# Batch processing for the SAF network
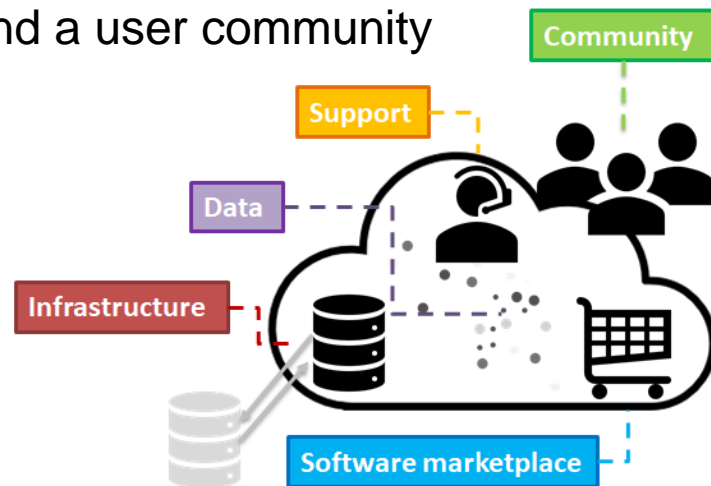
J.W. Kaiser and colleagues from the CM SAF, EUMETSAT & Accso

# THE VISION

SAFs will benefit from processing products in the EUMETSAT Cloud Infrastructure.
- access to input satellite data in Data Lake at EUM
  - efficient for large volumes, e.g. for climate data records
  - quick for real-time production
  - comprehensive
- common environment, set of ready tools, and a user community
  to accelerate collaborative developments
  - within SAFs
  - across SAFs

(Users will benefit from SAF products
being available in the EWC, too.)

**EUMETSAT**
**CM SAF**
CLIMATE MONITORING

**DWD**
**Deutscher Wetterdienst**
Wetter und Klima aus einer Hand

# Proof of concept: CLAAS TCDR production

- The processing of a 17-year climate data record based on all SEVIRI observations was accomplished on the ECI during July-October 2021.
- The infrastructure and processing chain proved to be stable and suitable for bulk processing.
- The CDR will become an operational product of CM SAF.
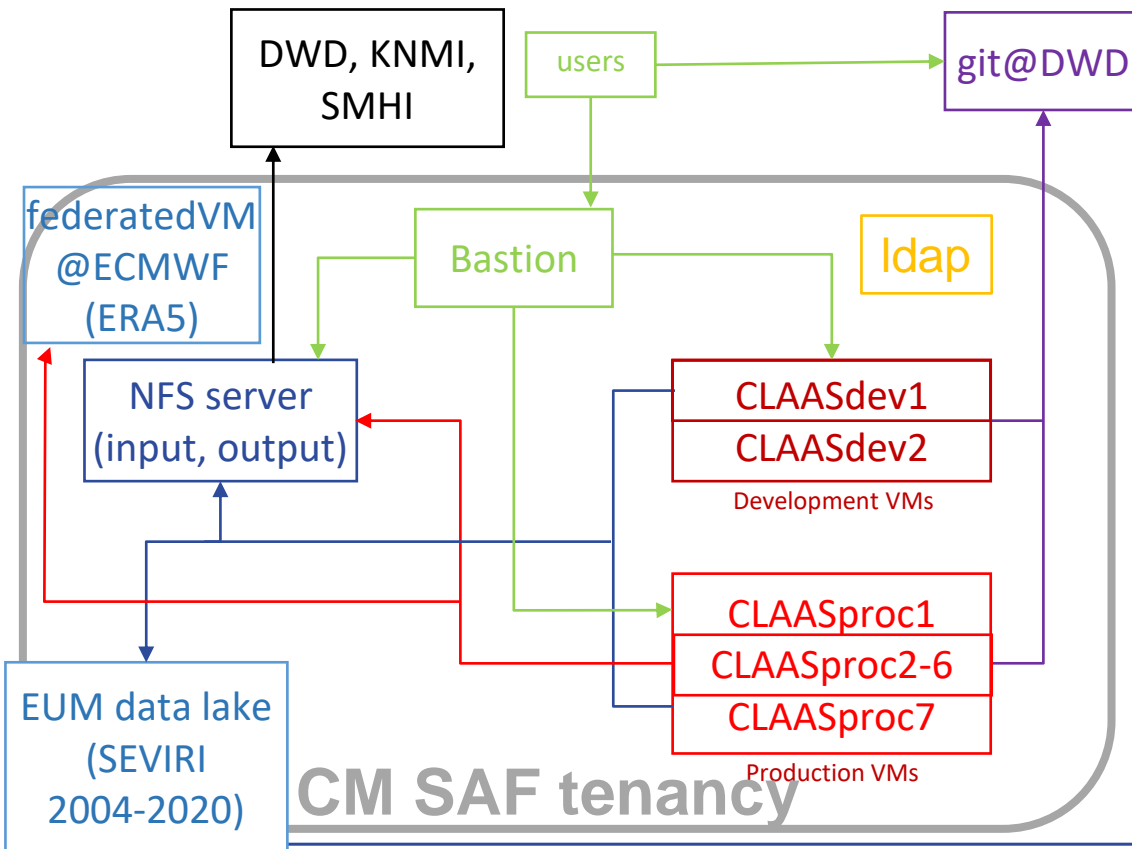
*mean ice water path [g/m2] during Feb 2004*
*(preliminary data, graphics by Irina Slodovnik, DWD)*
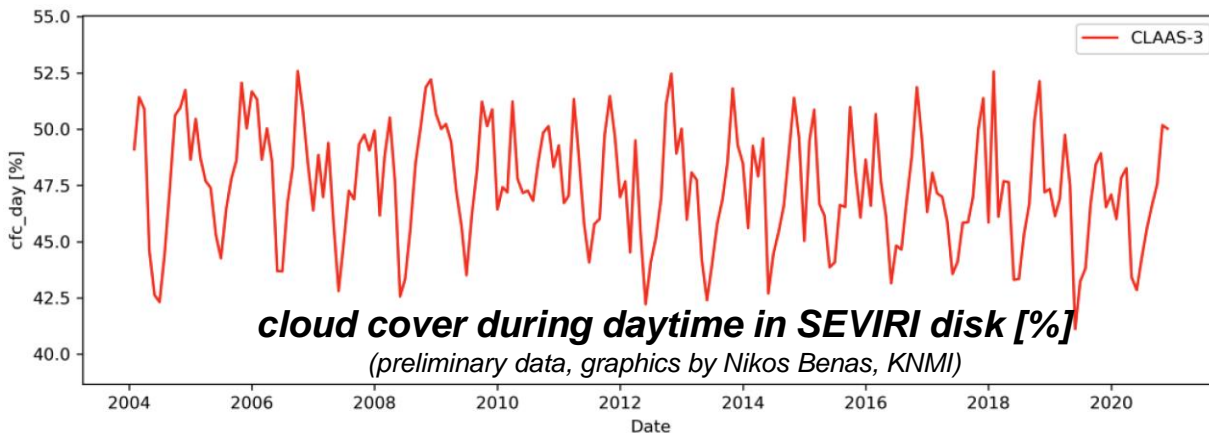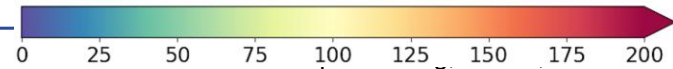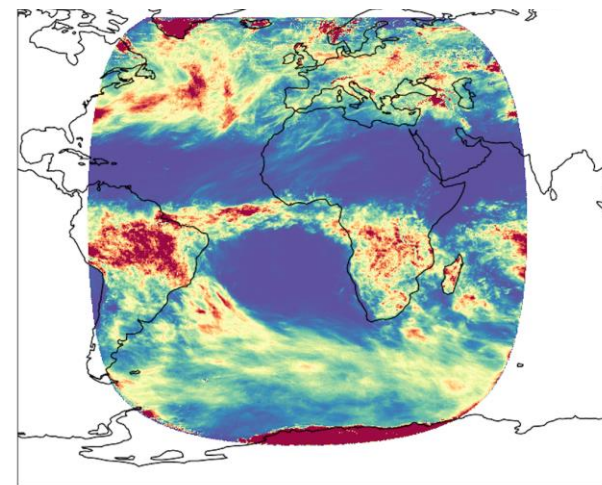




*cloud cover during daytime in SEVIRI disk [%]*
*(preliminary data, graphics by Nikos Benas, KNMI)*

# Setup for Development & Production of CLAAS TCDR



| VMs in CM SAF tenancy | Resources |
|---|---|
| Development | 8 cores, 16 GB RAM 100 GB disk |
| Production | 32 cores 128 GB RAM 150 GB disk |
| NFS server | 2 cores 4 GB RAM 8 TB disk |
| Bastion | 1 core 1 GB RAM 10 GB disk |

# Example CLAAS-3 Production VM Usage



**CPU**

58%

**RAM**

100%
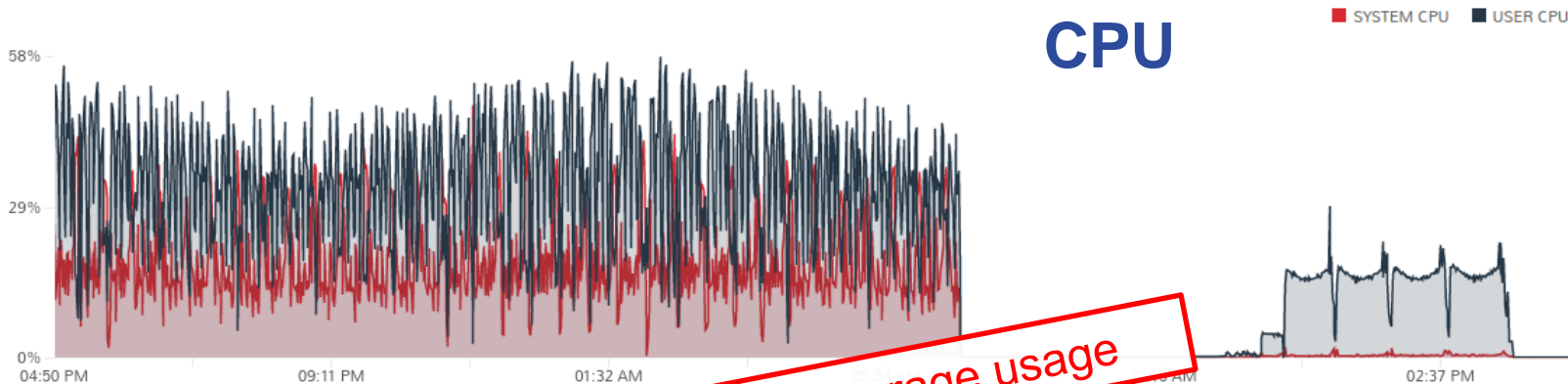
Need to maximise average usage

EUMETSAT
CM SAF
CLIMATE MONITORING

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

DWD

# THE CHALLENGE

- 2020 estimate of peak requirements for 12 product families by <u>CM SAF:</u>



- committed infrastructure provision for the entire SAF network (8 SAFs):
  - 1024 cores
  - 4 TB memory
  - 400 TB disk capacity
  - (object storage)
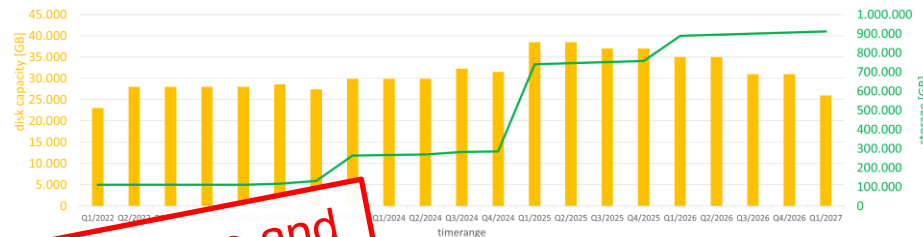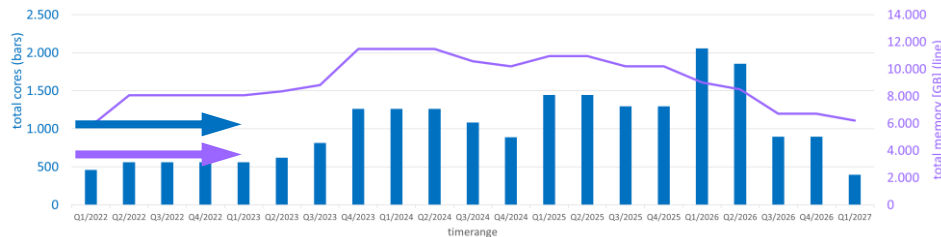
> Needs to (1) reduce peak usage and (2) maximise average usage require **flexible resource allocation** in SAFs and between SAFs.

# THE APPROACH

<u>technical idea:</u>
- use load balancing with a **batch system** for the entire SAF network

<u>administrative form:</u>
- WP in CM SAF and external contract run by DWD
  1. consolidation of user **requirements** from all SAFs
     - and technical requirements from EUM Sec
  2. **evaluation** of available solutions
  3. **implementation** of minimum viable solution
  4. (training of users and administrators)

# Tentative concept for batch in single tenancy
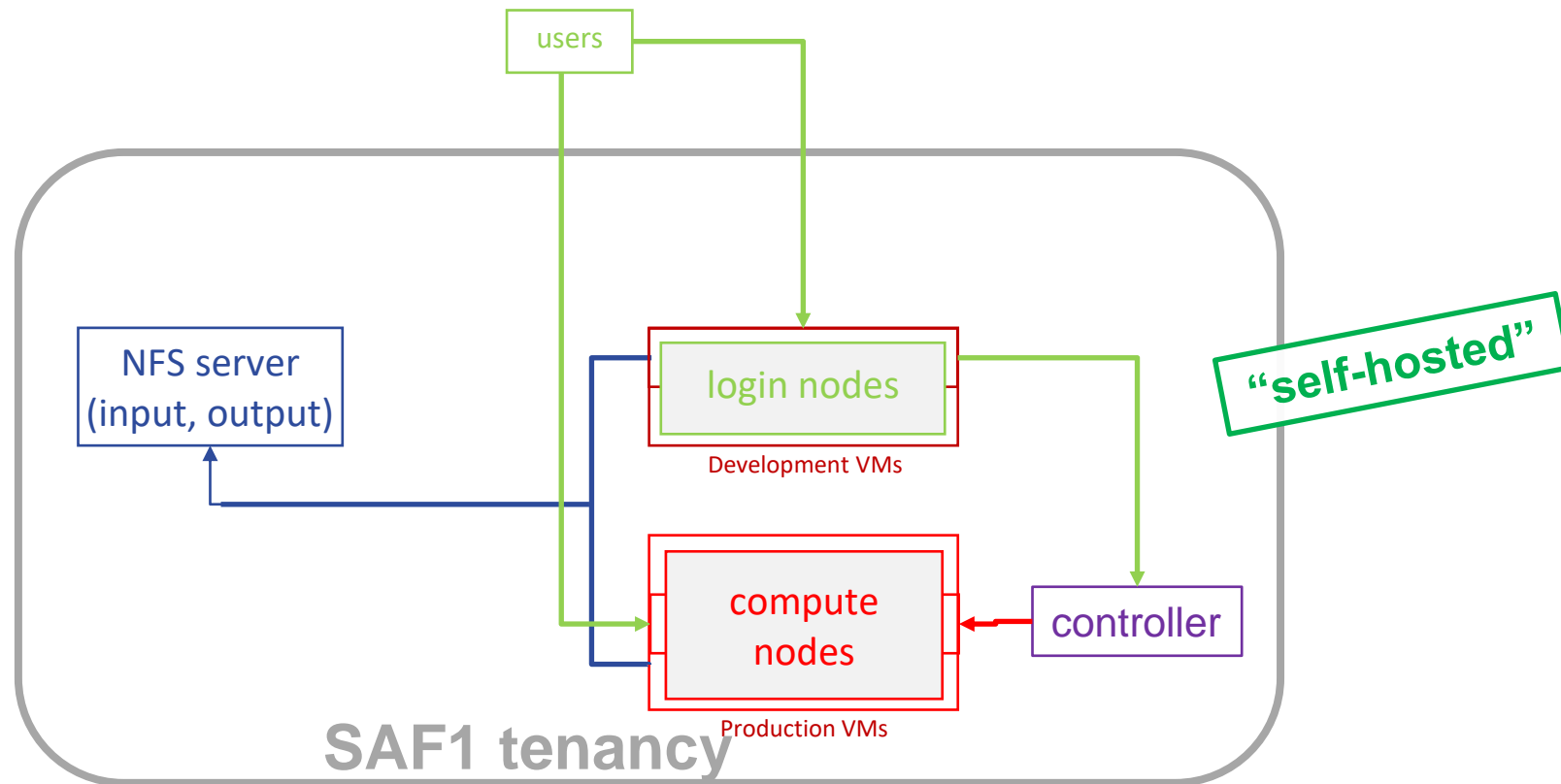
# Tentative concept for batch across SAF network

# 1. Requirement collection & prioritisation

- in 3 workshops with CM SAF, EUMETSAT Sec, SAF Network, respectively, on
  - 26 Oct, 10 Nov, after 15 Nov
- current status: 26 requirements

key constraints
- file exchange
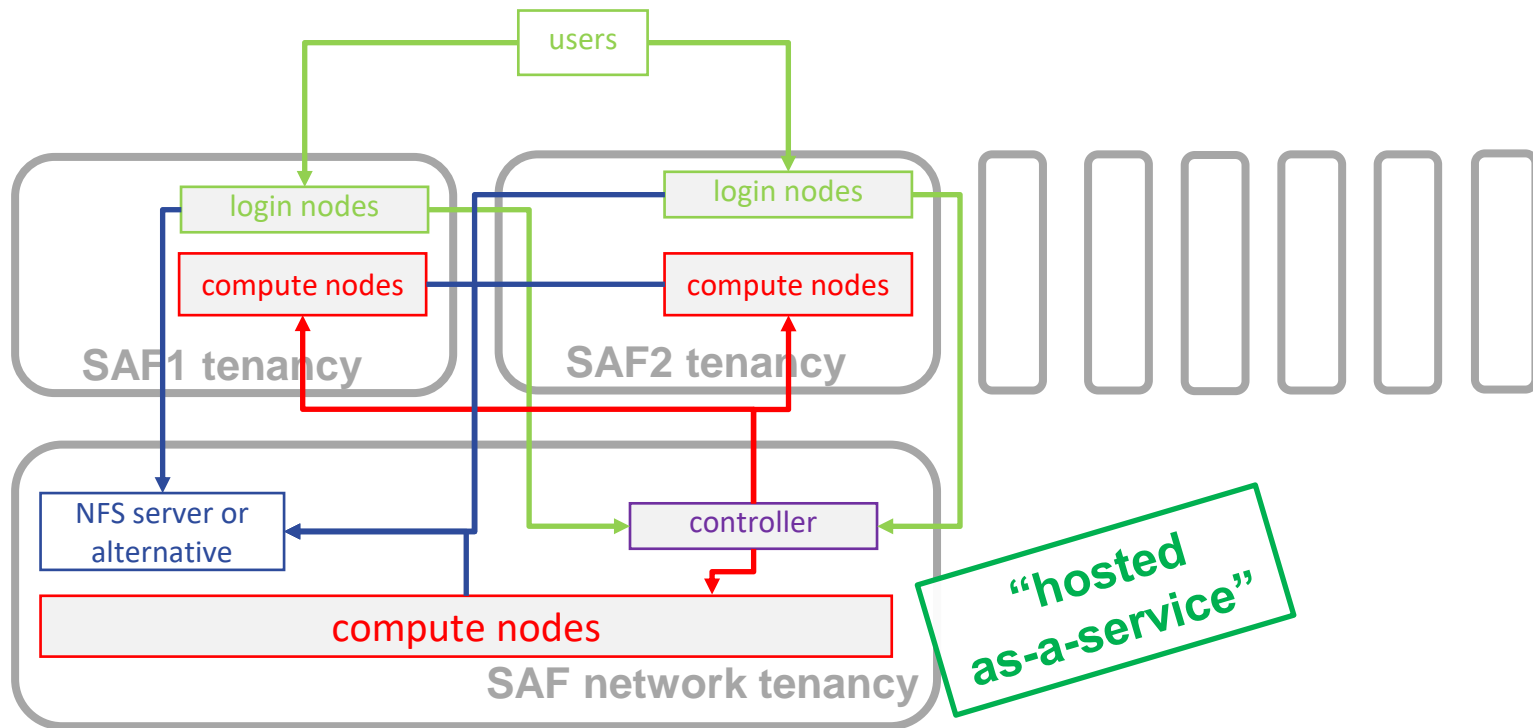- compute node
    standard system
- schedulers
    like ecFlow

| Requirement | Status | Explanation | Category | Subcategory |
|---|---|---|---|---|
| Jobs should be sent and displayed via the command line | (READY) | 26.10.2021<br>- How many users will be using the system? 50-100 users (only a guess from Mike); Johannes expects 25 users at a time. Differentiate between "generic" and "individual" users.<br>- What does "display a job" mean? By means: What is the current state of the job and who used what percentage of the systems over the last x months (statistics). | Batch-System | Execution |
| Jobs should be organized in queues. We need the following queues:<br>a. Time-critical production (access limited to user groups, high priority)<br>b. Non-time-critical production (access restricted to user groups, low priority)<br>c. Development (medium priority) | (DISCUSSION) | 26.10.2021<br>- What is meant by queue? HTCondor: accounting group | Batch-System | Scheduling |
| The number of jobs per user should be unlimited | (READY) | | Batch-System | Resource-Management |
| Load balancing should be done according to the following criteria:<br>a. Queue priority: High priority first<br>b. Resource requirement: Low requirement first | (DISCUSSION) | 26.10.2021<br>- Should be able to change the weighting of the different criteria on the fly by an operator<br>- Each user can define which of HIS jobs is the most important, the available priorities are defined by the accounting group the user is assigned to<br>- separated criteria c to #4a | Batch-System | Scheduling |
| Load balancing should be done according to the history: If you submit a lot of jobs, you sometimes have to wait | (READY) | | Batch-System | Scheduling |
| A compute node should be able to run multiple jobs simultaneously | (READY) | | Batch-System | Scheduling |
| The VM on which a job is executed is to be determined based on the job properties. In particular, jobs with high memory requirements are to be distributed across different VMs so that the remaining CPUs on these machines can be used by jobs with low memory requirements. | (READY) | | Batch-System | Scheduling |
| The system shall be adaptable when<br>a. VMs are added to the pool<br>b. VMs are removed from the pool<br>c. VMs are reconfigured (RAM & CPUs) | (READY) | 26.10.2021<br>- Resources are distributed between the 8 SAF projects + 1 central tenancy (operated by EUMETSAT)<br>- in case resources (by means CPU/memory used by a node) are idle, the projects can decide if they contribute these resources to the central tenancy<br>- if so the node will register itself (self-registering) at the central tenancy without the need to register the node manually<br>- boundary: dynamic resource allocation (the system identifies idle resources | Batch-System | Scalability |

EUMETSAT
CM SAF
CLIMATE MONITORING

DWD
Deutscher Wetterdienst
Wetter und Klima aus einer Hand

# 2. Evaluation of available solutions

- short list: HTCondor, OpenPBS, AirFlow

| | System | Priorities | Queues | Active Development | Open Source and free | Containerization | Documentation Quality | Remarks |
|---|---|---|---|---|---|---|---|---|
| Candidates for analysis | HTCondor | User, group, job | Accounting Groups | Yes | Yes | Yes | Good | |
| | Slurm | Only based on fair-share | Partitions | Yes | Yes | Not native | Good | |
| Options for third candidate | OpenPBS | job, queue | Yes | Yes | Yes | Yes (docu chapter 18) | Good (+ active forum) | Documentation: https://openpbs.atlassian.net/wiki/spaces/PBS Pro/pages/5537831/User+Documentation |
| | Apache Mesos | Quota, fair share and reservations; no individual job priorities | ? | Yes | Yes | Yes | Ok (no search) | Reservations and quota are nice features |
| | Apache Airflow | only for tasks, not e.g. user-wide | Yes | Yes | Yes | ? | Good | |
| | Luigi | Free configuration, but seems like it has to be coded manually | ? | Yes | Yes | Possible running a docker container as a task | Ok (not realy detailed, no big community) | While you can probably schedule a few thousand jobs, it's not meant to scale beyond tens of thousands. Luigi does not support distribution of execution. When you have workers running thousands of jobs daily, this starts to matter, because the worker nodes get overloaded. There are some ways to mitigate this (trigger from many nodes, use resources), but none of them are ideal. |

# 3. Implementation of selected batch system

- "Minimum viable system" expected by Jan 2022
  - according to requirement prioritisation
  - some scope for additional features

  - Software-as-infrastructure should ensure
    - future maintainability and adaptability
    - flexible deployment in two operating models
      - "self-hosted" by each tenancy, i.e. SAF
      - "hosted as-a-service"
        - pooled resources from all SAFs
        - usage-based accounting
        - central administration of batch system (by EUMETSAT Sec.)

# SUMMARY

- The EUMETSAT Cloud Infrastructure will facilitate efficient bulk processing in the SAF Network.

- Flexible resource allocation in SAFs and between SAFs will be needed to use the limited resources efficiently.

- CM SAF is working with Accso, EUMETSAT Sec and the SAF Network to establish a suitable batch system for this.

Outlook
- The same software stack will be made available for EWC users as well.
- Future operating model — all EWC users are considered.

# What didn't fit on the slides

One wish:

- Providing fast federated access to MARS for VMs at EUMETSAT should be high priority.

One acknowledgement:

- Many thanks to the most helpful staff at EUMETSAT and ECMWF supporting us!