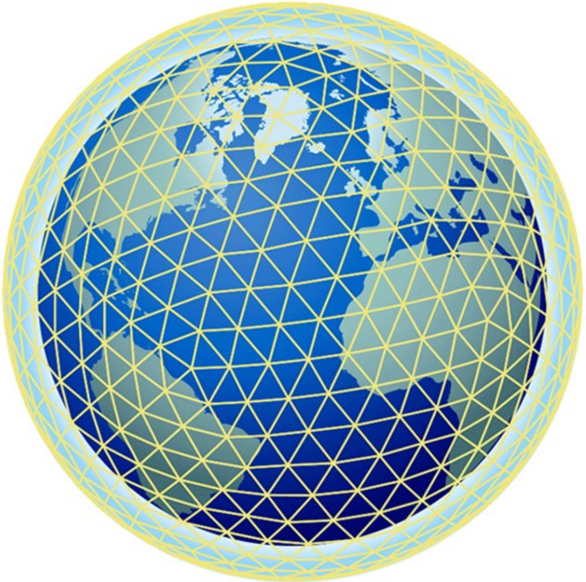


EARTH SYSTEM MODELING ON MODULAR SUPERCOMPUTING ARCHITECTURES: COUPLED ATMOSPHERE-OCEAN SIMULATIONS WITH ICON

OLAF STEIN, FZJ/JSC, SDL CLIMATE SCIENCE | OCTOBER 10, 2023

A. BISHNOI, C.I. MEYER, R. REDLER, N. EICKER, H. HAAK, L. HOFFMANN, D. KLOCKE, L. KORNBLÜH, E. SUAREZ

WARM WORLD

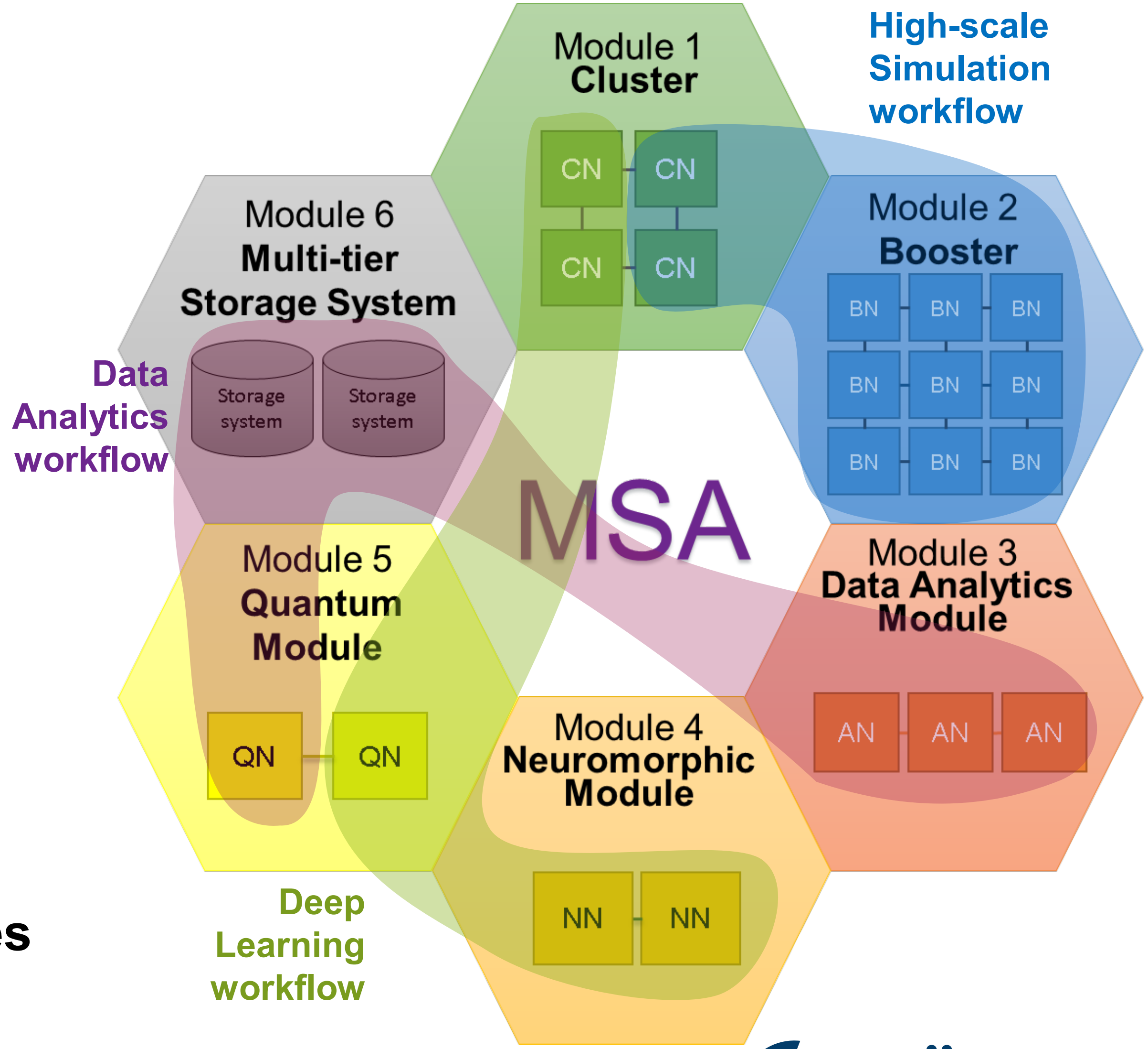
ICON

MAX-PLANCK-INSTITUT FÜR METEOROLOGIE

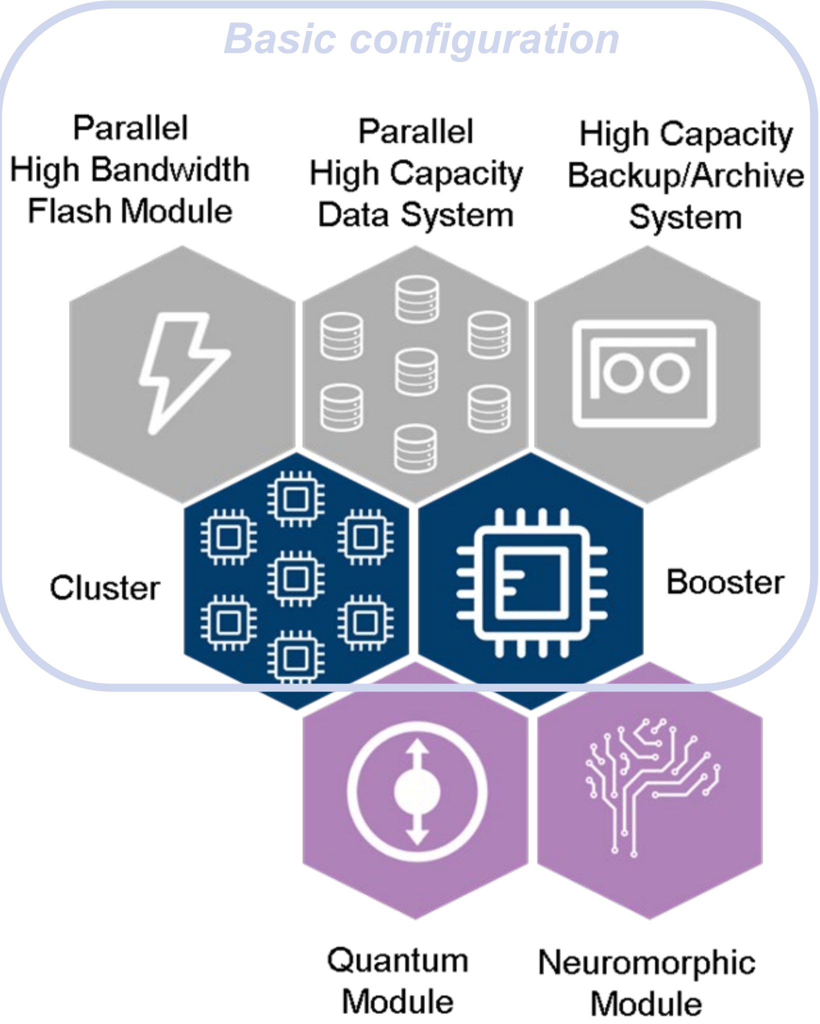
JÜLICH Forschungszentrum

MODULAR SUPERCOMPUTING ARCHITECTURE - MSA

- **Composable heterogeneous resources to serve diverse applications**
- **Segregate hardware resources into seperate groups such that the components within each group are maximally homogeneous**
- **Each application can select any number of nodes on the different partitions**
- **Segregation into compute partitions makes it much easier to dynamically allocate resources**



JÜLICH SUPERCOMPUTING CENTRE – ROADMAP TO JUPITER



- **JSC will host Europe‘s first exascale modular supercomputer by end of 2024 – JUPITER**
- **The system will be acquired by the European supercomputing initiative EuroHPC JU**
- **The current HPC system JUWELS (Cluster & Booster) is a blueprint for JUPITER**

CHANCES AND CHALLENGES FOR ESM

- **Modular applications like Earth System Models naturally fit MSA**
- **Computational requirements for ESM model components do vary (e.g.: Atmosphere - Ocean - Land - Chemistry - I/O - Interfaces)**
- **Long development cycles prevent model components to be ported to GPUs**
- **Some components are better suited to run on classical CPUs due to their communication and memory usage patterns**
- **Need to find load balance between components and HPC partitions they are running on**

Bishnoi et al. *Earth system modeling on Modular Supercomputing Architectures: coupled atmosphere-ocean simulations with ICON 2.6.6-rc*, EGUsphere [preprint], doi.org/10.5194/egusphere-2023-1476, 2023.

ICON - ICOSAHEDRAL **N**ONHYDROSTATIC

- **Weather and Climate model jointly developed by MPI-Met, DWD, KIT, DKRZ, ETH**
- **Growing user community all over Europe**
- **Flexible configuration from global to regional domains, with regional refinement, or zoom options**
- **used for operational weather forecasts at DWD**
- **Extension to an Earth System Model with compartments:**
  - **ICON-A: Atmosphere**
  - **ICON-O: Ocean (including sea ice)**
  - **ICON-Land: Integral part of ICON-A**
  - **Option to include the full carbon cycle & bio-geochemical processes in the ocean**

JÜLICH Forschungszentrum

WARMWOLD

Goals of the BMBF funded project (2022-2025)

An ICON based coupled model configuration capable of being run, with an acceptable simulation quality, on km-scale* with a throughput of >0.5 SYPD

A first release of ICON-consolidated: an open source version of the fully coupled (land, ocean, atmosphere) ICON system refactored to enable its scalable development

An integrated workflow that exposes the information content of the ICON system alongside IFS-based solutions and observational data in ways that leverage the intuition of users and supports innovation

**2.5km global grids or finer*

ICON SET-UP FOR MSA

ICONv2.6.6-rc Sapphire model configuration

Hohenegger et al. *ICON-Sapphire: simulating the components of the Earth System and their interactions at kilometer and subkilometer scales*, doi.org/10.5194/gmd-2022-171, 2023.

- **Global simulations R2B09 → 5 km resolution**

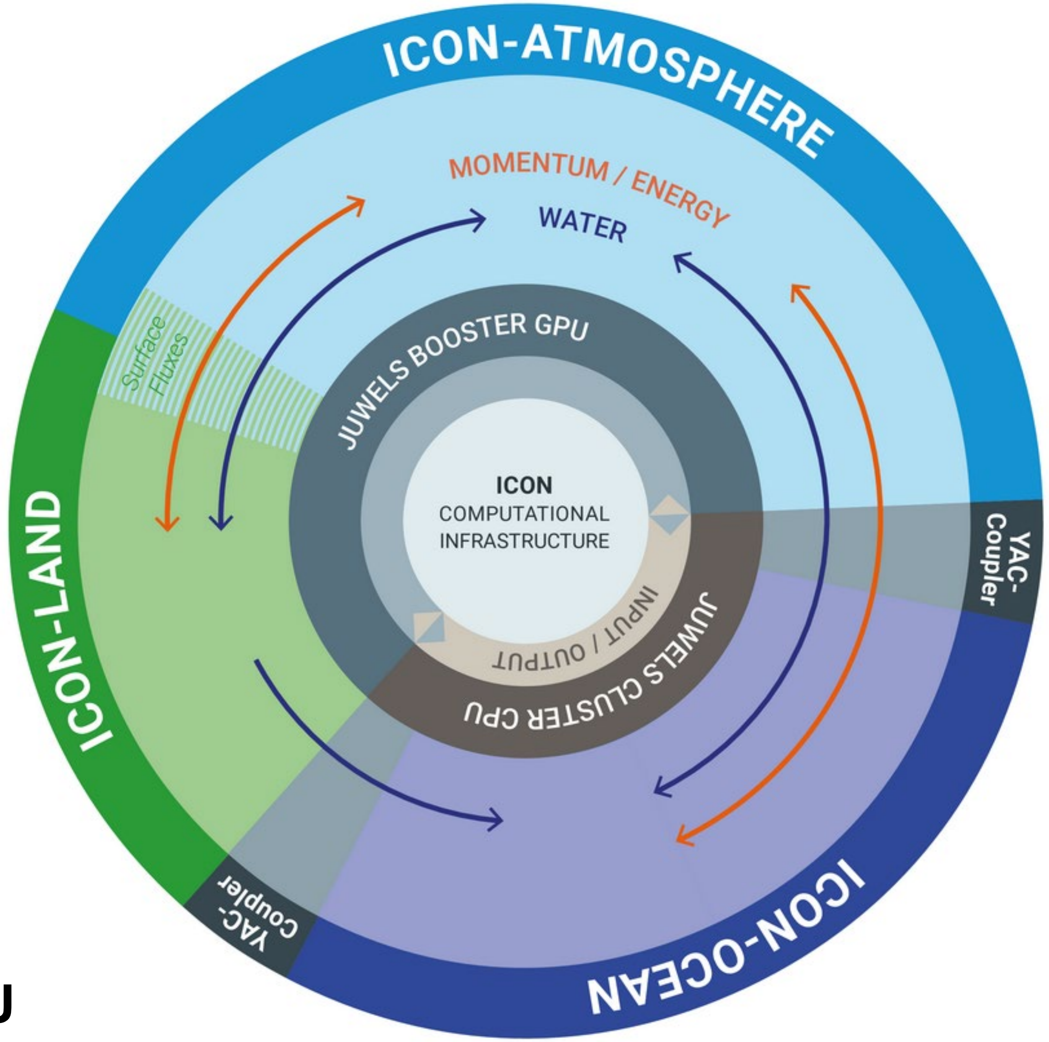
  ~ 21 Mio. amospheric grid cells

  ~ 15 Mio. oceanic grid cells

- **ICON-A running on JUWELS Booster GPUs**
  - **Each node with 4× NVIDIA A100 GPU**

Giorgetta et al. *The ICON-A model for direct QBO simulations on GPUs (version icon-cscs:baf28a514)*, doi.org/10.5194/gmd-15-6985-2022, 2022.

- **ICON-O running on JUWELS Cluster CPUs**
  - **Each node with 2× Intel Xeon Platinum 8168 CPU**
- **I/O running on JUWELS Cluster CPUs (17 dedicated nodes)**
- **Components coupled with YAC coupling library**



EXTRACT OF SLURM SCRIPT (PART 1)

```
#! /bin/bash
#SBATCH --job-name=hybCpl_R2B4          # Job name
#SBATCH --time=00:15:00                 # Time limit
# ------------------------------------------------------------------------------------------
#SBATCH --account=highresmonsoon        # User account
#SBATCH --partition=develbooster        # Choose Booster partition
#SBATCH --constraint=gpu                # Select GPU
#SBATCH --gres=gpu:1                    # Request 1 GPU
#SBATCH --cpus-per-gpu=1                # Request 1 CPU
#SBATCH --nodes=1                       # Request 1 node
#SBATCH --ntasks-per-node=1             # Execute 1 task per node
#SBATCH hetjob          # Separator between parts of heterogeneous Job
# ------------------------------------------------------------------------------------------
#SBATCH --account=esmtst                # User account
#SBATCH --partition=batch               # Select Cluster partition
#SBATCH --nodes=1                       # Request 1 node
#SBATCH --ntasks-per-node=8             # Run 8 tasks per node
#SBATCH --distribution=block:block      # Thread binding (distrib. over cores)
# --- The script continues with the configuration of the use-case itself
```

global settings

Atmosphere computation on JUWELS Booster (het_group_0)

Ocean computation on JUWELS Cluster (het_group_1)

EXTRACT OF SLURM SCRIPT (PART 2)

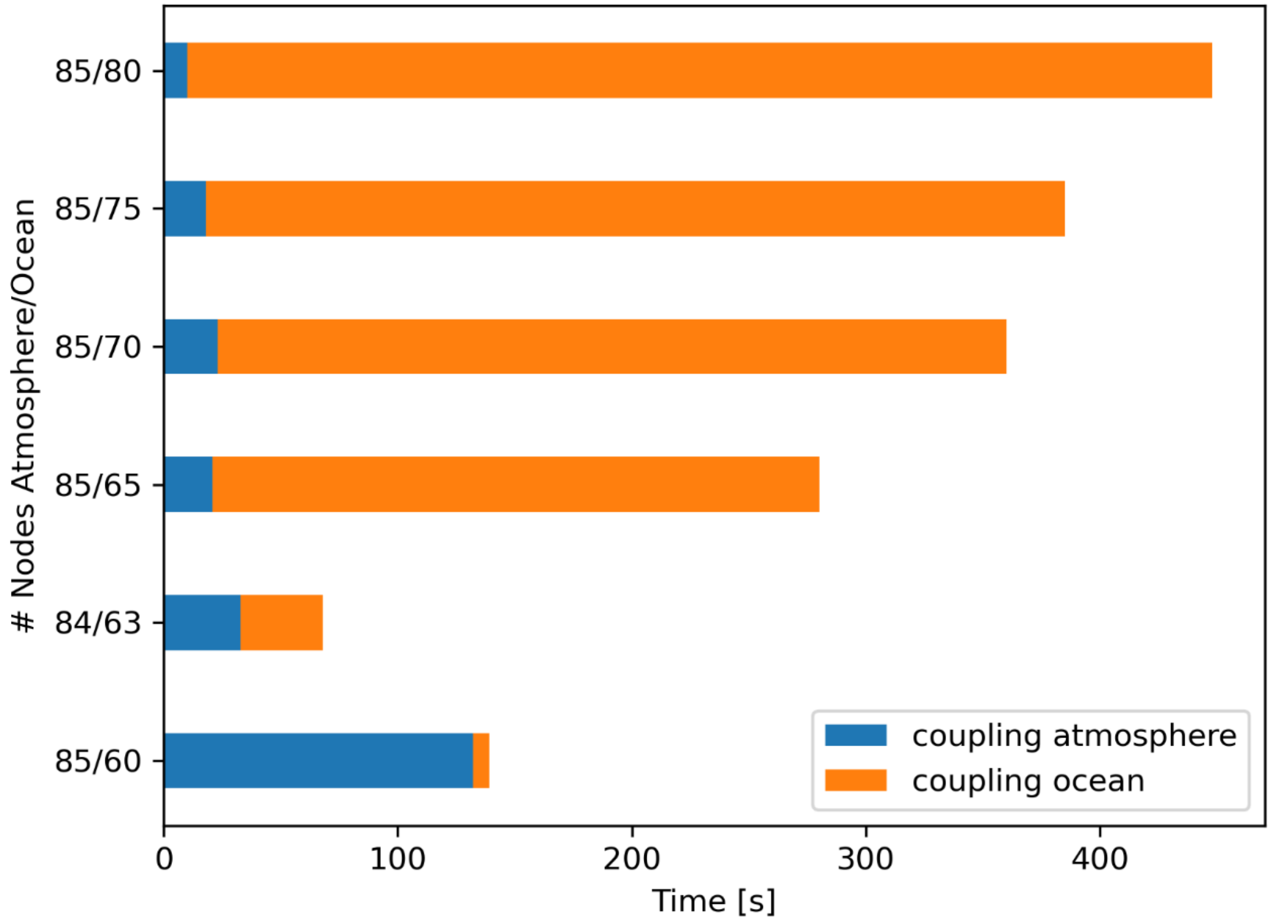
```
srun --export=ALL --propagate=STACK,CORE --label \
    -n $atm_compute_tasks --cpu-bind=verbose --accel-bind=v,g,n \
    xenv -L Stages/2020 -L NVHPC/21.5-GCC-10.3.0
         -L ParaStationMPI -L UCX/1.10.1 \
         -L netCDF/4.7.4 -L netCDF-Fortran/4.5.3 \
    env UCX_RNDV_SCHEME=get_zcopy UCX_RNDV_THRESH=8192 \
        UCX_TLS=rc_x,mm,cuda_ipc,cuda_copy,gdr_copy \
    $(pwd)/icon.nvhpc                    # Executable for the GPUs (JUWELS Booster)
    : \             # Separator between two parts of the heterog. Job
    -n $oce_compute_tasks --cpu-bind=verbose -B1:8:1 \
    xenv -L Stages/2020 -L Intel/2021.2.0-GCC-10.3.0
         -L ParaStationMPI -L UCX/1.10.1 \
         -L imkl/2021.2.0 -L netCDF/4.7.4 -L netCDF-Fortran/4.5.3 \
    env UCX_RNDV_SCHEME=get_zcopy \
        UCX_RNDV_THRESH=8192 UCX_TLS=rc_x,self,sm \
    $(pwd)/icon.intel.compute            # Executable for Intel CPU (JUWELS Cluster)

echo "Script run successfully"
```

Atmosphere computation on JUWELS Booster (het_group_0)

Ocean computation on JUWELS Cluster (het_group_1)

SWEET SPOT ANALYSIS

Aim: Figure out a configuration which minimizes idle times during coupling



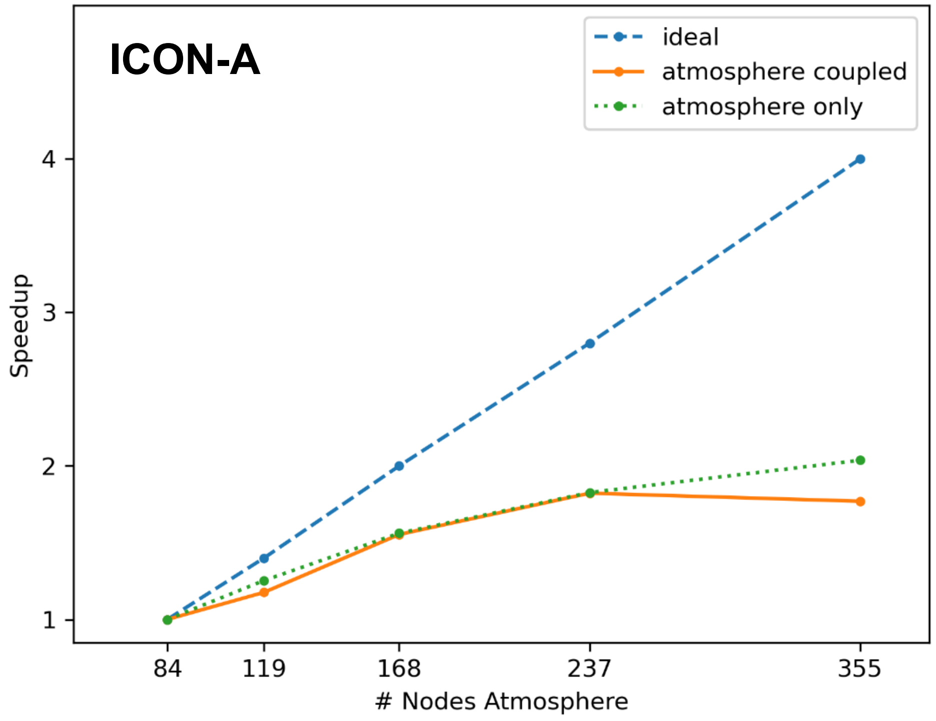
Sweet spot:

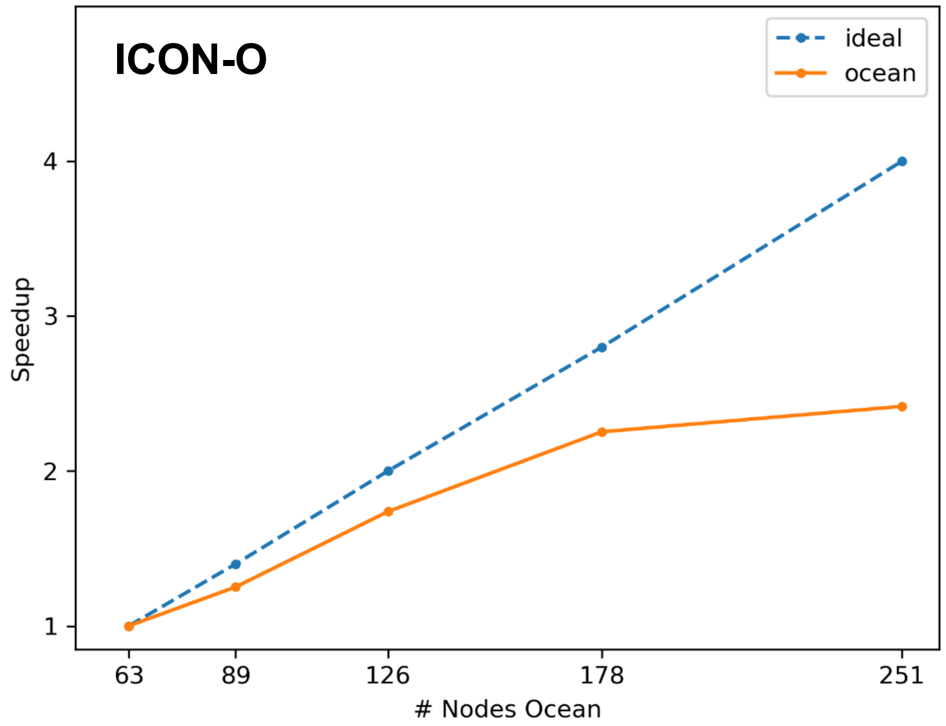
63 CPU nodes (ICON-O)

84 GPU nodes (ICON-A)

Note: ≥ (82 CPU nodes & 60 GPU nodes) are needed due to memory requirements

STRONG SCALING OF THE COMPONENTS





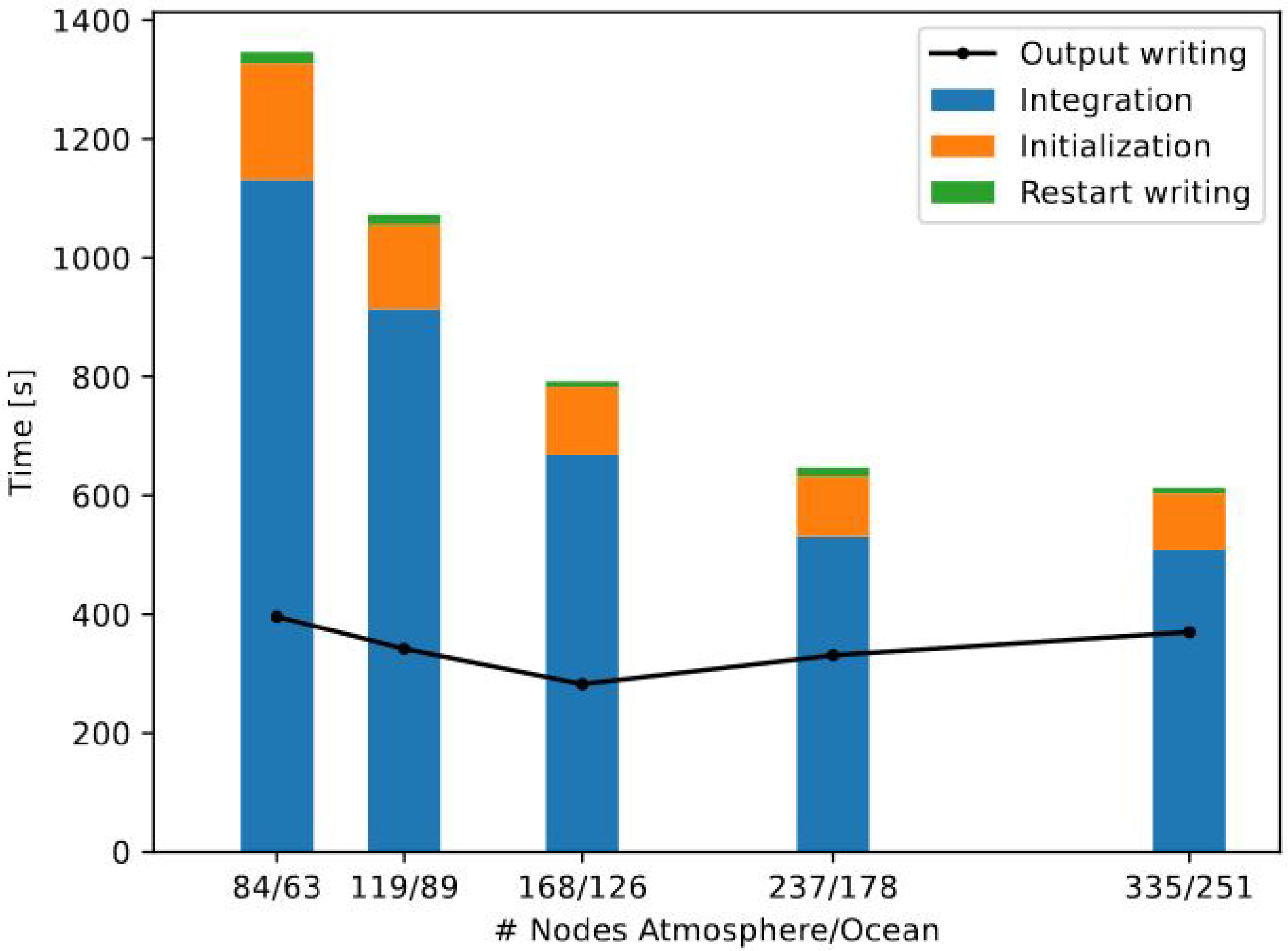
We scaled our application, keeping the relationship Booster/Cluster nodes constant

- **Strong scaling can be achieved for a node range of factor 2-3**
- **Higher model resolution will require larger node counts**
- **Coupling overhead is rather small**

JÜLICH Forschungszentrum

mehrText?

STRONG SCALING OF THE COUPLED MODEL



- **Scaling comparable to its components**
- **15% initialization phase due to short model runs of 1 simulation day**
- **Writing of output fields is done asynchronously (2.2 TB grib in total)**

PERFORMANCE

Model Throughput for the maximum of the integration times of ICON-A and ICON-O

# nodes Booster	# nodes Cluster	Integr. time ICON-A (s)	Integr. time ICON-O (s)	SDPD*
84	63	1131	779	76
119	89	931	733	95
168	126	668	514	130
237	178	531	399	163
335	251	508	283	170

* SDPD =
Simulation Days
Per Day

- **0.5 SYPD can be reached with ICON 5 km resolution**
- **On JUPITER, 0.5 SYPD is within reach for ICON-C (2.5 km)**
- **Less performant than in *Giorgetta et al. (2022)* and *Hohenegger et al. (2023)* due to additional coupling overhead and MPI choices**

ENERGY CONSUMPTION

- **Total energy consumption of all nodes involved in a particular run is used as a measure of resource efficiency**
- **absence of energy meters on each node → Use proxy for total energy consumption**
  - **Average total runtime * total Thermal Design Power (TDP) of all nodes involved**
  - **Total TDP: Adding the maximum Thermal Design Power of the CPU and if applicable all GPUs connected to a node**

COMPARISON NON-MSA VS. MSA:

# nodes ICON-A	# nodes ICON-O	Runtime ICON-A	Runtime ICON-O	Total TDP	Total Energy
780	63	1336 s	1275 s	352600 W	471 MJ
84	63	1321 s	1291 s	167200 W	221 MJ

MSA case is 1% faster than non-MSA, but needs 53% less energy

Note: Power consumed by the network is ignored

JÜLICH Forschungszentrum

CONCLUSIONS

- **With ICON-Sapphire, good scalability on MSA can be achieved**
- **Significant amount of energy can be saved**
- **For efficient simulations in the km range, exaflop HPC systems will be needed (→ JUPITER)**

The ICON use case shows the potential using MSA for Earth System Modelling

- **km-scale models cannot rely on CPU clusters only**
  - ➢ **availability of nodes**
  - ➢ **energy consumption**
- **Not all model parts will be ported to GPUs easily**
  - ➢ **Long-term development efforts needed**
  - ➢ **Some workflows may be better suited for CPUs**