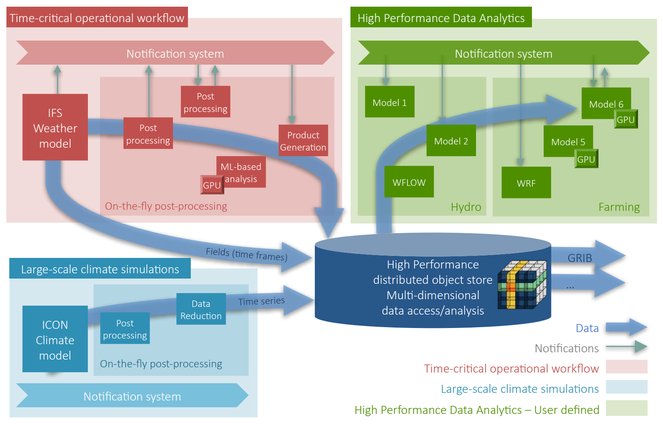
A framework for Better ORGanized Earth System model data — BORGES

Sven Willner*, Jairo Segura**, Reinhard Budich, Luis Kornblueh

Max-Planck-Institute for Meteorology, Hamburg

*sven.willner@mpimet.mpg.de, **jairo.segura@mpimet.mpg.de



Dannovaro, 2021

Abstract

BORGES aims to provide a high-performance semantic data management workflow for Earth System Models, which strive for ever higher resolutions in their simulations in space and in time — thereby producing a vast amount of data. BORGES is part in the *Weather, Climate, Hydrological and Farming* pilot project in the ACROSS project on *HPC Big DAta ArtifiCial intelligence cross Stack PlatfoRm TOwardS ExaScale*. It is being developed at the Max Planck Institute for Meteorology, Hamburg, focusing on large-scale climate simulations with the Earth System Model ICON.

Project overview

- Project runtime: 10/2021 to 02/2024
- State: Architecture defined after first simple prototypes
- Next steps: First minimal viable product version to be tested and benchmarked with climate researchers as main targeted users
- In: EuroHPC-project ACROSS (`https://acrossproject.eu`)

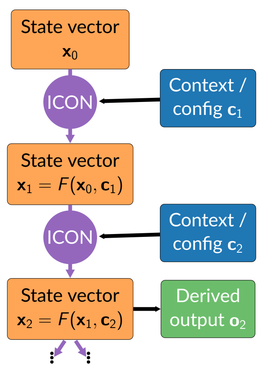
1 Main ideas

As a data system BORGES combines two worlds: the service architecture of a classical database service and the decentralized, highly dynamic multi-user environment of HPC. In it, high-volume model output as well as corresponding metadata is to be stored in a consistent way to be easily and efficiently retrievable. To ensure data consistency and optimal data organization the system is kept strictly separated from the actual model code but uses the fastest data pathways possible nevertheless.

For that, the system is a collection of distributed processes and only accessible via well-defined interfaces. In particular, for each high-volume writing and reading process, the system provides an “interface service”. This is a process under control of the BORGES administrators, running as a dedicated BORGES user (effective user id), and as part of the user’s HPC job as an “ephemeral service”.

2 ICON

The ICON (ICOsahedral Nonhydrostatic) model is a state-of-the-art Earth System model used for climate research as well as for global numerical weather prediction. Currently, many researchers at MPI-M and other research institutions employ ICON covering a variety of research questions, from slow low-resolution long-timescale processes to quick high-resolution short timescale ones. Developed in Fortran and C, the model is highly parallelized with dedicated MPI ranks used for I/O.



3 Storage system

From the perspective of the user, data should be retrievable in an easy, semantic way: instead of organizing filenames themselves, respective attributes of the data are sufficient to search and retrieve it. Thus, at the heart of the BORGES data system, the Field Data Base (FDB) data catalog library is used as a domain-specific object store for core data.

Longer-term storage data can seamlessly be transferred between a hierarchy of storage backends with several levels of volumes and access times. In the ACROSS project, FDB is further advanced to take advantage of more modern non-volatile storage devices as well.

4 Metadata

In an abstract way, a physics model such as ICON can be perceived as a process transforming a single state vector — a representation of the full state of the modeled system — into another. To ensure reproducibility and for better data provenance the configuration and context of the model run needs to be stored along side the generated data — be it a full state vector or some data derived from that.

Accordingly, BORGES keeps track of experiment metadata as well as parameter and grid definitions according to WMO and intra-institute standards. This is done in a central server with a defined interface to users as well as to other parts of the system. The semi-centralized nature of the data system thereby allows to keep data consistency at check.

5 Post-Processing

Dedicated interface service processes are also used for reading and other post-processing tasks as part of the respective user HPC job. This allows high throughput in the data transfer and can also include simple post-processing tasks as part of the interface service, whose resource use is accounted for directly by the respective workload manager (e.g. Slurm).

During actual model runtime the interface service acts as a data “multicaster” streaming the model output data to storages and several tasks for “in-situ” post-processing. Additionally, interface services can be shared between different post-processing tasks across users. That way, High-Performance Data Analysis can be done much more dynamically than with classical workload managers. This is, in particular, useful for the planned Digital Twins of the Earth system in the DestinE project. Here, post-processing tasks (or whole trees of these) “observe” the looping model and attach/detach in a dynamical way.

The coordination of different interface services scattered across users, jobs, and nodes in the HPC system is thereby centrally coordinated by a dedicated server.

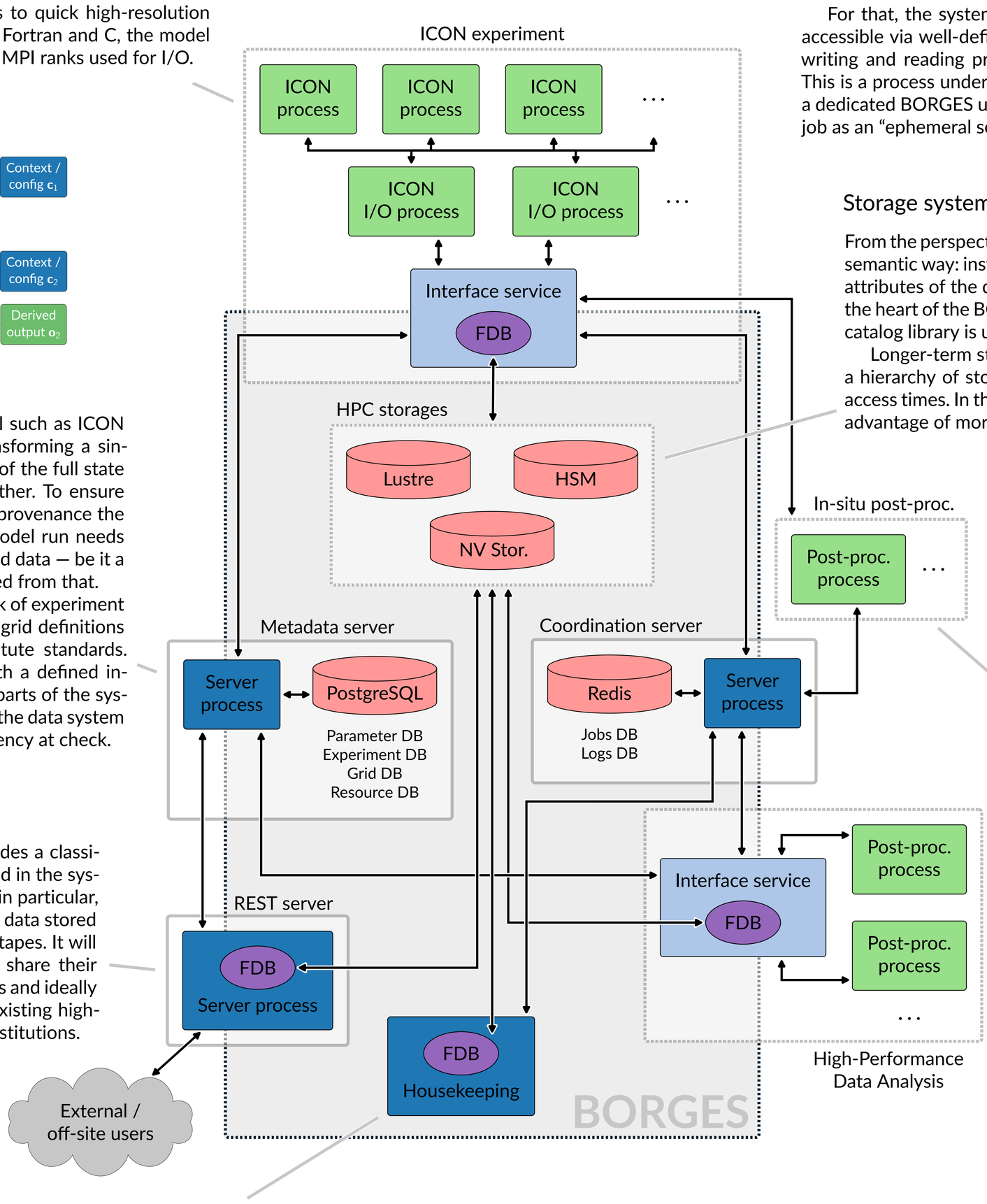
6 REST interface

For external users the system provides a classical REST server to access data stored in the system. Thereby, the service needs to, in particular, take care of asynchronous access to data stored in slower storage backends such as tapes. It will further allow users to publish and share their data in BORGES with other scientists and ideally in a later stage take advantage of existing high-speed connections between HPC institutions.

7 Housekeeping

The semi-centralized structure and strict separation between data storage and using processes allows for improved control of the data inside the system. In particular, a housekeeping process will be run regularly to ensure that

- stored data is consistent and residuals from failed concurrent writes are cleaned up,
- data with limited time-to-live is removed (a default TTL is planned to ensure that users only keep the data they really need longer-term),
- data is seamlessly transferred between storage backends (e.g. between tape archives and cache disks),
- data layout is adjusted to ensure high performance according to actual usage patterns.



Legend

ACROSS HPC Big DAta ArtifiCial intelligence cross Stack PlatfoRm TOwardS ExaScale

EuroHPC Joint Undertaking

SPONSORED BY THE Federal Ministry of Education and Research

MAX-PLANCK-INSTITUT FÜR METEOROLOGIE