provide tools for recording (digital power meters, grafana monitoring package)
  - Or provide incentive/alternative offers, e.g. renting slots in science cloud
- All data centers used by ErUM communities should provide power/CO2 tracking
  - Ideally in standardized format
- Aim to make fine-grained footprint information available
  - job/task/dataset level