EGU2020-22618
https://doi.org/10.5194/egusphere-egu2020-22618
EGU General Assembly 2020


# Towards easily accessible interactive big-data analysis on supercomputers

**Katharina Höflich**[1], Martin Claus[1], Willi Rath[1], Dorian Krause[2], Benedikt von St. Vieth[2], and Kay Thust[2]
[1]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany (khoeflich@geomar.de; mclaus@geomar.de; wrath@geomar.de)
[2]Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany (d.krause@fz-juelich.de; b.von.st.vieth@fz-juelich.de; k.thust@fz-juelich.de)

Demand on high-end high performance computer (HPC) systems by the Earth system science community today encompasses not only the handling of complex simulations but also machine and deep learning as well as interactive data analysis workloads on large volumes of data. This poster addresses the infrastructure needs of large-scale interactive data analysis workloads on supercomputers. It lays out how to enable optimizations of existing infrastructure with respect to accessibility, usability and interactivity and aims at informing decision making about future systems. To enhance accessibility, options for distributed access, e.g. through JupyterHub, will be evaluated. To increase usability, the unification of working environments via the operation and the joint maintenance of containers will be explored. Containers serve as a portable base software setting for data analysis application stacks and allow for long-term usability of individual working environments and repeatability of scientific analysis. Aiming for interactive big-data analysis on HPC will also help the scientific community in utilizing increasingly heterogeneous supercomputers, since the modular data-analysis stack already contains solutions for seamless use of various architectures such as accelerators. However, to enable day-to-day interactive work on supercomputers, the inter-operation of workloads with quick turn-around times and highly variable resource demands needs to be understood and evaluated. To this end, scheduling policies on selected HPC systems are reviewed with respect to existing technical solutions such as job preemption, utilizing the resiliency features of parallel computing toolkits like Dask. Presented are preliminary results focussing on the aspects of usability and interactive use of HPC systems on the basis of typical use cases from the ocean science community.