A B C / J geoverbund
aachen bonn cologne jülich

HPSC TerrSys

JÜLICH Forschungszentrum

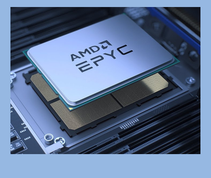
Towards Exascale through Modularity And I/O Management

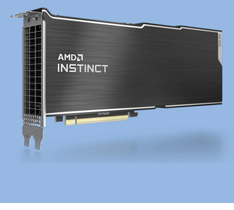
Performance study under Modular computing with TSMP

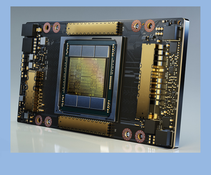
In the journey to exascale Earth System Modeling, adapting software to evolving hardware is a key challenge. We focuse on performance-portability solutions to ensure software compatibility with new capabilities.

Software development challenges

Performance
Scalability
Portability
Readability
Sustainability
Productivity



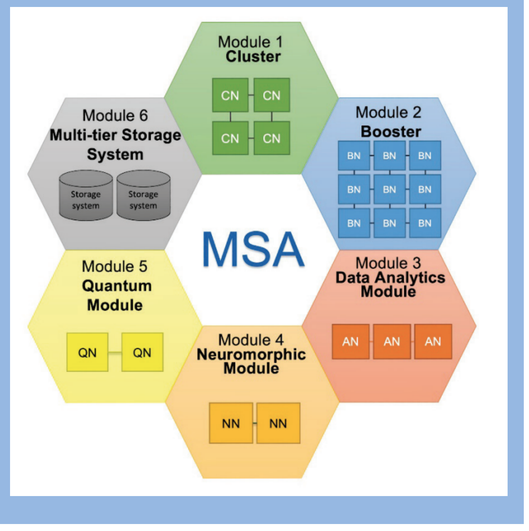




Adapting to evolving hardware poses a challenge for HPC architects and administrators, who must meet diverse needs. JSC employs a Modular System Architecture (MSA) to tackle this. This approach, especially beneficial for ESM software, strategically aligns modular codes with MSA modules. This smart mapping is a key step toward exascale efficiency.

ESM workflows stand to gain by simultaneously utilizing CPU clusters and accelerated booster modules for diverse model components. Moreover, employing data analytics allows for in situ processing, executing ML algorithms, and even implementing model surrogates, enhancing the workflow's capabilities.

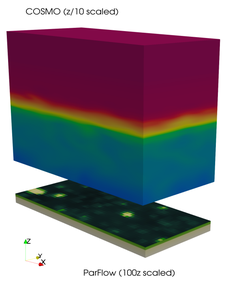
we explore tthe use of heterogeneous and modular computing with the Terrestrial Systems Modelling Platform.

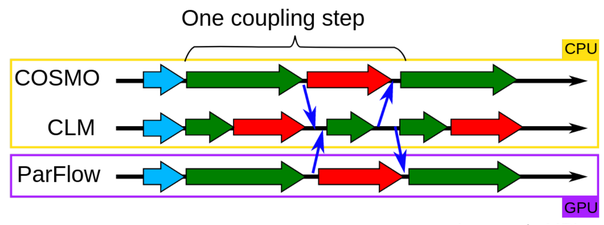


Scalability study of an idealised planetary boundary : concurrent coupled models on different hardware

The Terrestrial Systems Modeling Platform (TSMP) is a Multiple Program Multiple Data (MPMD) system model that seamlessly integrates into a Modular Supercomputing Architecture (MSA). Within the DEEP - SEA project, our exploration revolves around understanding the performance, scalability, and load-balancing characteristics of TSMP in response to heterogeneous and modular configurations, as opposed to homogeneous ones.

Both TSMP and the JUBE workflow presently possess the capability to submit jobs of both homogeneous (CPU only) and heterogeneous nature. In the latter scenario, ParFlow is executed on the GPU, while COSMO and CLM run on the CPU.



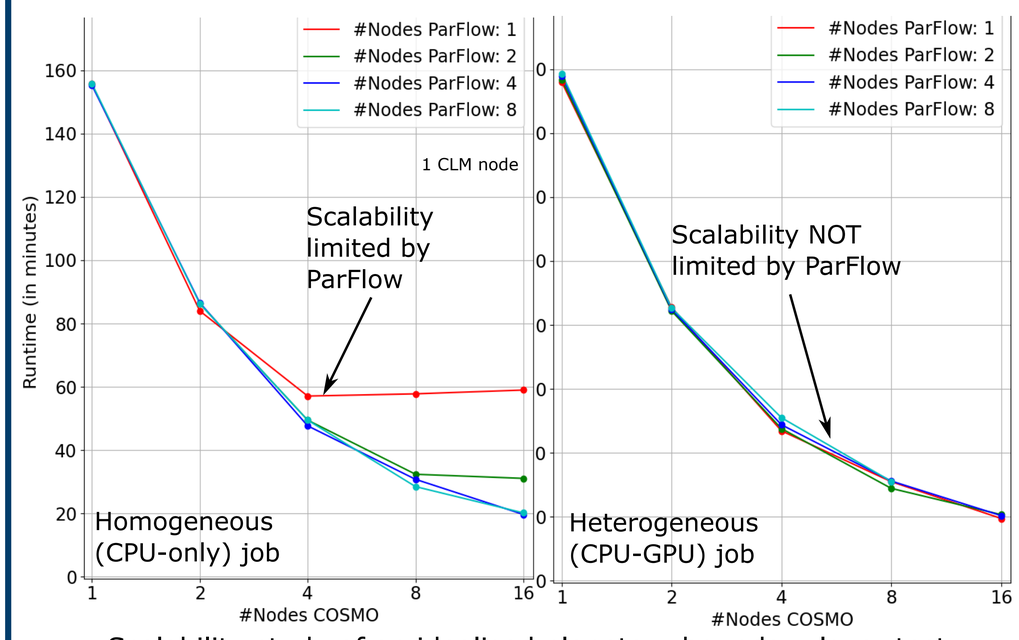


DEEP-SEA

In our investigation, we thoroughly examined both strong and weak scaling, employing the JUBE software for meticulous evaluation across diverse domain reconfigurations on the JUWELS and DEEP supercomputers. Our primary objective was to assess parallel efficiency. Additionally, we conducted a detailed analysis of traces from specific cases to understand behavioral variations across distinct allocation configurations.

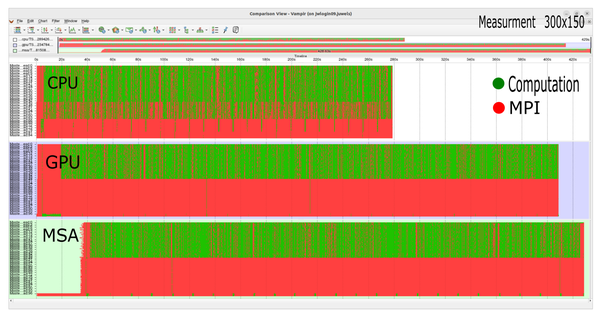
Our observations revealed that the coupling of TSMP with Score -P demonstrated significant speedup in CPU runs compared to GPU/MSA runs, particularly noticeable in the "Richards Exclude 1st" scenario.

Furthermore, our study highlighted that TSMP's component models exhibit distinct computational loads, potentially leading to inefficient use of HPC resources. We proposed that heterogeneous runs could mitigate these inefficiencies, enhancing overall resource utilization.

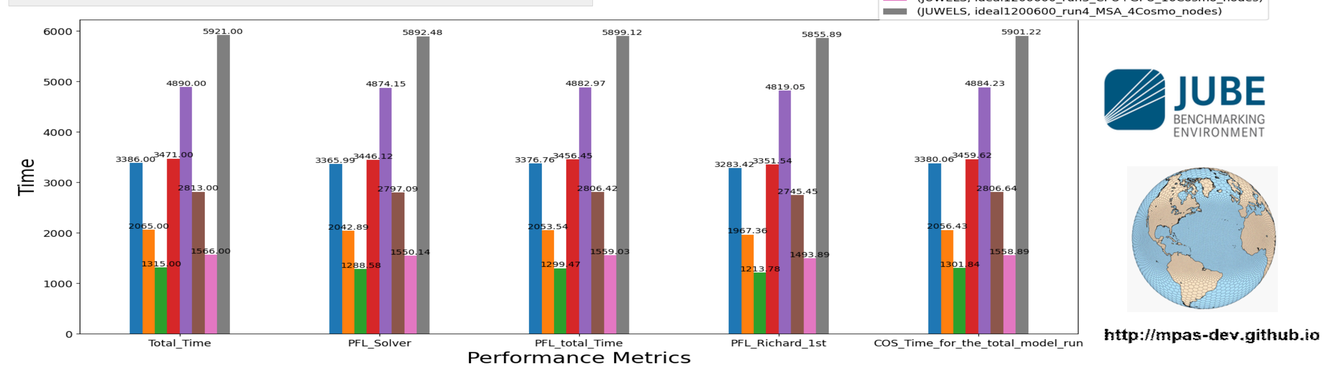


Scalability study of an idealised planetary boundary layer test

https://www.deep-projects.eu/



Score-P
Scalable performance measurement infrastructure for parallel codes



JUBE BENCHMARKING ENVIRONMENT

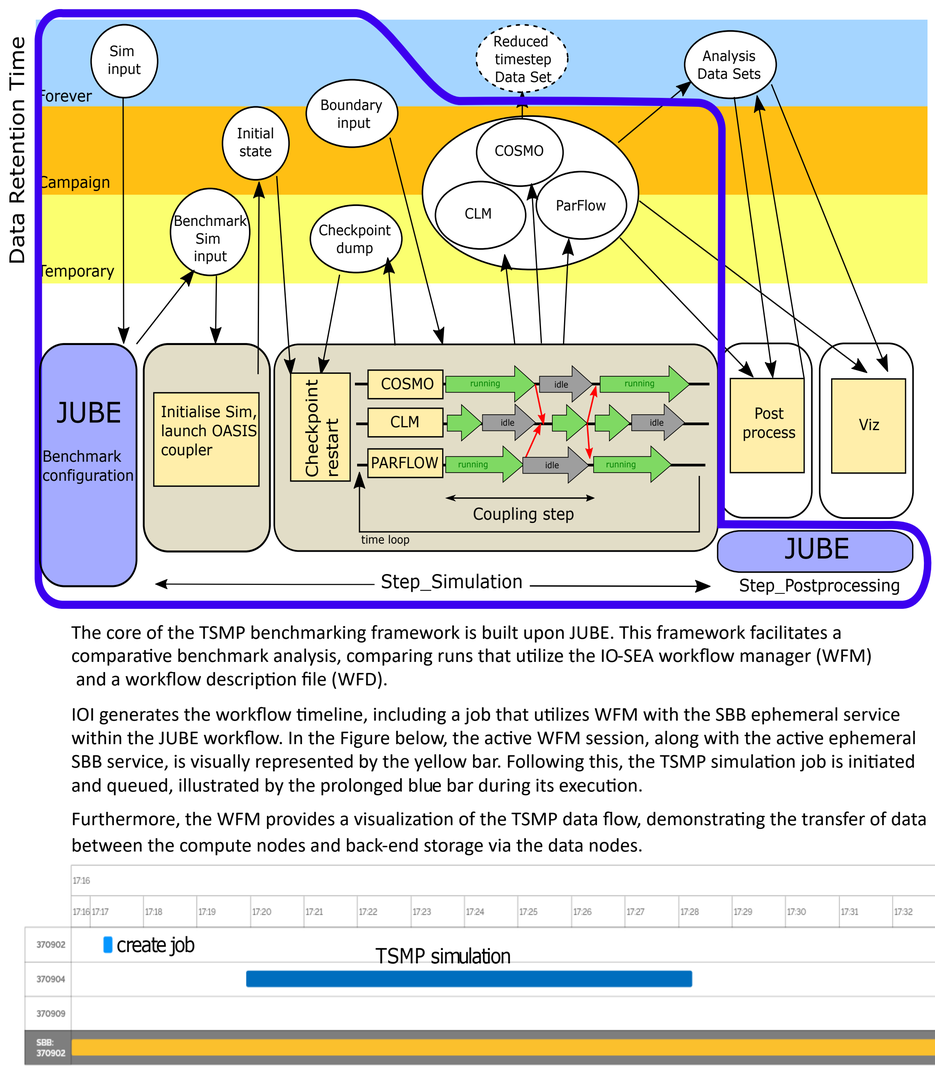
http://mpas-dev.github.io

IO – Software for TSMP Exascale benchmarking study:SBB ephemeral service

The primary objective of IO-SEA is to leverage the Smart Burst Buffer (SBB) service to optimize data movement, particularly beneficial for large-scale workflows on a busy HPC platform. IOI Instrumentation captures I/O -related metrics, organizing them into counters for each five-second timeslice during the workflow runtime. Some of these counters are structured as histograms.

During the IOI job overview, a comparison was made between read times and operation numbers, both with and without SBB. The analysis aimed to highlight the shift in the distribution of read operation durations. Conversely, a detailed examination of writing operations, with and without SBB, uncovered distinct patterns. Notably, there were significant write events at the beginning, corresponding to the initial output of static fields and initial states. Subsequently, six well -defined writing events aligned wi th state snapshots. In jobs employing SBB, the temporal distribution of writing events remained consistent but featured notably shorter writing operations.

https://iosea-project.eu7



IO-SEA

The core of the TSMP benchmarking framework is built upon JUBE. This framework facilitates a comparative benchmark analysis, comparing runs that utilize the IO-SEA workflow manager (WFM) and a workflow description file (WFD).

IOI generates the workflow timeline, including a job that utilizes WFM with the SBB ephemeral service within the JUBE workflow. In the Figure below, the active WFM session, along with the active ephemeral SBB service, is visually represented by the yellow bar. Following this, the TSMP simulation job is initiated and queued, illustrated by the prolonged blue bar during its execution.

Furthermore, the WFM provides a visualization of the TSMP data flow, demonstrating the transfer of data between the compute nodes and back-end storage via the data nodes.

Contact: g.tashakor@fz-juelich.de | website: http://www.hpsc-terrsys.de/