Advances and pitfalls in climate modelling on the NEC SX-Aurora TSUBASA

Panagiotis Adamidis, Natalja Rakowsky
Deutsches Klimarechenzentrum (DKRZ)

People who contributed

- Dominik Zobel (DKRZ)
- Marek Jacob (DWD)
- Jens-Olaf Beismann (NEC)

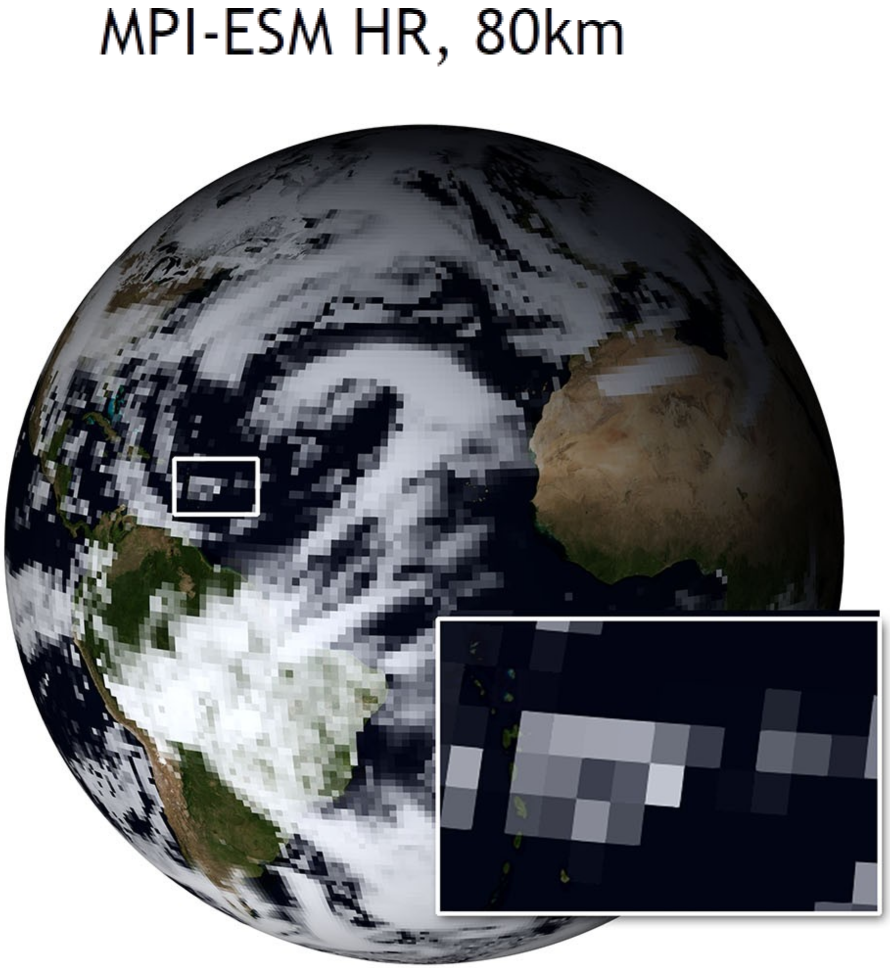
ICON High Resolution Simulations

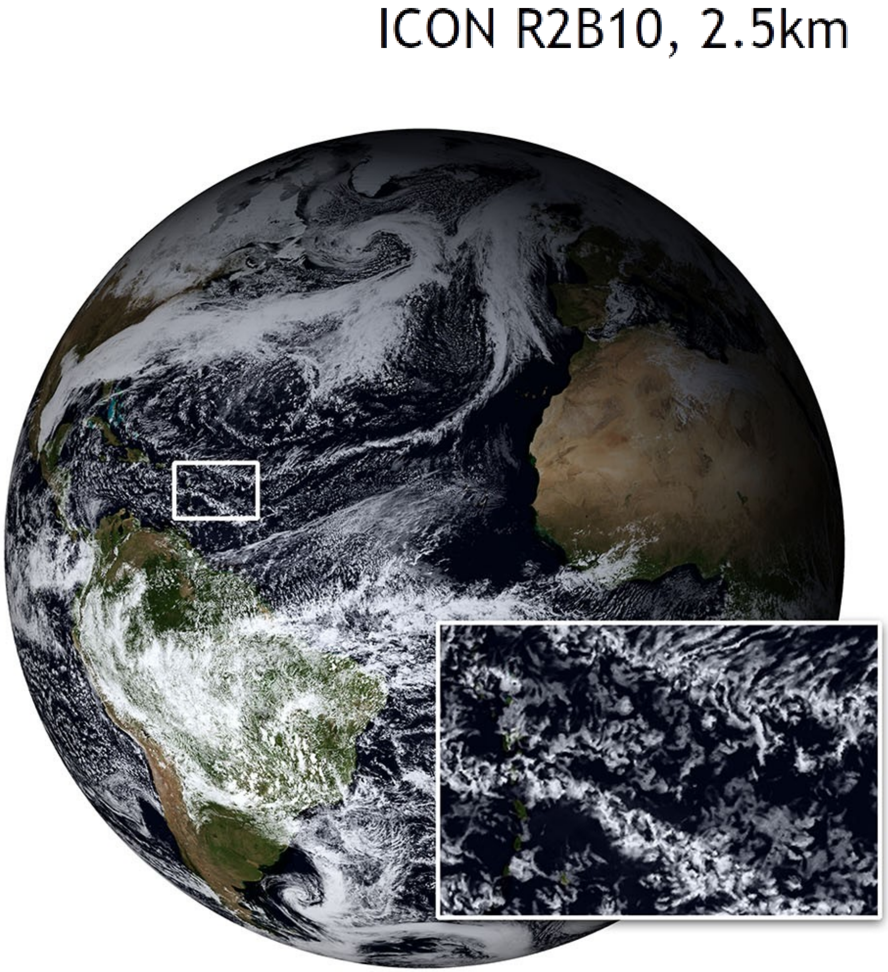
- ICON is being used in climate simulations with high resolution grids, in order to resolve small-scale physical processes.
- In this way, parameterisation and the inherent uncertainty can be avoided, thus improving significantly climate change projections.

ICON Grid Resolutions

grid	number of cells	avg. resolution
R2B04	20480	158 km
R2B05	81920	79 km
R2B06	327680	40 km
R2B07	1310720	20 km
R2B09	20971520	5 km
R2B10	83886080	2.5 km
R2B11	335544320	1.25 km

ICON Resolving Clouds

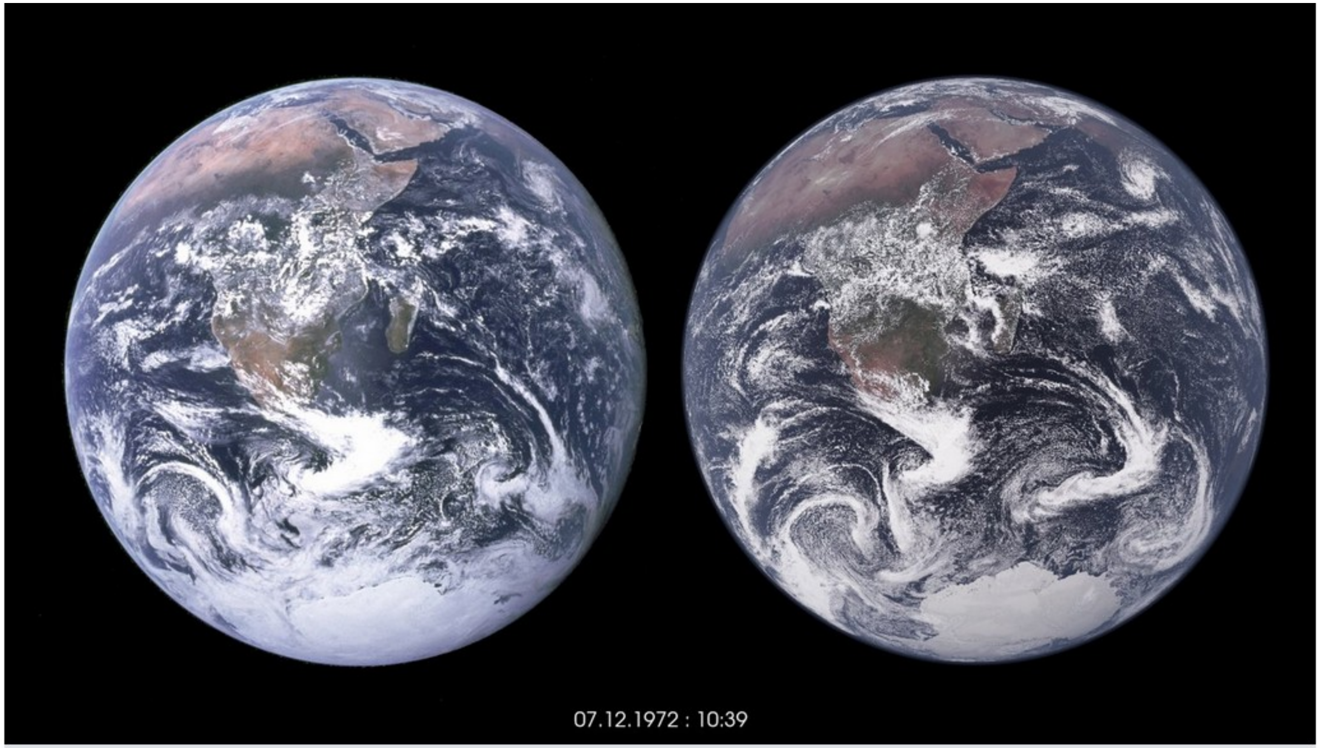




Florian Ziemen DKRZ

Blue Marble

ICON simulating the coupled climate system at 1 km



07.12.1972 : 10:39

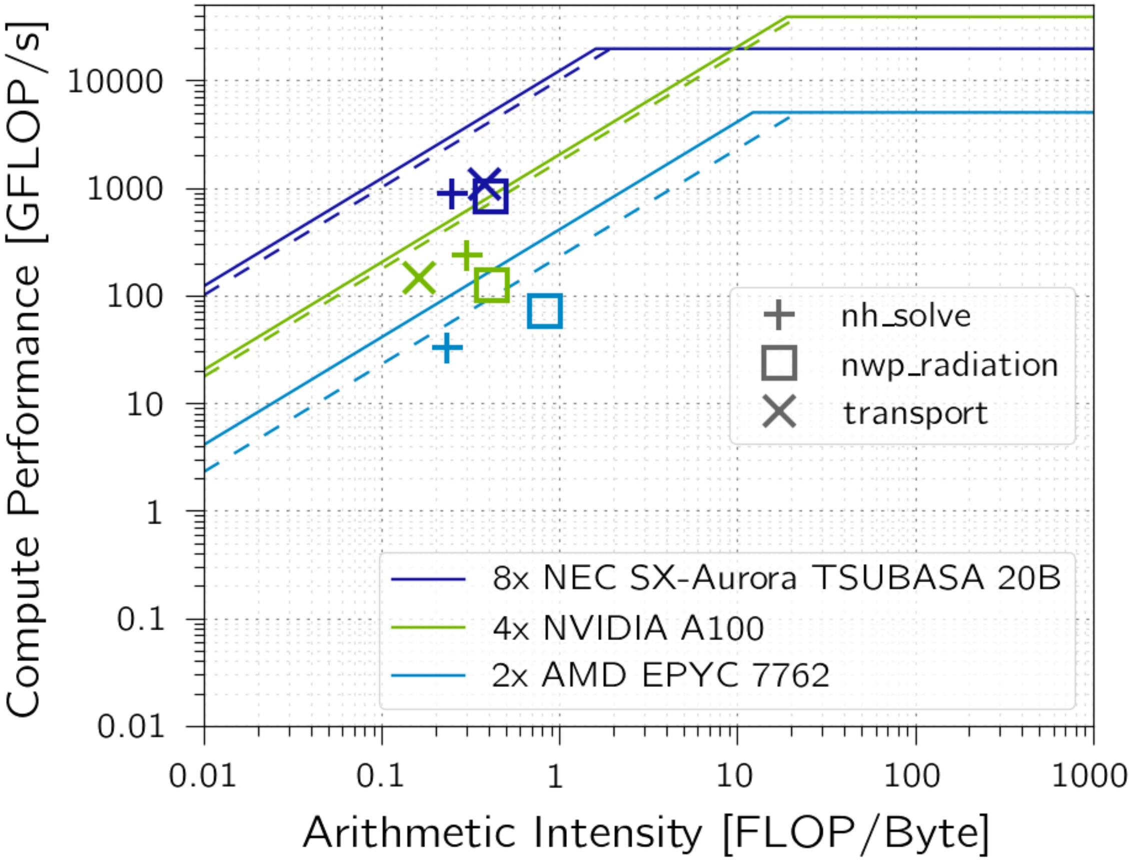
Original NASA Blue Marble photo left, visualization right. Credit: MPI-M, DKRZ, NVIDIA

https://mpimet.mpg.de/en/communication/detail-view-news-ii

Coupled Climate System @1km R2B11

- ICON-2.6.6-rc
- 90 vertical levels in the atmosphere 335544320 grid points per level
- 128 vertical levels in the ocean 237102291 surface grid points
- Time step atmo=8s ocean=45s
- 900 nodes of Levante (128 cores per node, AMD EPYC Milan CPUs) at DKRZ and use a split of 24:8 (atm:oce) mpi tasks per node, with 4 openMP threads
- Total Throughput = 3 SDPD on 900 nodes (about 1/3 of Levante).

Rooflines Single Node : Experiment R2B6N7



Comparison VE2 .vs. VE3 Exp. R2B6N7

VE2 = 8 cores/VE
#VEs = 8
Wallclock = 155.8 sec

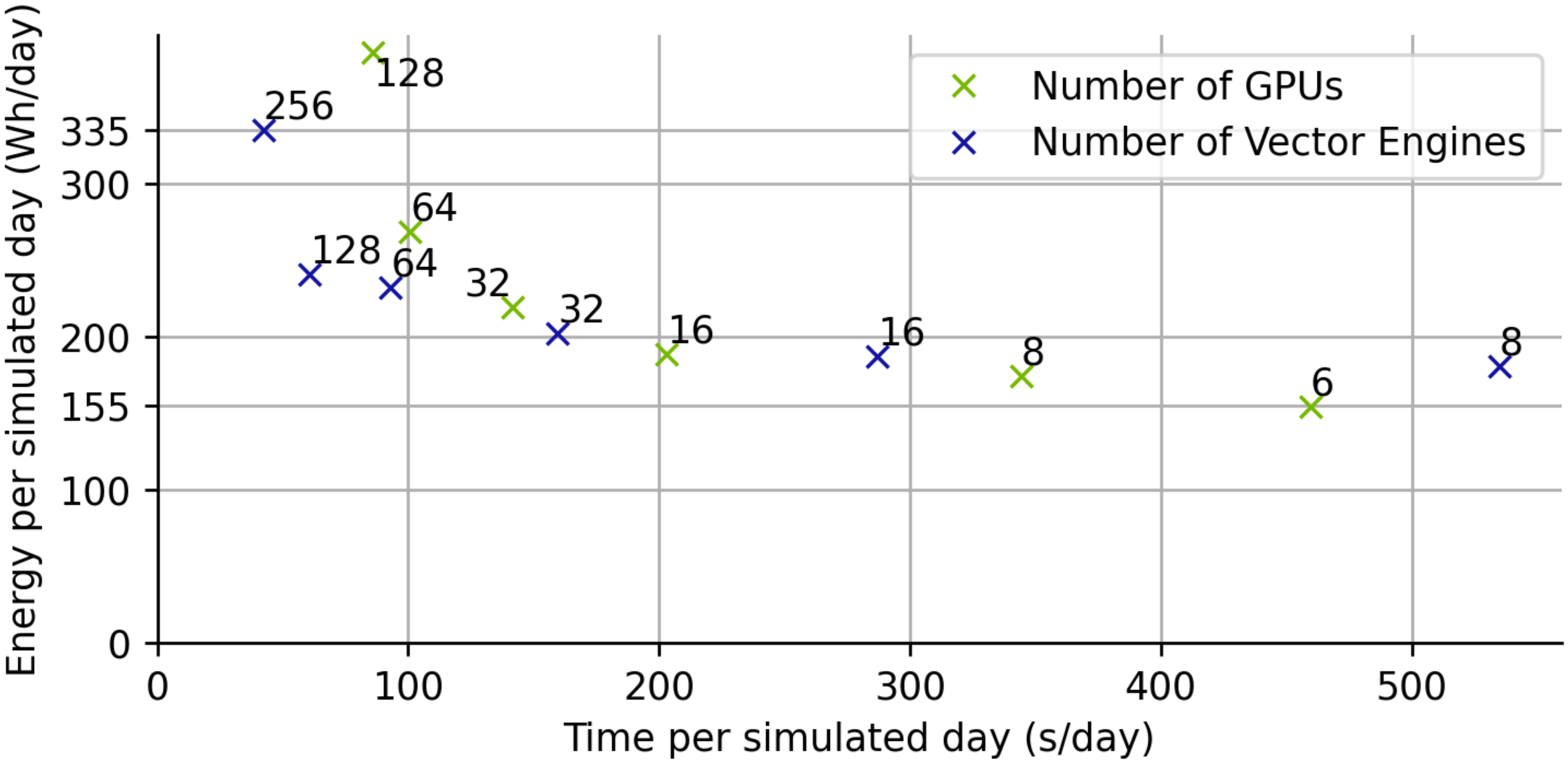
VE2	MFLOPS	ACT. B/F	PROC.NAME
	13939.3	4.01	solve_nh
	13127.0	2.45	nwp_radiation
	16928.1	2.66	transport

VE3 = 16 cores/VE
#VEs = 4
Wallclock = 144.8 sec

VE3	MFLOPS	ACT. B/F	PROC.NAME
	16015.5	**3.53**	solve_nh
	13124.4	2.29	nwp_radiation
	19002.8	**2.33**	transport

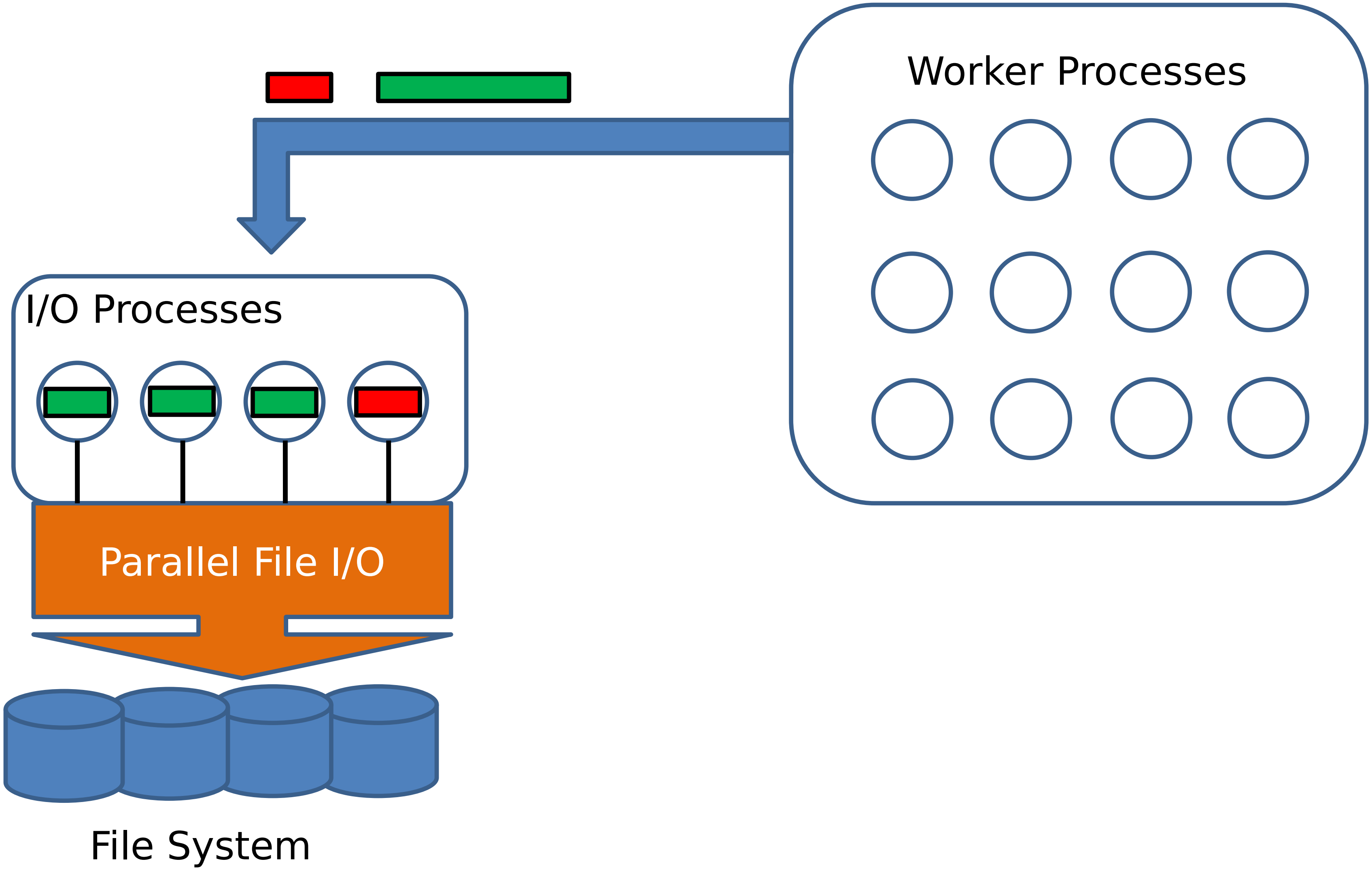
Jens-Olaf Beismann (NEC)

Energy Efficiency: Experiment @R2B07

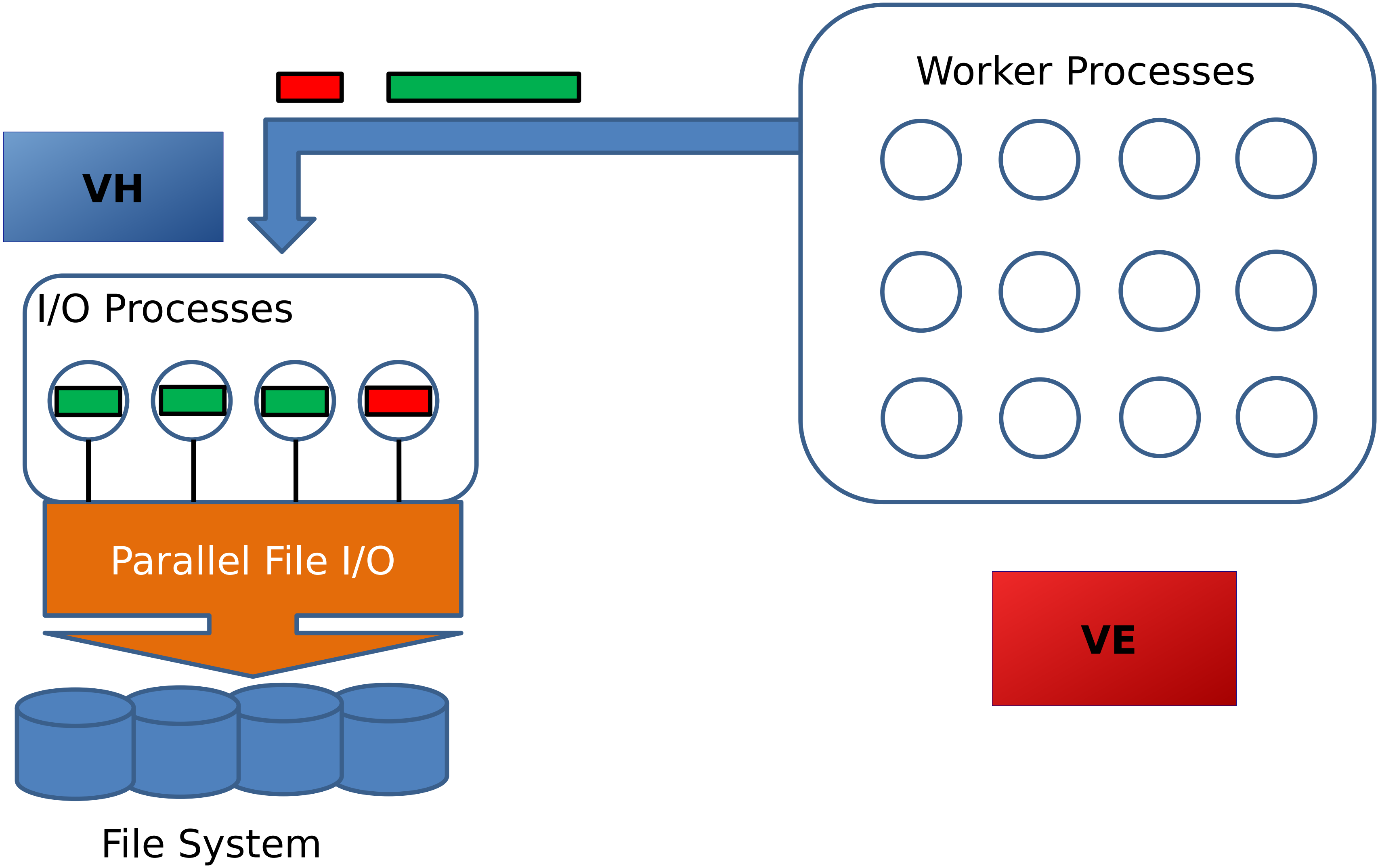


M. Jacob et al., "ICON-GPU for Numerical Weather Prediction – A Status Report", PASC 23, Davos, 2023 https://pasc23.pasc-conference.org/posters/

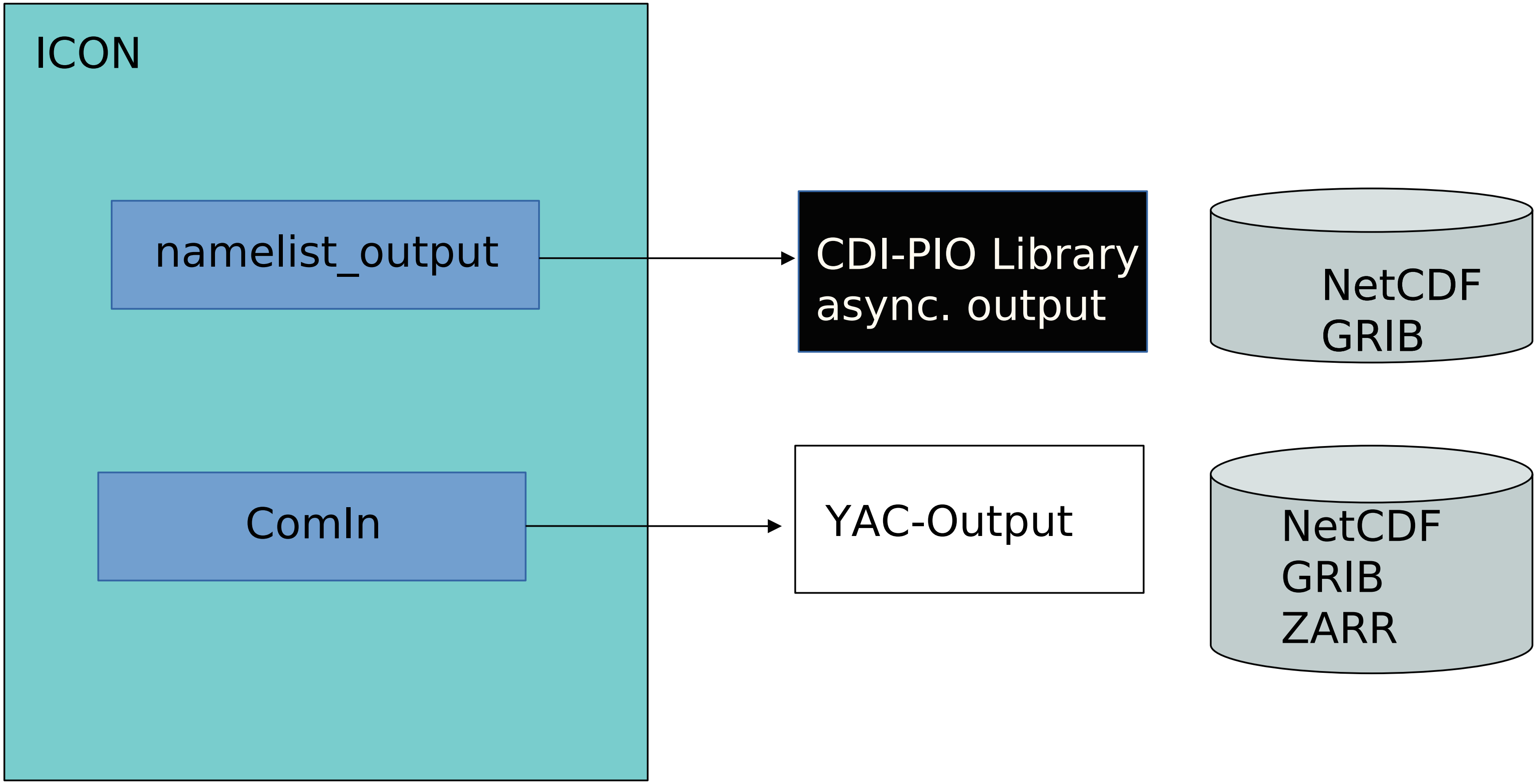
Parallel I/O in ICON



Parallel I/O in ICON



ICON : Asynchronous/Parallel Output



Test system at DKRZ

- 2x SX-Aurora TSUBASA A300 with 2VE each
- Rocky Linux 8.6
- NEC programming environment with node locked licenses
- Lustre Filesystem from Levante mounted via TCP/IP
- Slurm 22.05.2 with add-ons to handle VE hardware
  `https://sxauroratsubasa.sakura.ne.jp/Special:WikiForum/Preview_release_2_of_SLURM_for_VE`

Remarks on slurm

- Installation: compile source code, follow the provided README to setup slurm.conf, gres.conf,… not complicated
- Pitfall: Unique host names are crucial – we had aurora1,2 in DNS, internally aurora01,02 => NEC MPI in Slurm failed with OOM.
  W/o Slurm: very high virtual mem
- NEC MPI on VE is officially supported,
  on VH and hybrid it works, too
  (though a bit old fashioned with hostfile)

Resource Specification for VE

```
#SBATCH --gres=ve:10b:1,hca:1 # 1 VE and 1 IB card
#SBATCH -overcommit           # required for NEC MPI
```

- Problem: our test cluster has only one hca card per node. Even with over-commit, only one MPI job runs at a time.
- But it is a small development system, we want shared usage...
- First tests: skipping hca:1 seems ok.

Conclusion & Outlook

- NEC SX-Aurora Tsubasa architecture has good potential to deliver energy efficient sustained performance
- Key to success is single node performance
- Good scaling over hundreds/thousands of nodes is necessary
- Efficient parallel I/O is vital for high resolution simulations
- The performance of the file system is very important

Thanks for your attention !

Questions ?