Sensitive data challenges on HPC

Powered by CSC

LUMI

European flagship supercomputer

EOSC-Nordic sensitive data workshop
4.10.2022
Jarno Laitinen (CSC) et al.

www.lumi-supercomputer.eu #lumisupercomputer #lumieurohpc

LUMI is an HPE Cray EX Supercomputer

LUMI

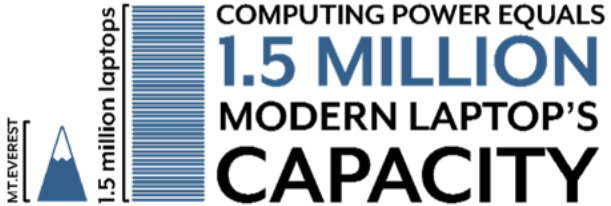
LUMI is 3rd fastest supercomputer in the world

LUMI

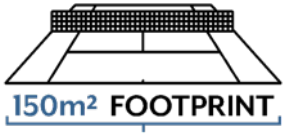
PEAK PERFORMANCE OVER

550 PETAFLOP/S

= performs 550 x 10^{15} calculations per second



2 x



High-performance computing

AI

Data analytics

LUMI Consortium

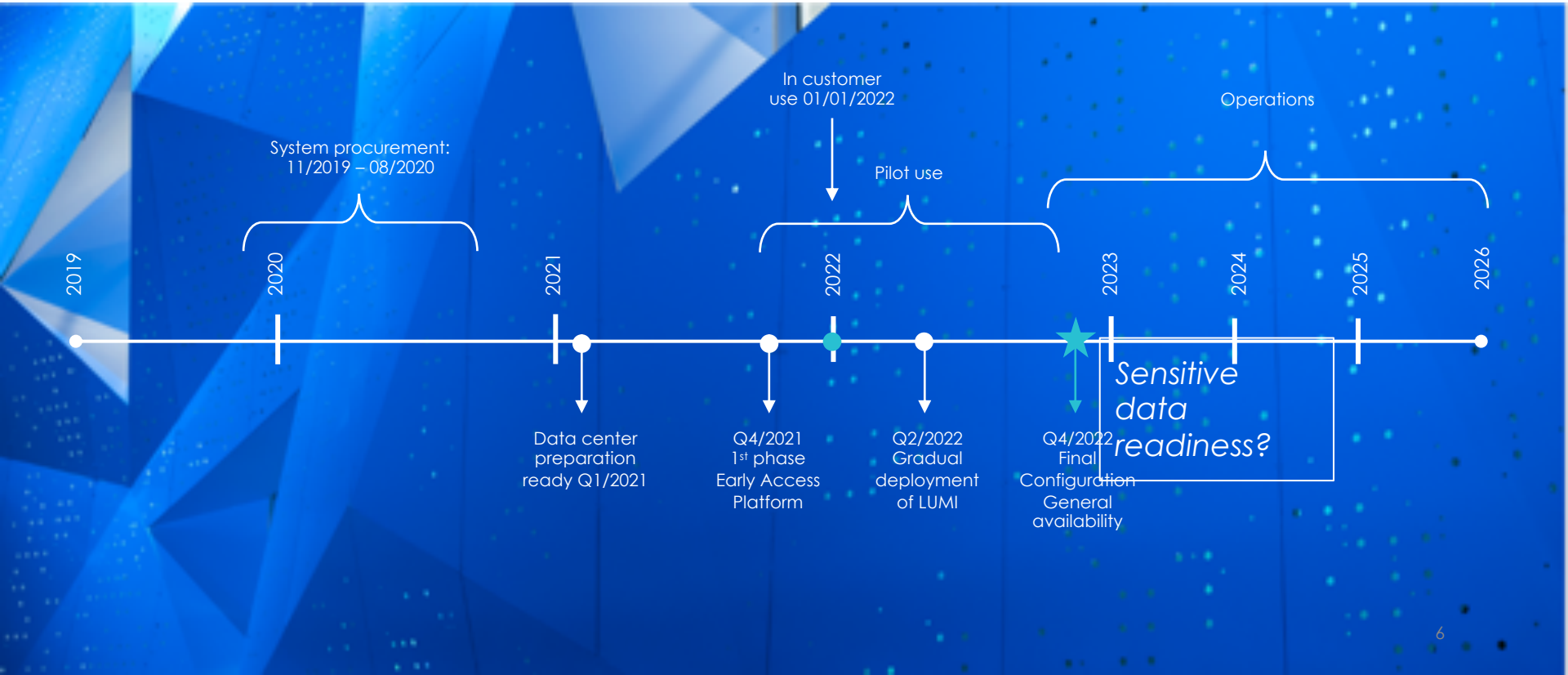
- LUMI consortium members are Finland, Belgium, Czech Republic, Denmark, Estonia, Iceland, Norway, Poland, Sweden and Switzerland
- The resources of LUMI will be allocated per the investments
- The 50% share of the EuroHPC Joint Undertaking (JU) will be allocated by a peer-review process and available for all European researchers
  - Also for industry and public sector

EuroHPC
Joint Undertaking

Puhuri – seamless access to LUMI and other resource providers

- Puhuri NeIC project has implemented a **distributed resource allocation and registration & authentication system**
- Users *register* to MyAccessID AAI service operated by GEANT eduTEAMS
  - Accepts Terms of Use of resource provider
  - Register public ssh key
    - The algorithm can be checked, but not if there is (good) password protection
  - Challenge: Identity proofing of all users
    - eIDAS proxy connected to Idp proxy covers currently only 13 EU countries
- *Authentication*: users can use their home organisation accounts
  - Identity provider proxy connects the identity providers and the services where users authenticate
    → Password policy is thus defined by their home organisation
- The Resource Allocation boards make the *authorisation* decisions by using their resource allocation portal (national or Puhuri provided)
- Research Project's manager (PI) invites the members to their group
- Outside of Puhuri's scope:
  - Data Protection Agreements (User as Controller of Personal Data processed in the service)
  - Export restrictions control of Service Providers

LUMI timeline



Sensitive data use case examples

Knowledge for treatment of diseases

Analyzing of human genomes
combined with health data
consented for secondary usage

Training of AI algorithms to diagnose pathological images, which may also contain pseudonymized metadata with diagnosis

TIME CRITICAL MODELLING

e.g. related to national or EU threat or other major crisis such as pandemics

Security in High Performance Computing (HPC) environment

Risks in supercomputing service

LUMI

- Confidentiality, integrity, availability of data
  - Personal Data → GDPR
- Unauthorised usage of resource and using to malicious activities
- Compromised system causes lots of work, interruption for service
- Denial of Service
- Reputation risks

Special security challenges in an HPC system

LUMI

- Large amount (thousands) of users in a shared environment (operating system, storage system, batch scheduling system)
- Complex software stack, various protocols for access and data transfer
- Users have operating system level access
  - Thus, risks for (unknown) vulnerabilities or misconfiguration exists
    - Elevation of privileges

- Users can also *accidentally* make misconfiguration e.g. on data sharing

Applying best security practices

LUMI

- Key elements of LUMI security framework are
  - Business continuity and disaster recovery plans
  - Security agreements and guidelines
  - Information and training on security
- ISO 27001 (security management) certification for LUMI hosting

- Physical Security based on zones, access controls and monitoring
- Network and system security is based on defence-in-depth, vulnerability management, adequate encryption and monitoring

- Procedures in place for change, capacity, and incident management
  - Communication with various stakeholders. Roles need to be defined.
- LUMI Security Interest Group supports the security management and networking LUMI countries

Sensitive data taskforce for LUMI

Sensitive data taskforce

LUMI

- Led by Claudio Pica (SDU,DK), ca. 10 active members from various LUMI counties (DK,FI, SE, CZ, PL) and vendor HPE.
- Goal is to find out technical solutions to enable sensitive data processing in LUMI by using the existing LUMI environment (e.g. SLURM)
- Various technical aspects focusing on data and workload isolation

  We are very interested to hear your ideas and best practices!

- Unfortunately there isn't commonly accepted standard(s) to define the requirements for sensitive data processing

What is the extra security for sensitive data?

LUMI

- Sensitive data processing job will be sent to a node dedicated for the job exclusively
- Data encryption must be user friendly and secure.
  - Data encryption must happen before sending it to LUMI.
  - Key management is needed to de/encrypt the data during the workflow.
- Secure work-storage (scratch) is needed.
  - LUMI has no local disks -> RAM (limited!) or Lustre based solution
  - Lustre's native encryption might be available in the future
  - Alternatively other encryption solution. Overhead and performance impact?
- Software environment security:
  - Users can bring their own singularity containers to isolated environment
  - Provided by support team and community developers (ensure the security and reliability of results)
- No Internet from the backend nodes? → License server challenges and no streaming of input data from Internet - except from trusted locations?
- Security assurance of jobs and nodes to ensure the configuration and trust?

Login and data access

LUMI

- Authentication: identify the user and to protect data from unauthorised users
- Services /clients support various technologies (ssh key, OIDC/SAML2, S3..)
  - Batch scheduling job

    → Secure credentials / encryption key management is essential

- Access to data repositories such as FEGA via data access middleware
  - To get data access credentials from Life Science AAI, user would need to authenticate with web browser and use their identity provider

GDPR Requirements for Resource Providers

GDPR collaboration between Controller and Processor

LUMI

- Data Controller is responsible on various GDPR requirements such as
  - allowing Data Subject to exercise their rights
  - to evaluate whether a system is suitable for processing and protecting the data (built-in security)
  - evaluate regularly effectiveness of the measures
- Processor needs to assist Controller to make documents
  - Data Protection Impact Analysis (DPIA) regarding risks for sensitive data.
  - Technical and Organisational Means (TOMS)
  - Privacy Notice

  … and also implement security:
- Art 28 "processor" e.g. 3(e) "taking into account the nature of the processing, assists the controller by appropriate technical and organisational measures, insofar as this is possible, for the fulfilment of the controller's obligation "
- Art 32 "security of processing "..the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, …

- Data Protection Agreement (DPA) is required between the Controller and Processor
  - e.g. instructions for processing, data breach notification, subprocessors..

LUMI

Thank you!

jarno.laitinen@csc.fi

Follow us

Twitter: @LUMIhpc

LinkedIn: LUMI supercomputer

YouTube: LUMI supercomputer

www.lumi-supercomputer.eu

contact@lumi-supercomputer.eu

EuroHPC Joint Undertaking

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States FI, BE, CH, CZ, DK, EE, IS, NO, PL, SE.

Leverage from the EU 2014–2020

European Union European Regional Development Fund

REGIONAL COUNCIL OF KAINUU

EURO