Parallel IO NHR Workshop

<Panagiotis Adamidis>
Deutsches Klimarechenzentrum (DKRZ)

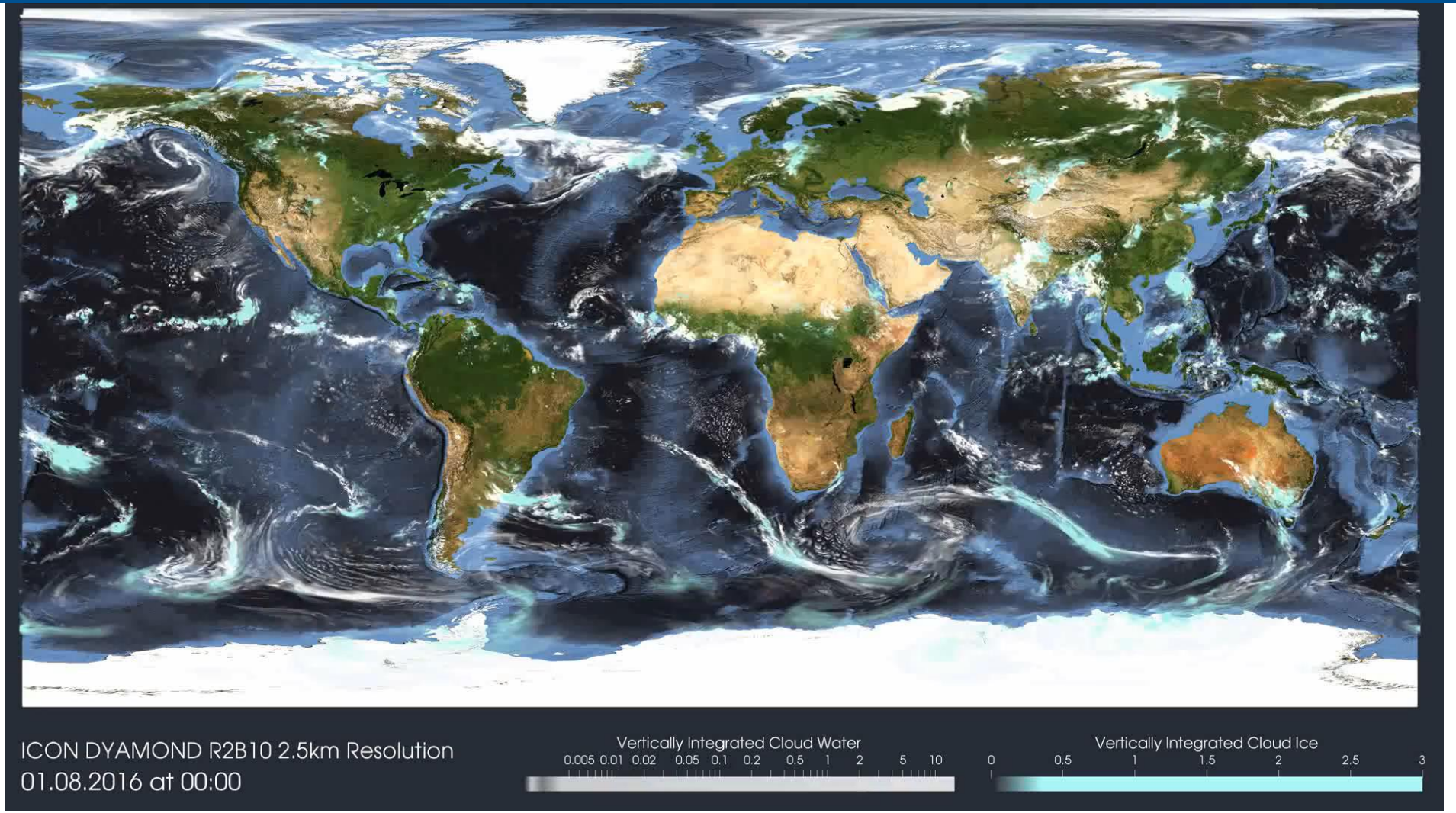

The German Climate
Computing Center
(DKRZ)

The German Climate Computing Center
Supercomputing for Earth System Research

Prof. Dr. Thomas Ludwig ***Deutsches*** *Klimarechenzentrum (DKRZ)*

Parallel IO NHR Workshop

- Anja Gerbes (TU Dresden)
- Anna Fuchs (University of Hamburg)
- Jannek Squar (University of Hamburg)


ICON DYAMOND R2B10 2.5km Resolution
01.08.2016 at 00:00
Vertically Integrated Cloud Water
0.005 0.01 0.02 0.05 0.1 0.2 0.5 1 2 5 10
Vertically Integrated Cloud Ice
0 0.5 1 1.5 2 2.5 3

Parallel IO NHR Workshop

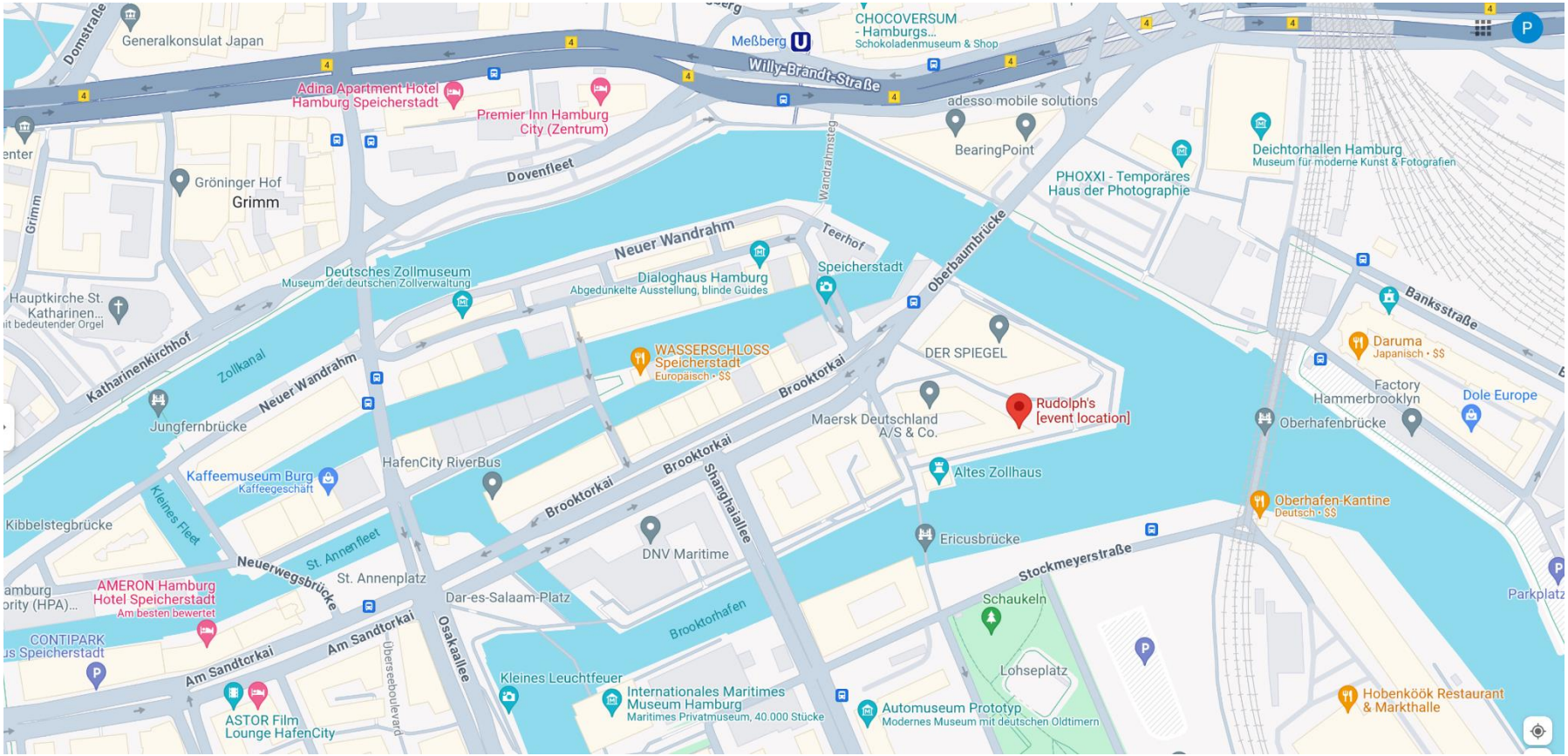
Dinner at Rudolph's in HafenCity

Ericusspitze 2-4 , 20457 Hamburg

https://www.rudolphs-hamburg.de/

7.5.2024 @ 7pm

Ericusspitze 2-4 , 20457 Hamburg



I/O in Climate Modeling

<Panagiotis Adamidis>
Deutsches Klimarechenzentrum (DKRZ)

Contributions

- Luis Kornblueh (MPI-Met)
- Thomas Jahns (DKRZ)
- Xingran Wang (DKRZ)
- Harald Braun (Eviden)

I/O in Climate Modeling

- A lot of data
- A lot of **data movement**

Climate Models @ DKRZ

- High Resolution
  - Resolving small scale physical processes
- Coarse Resolution
  - Simulating longer periods (80000 – 100000 years)
  - Complete glacial cycles

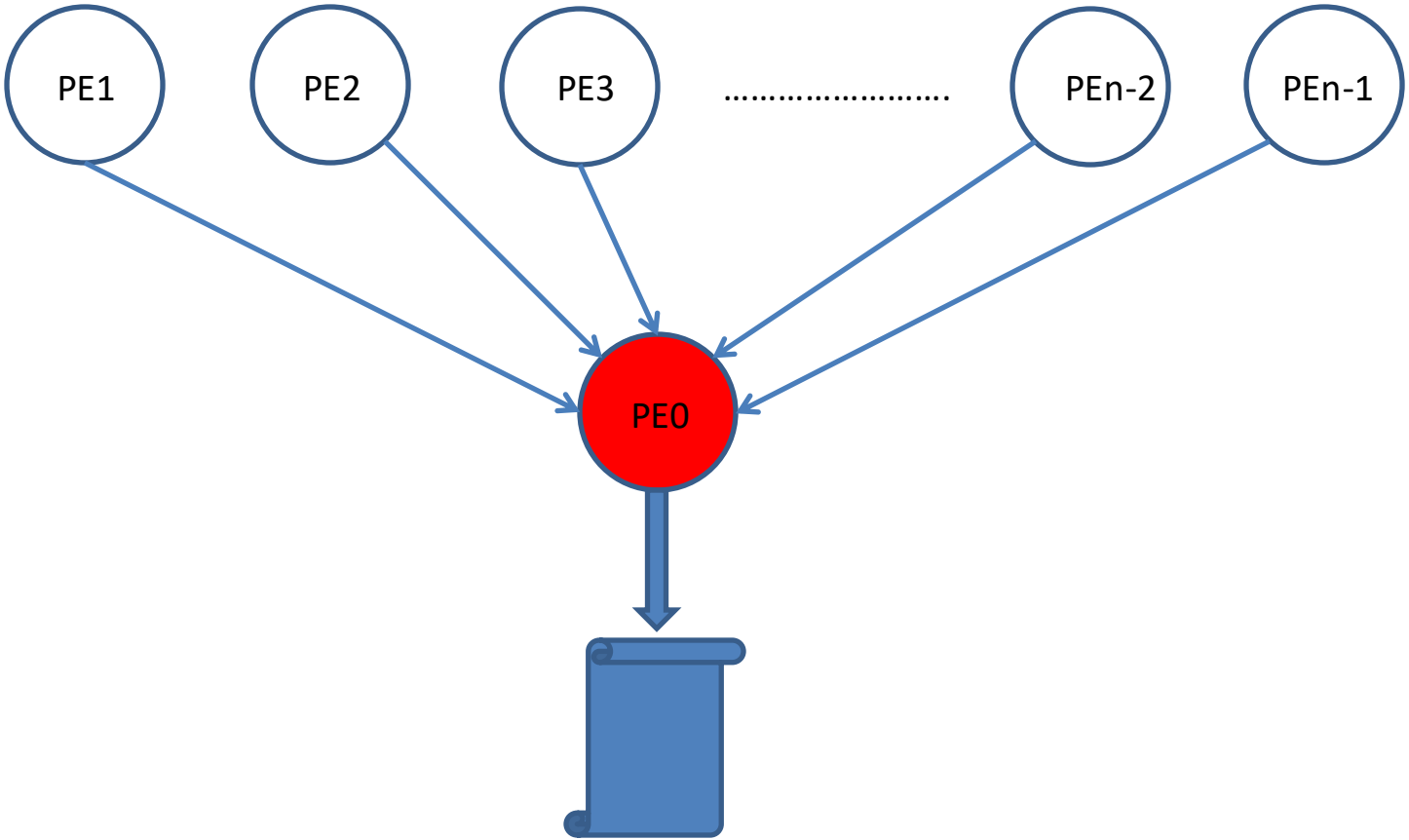
ICON Grid Resolutions

Global:

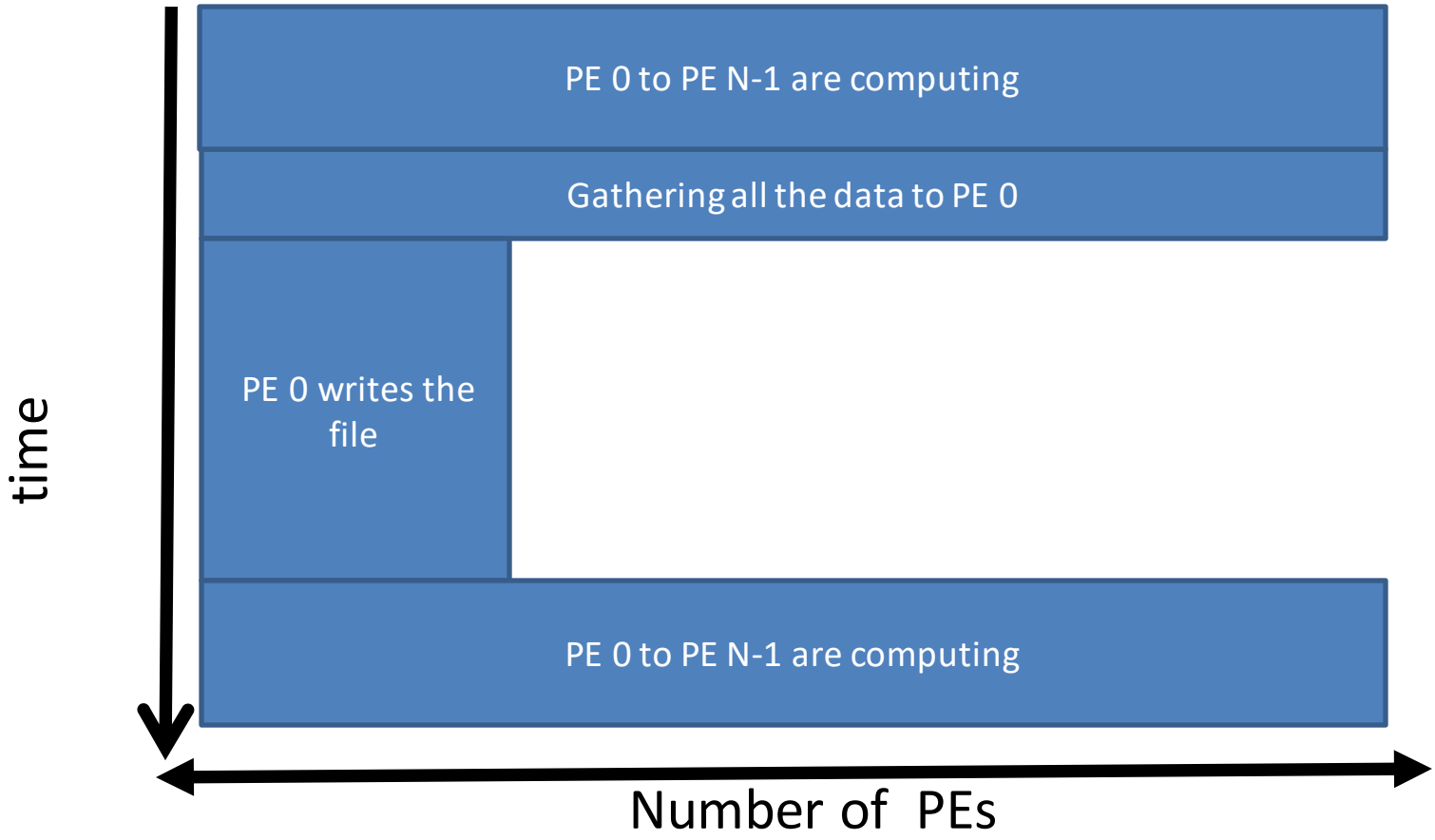
grid	number of cells	avg. resolution
R2B04	20480	158 km
R2B05	81920	79 km
R2B06	327680	40 km
R2B07	1310720	20 km
R2B09	20971520	5 km
R2B10	83886080	2.5 km
R2B11	335544320	1.25 km

Local Area : HD(CP)^2 nested grids over Germany 625m, 315m, 256m

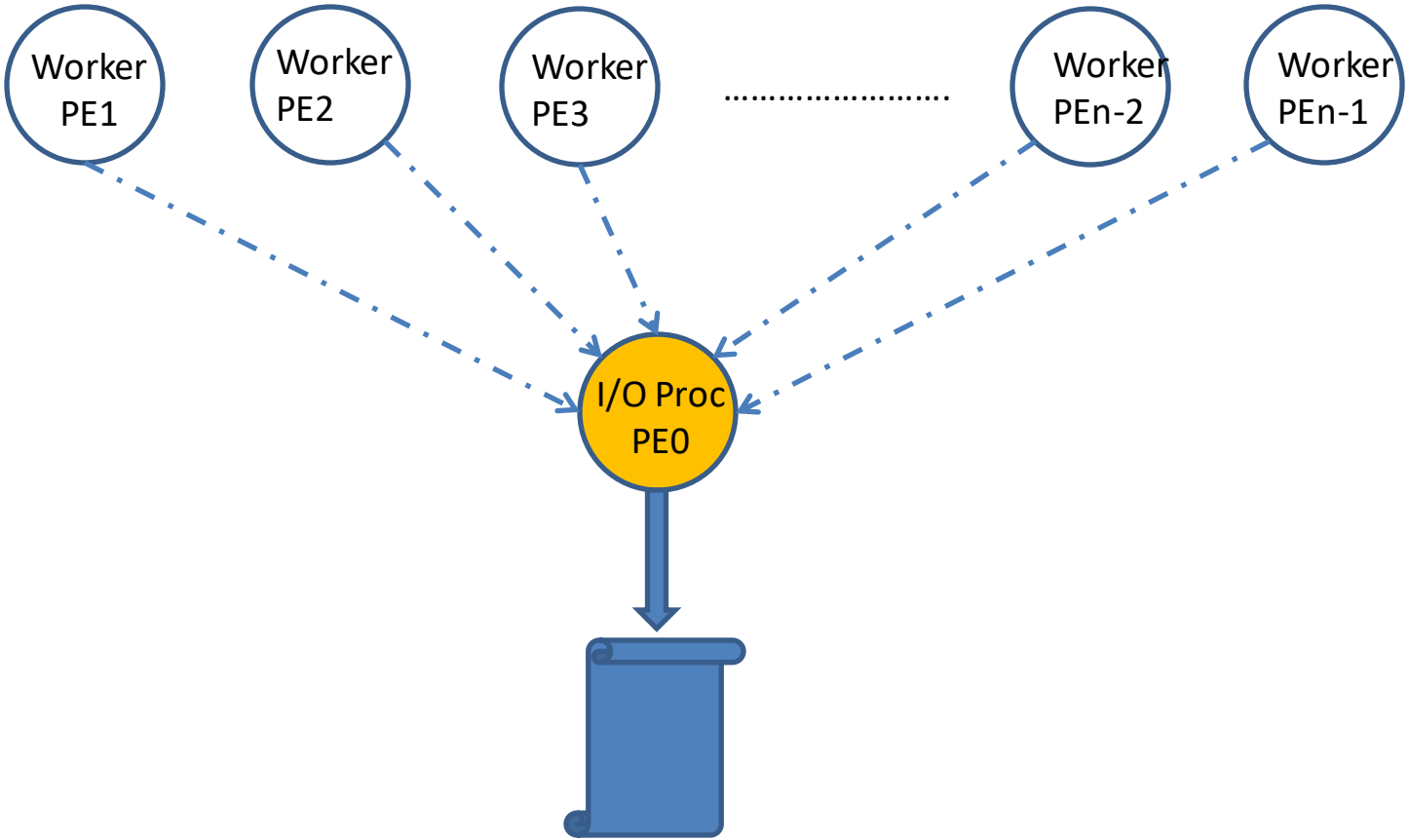
Serial I/O in ICON



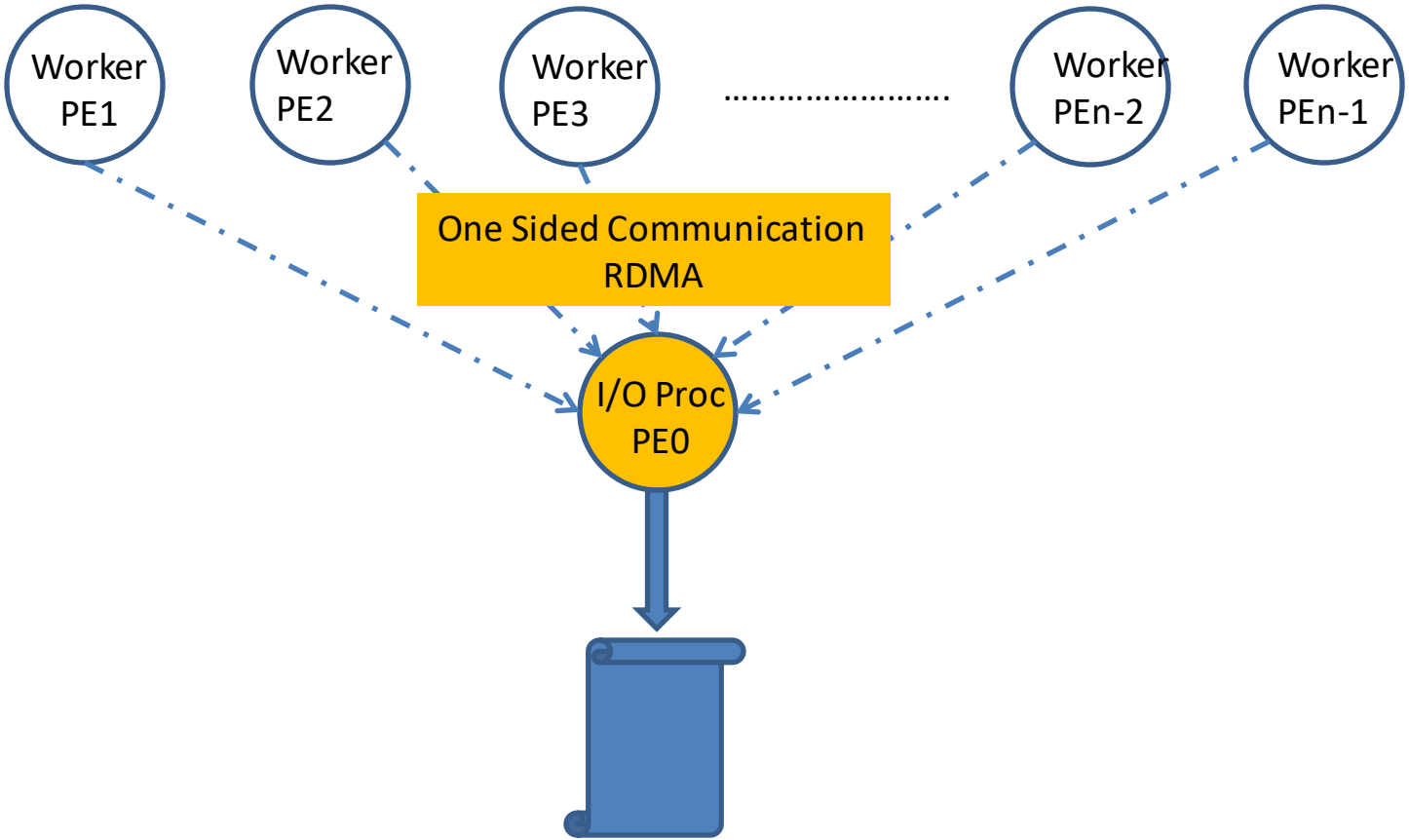
Serial I/O in ICON



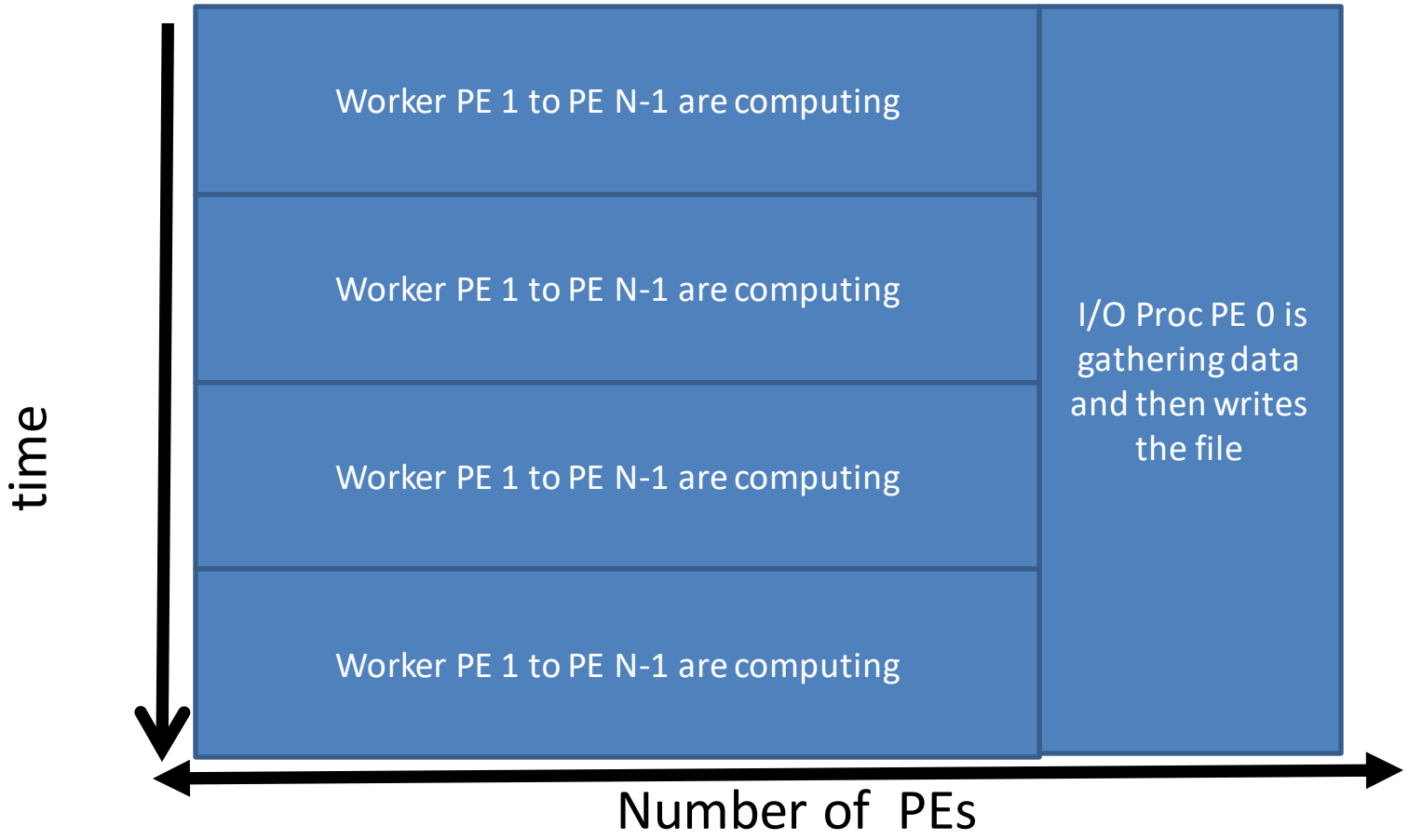
Asynchronous I/O in ICON



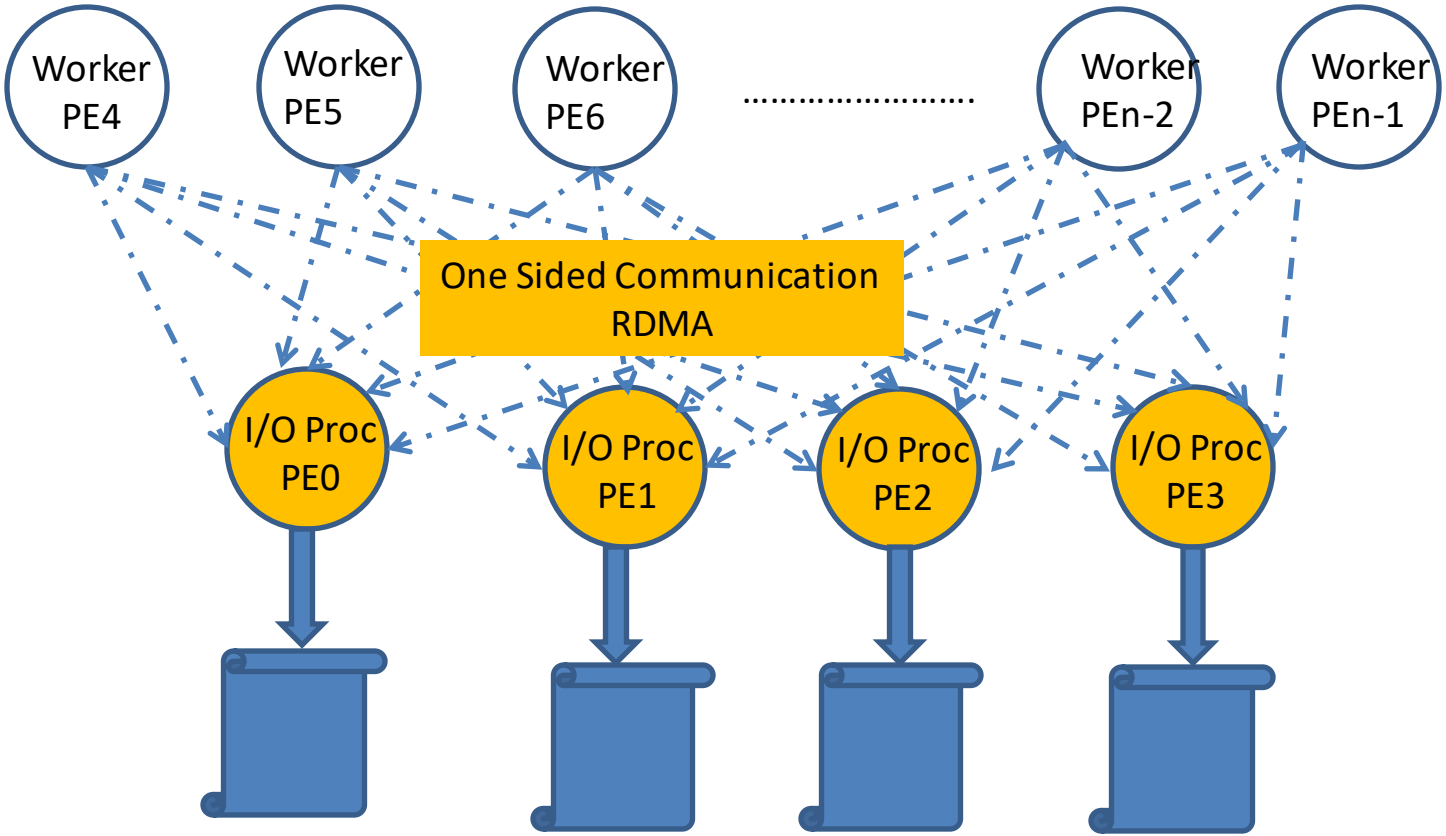
Asynchronous I/O in ICON



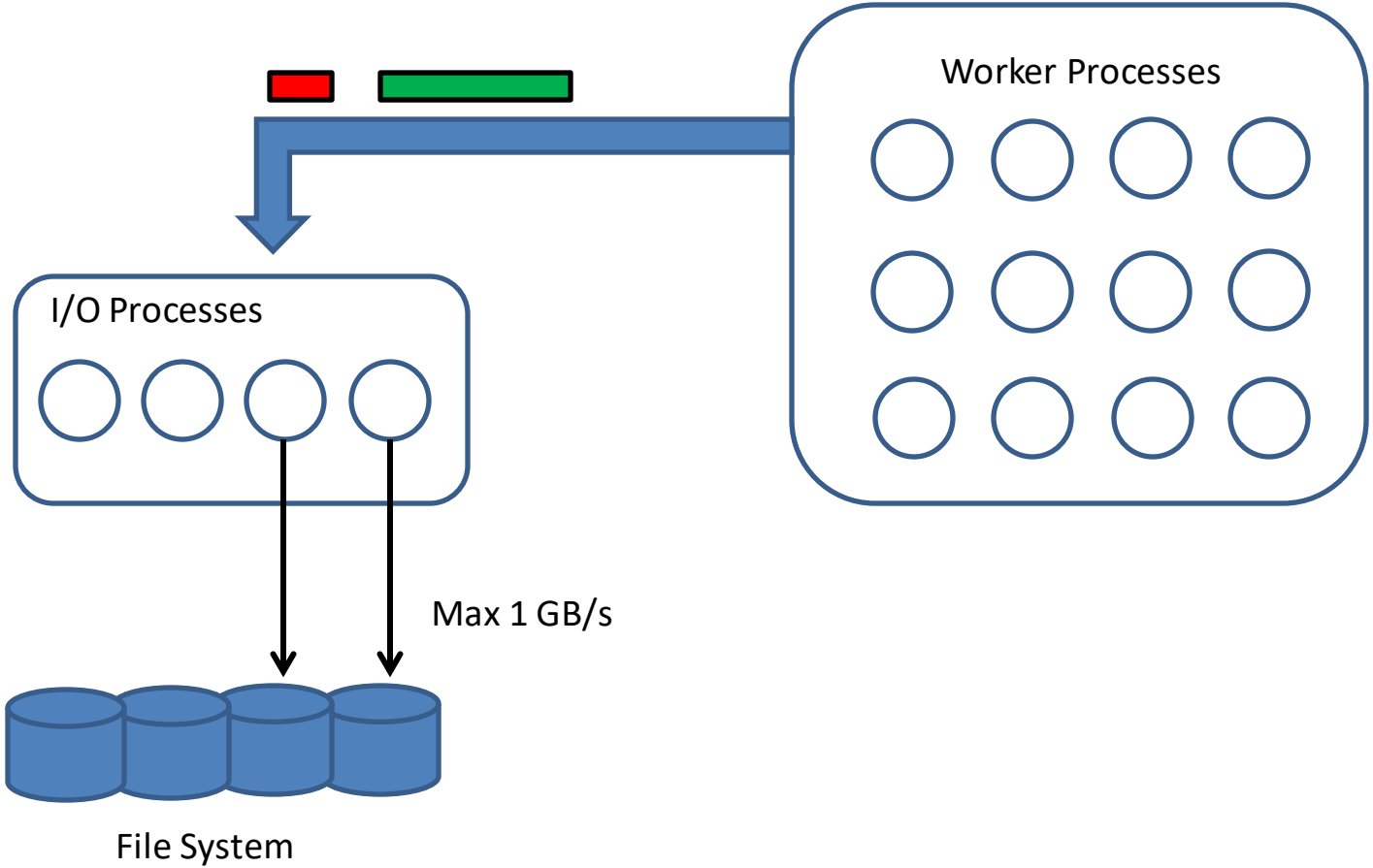
Asymchronous I/O in ICON



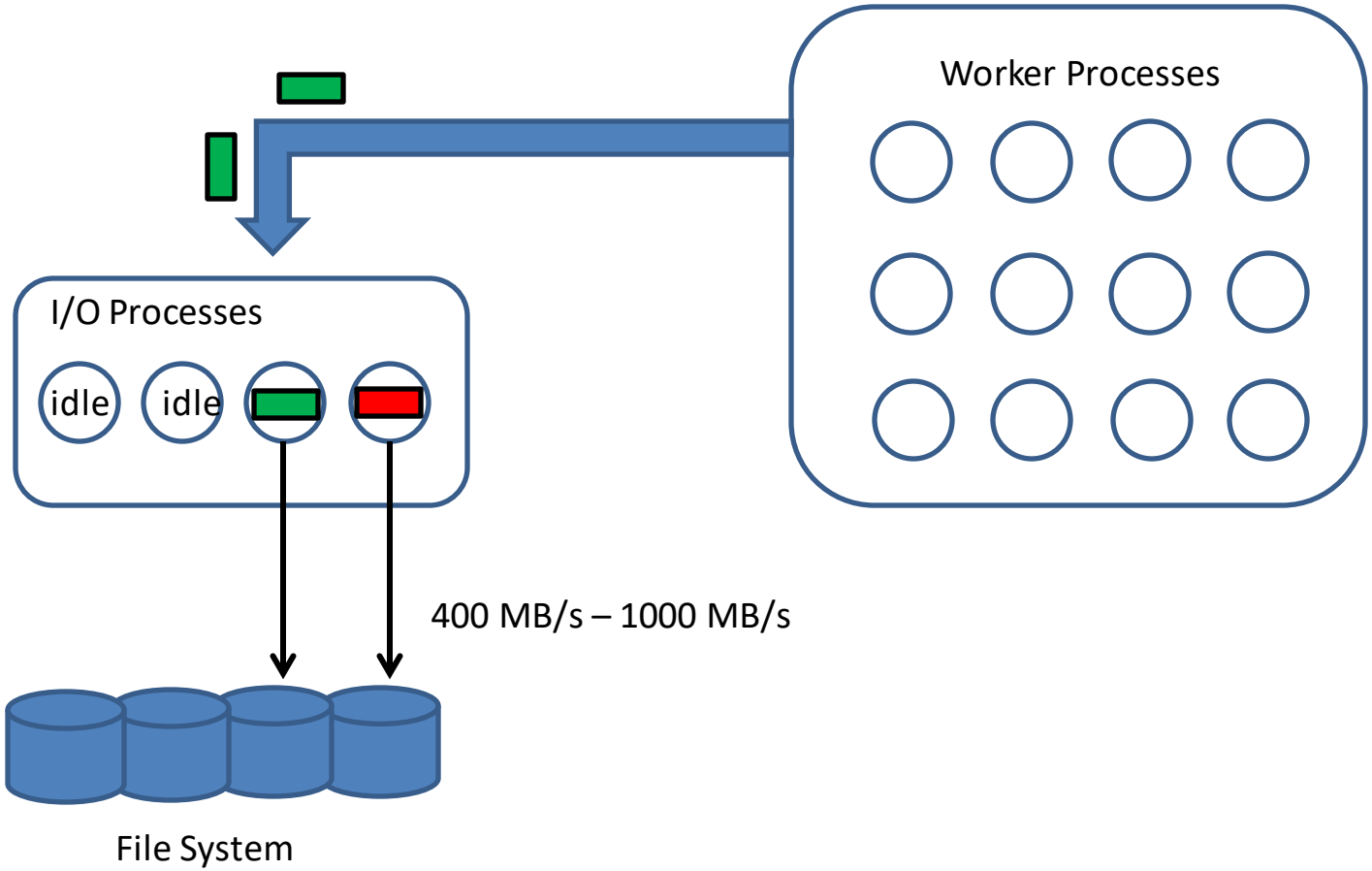
Asynchronous File Parallel I/O in ICON



Asynchronous File-Parallel I/O in ICON



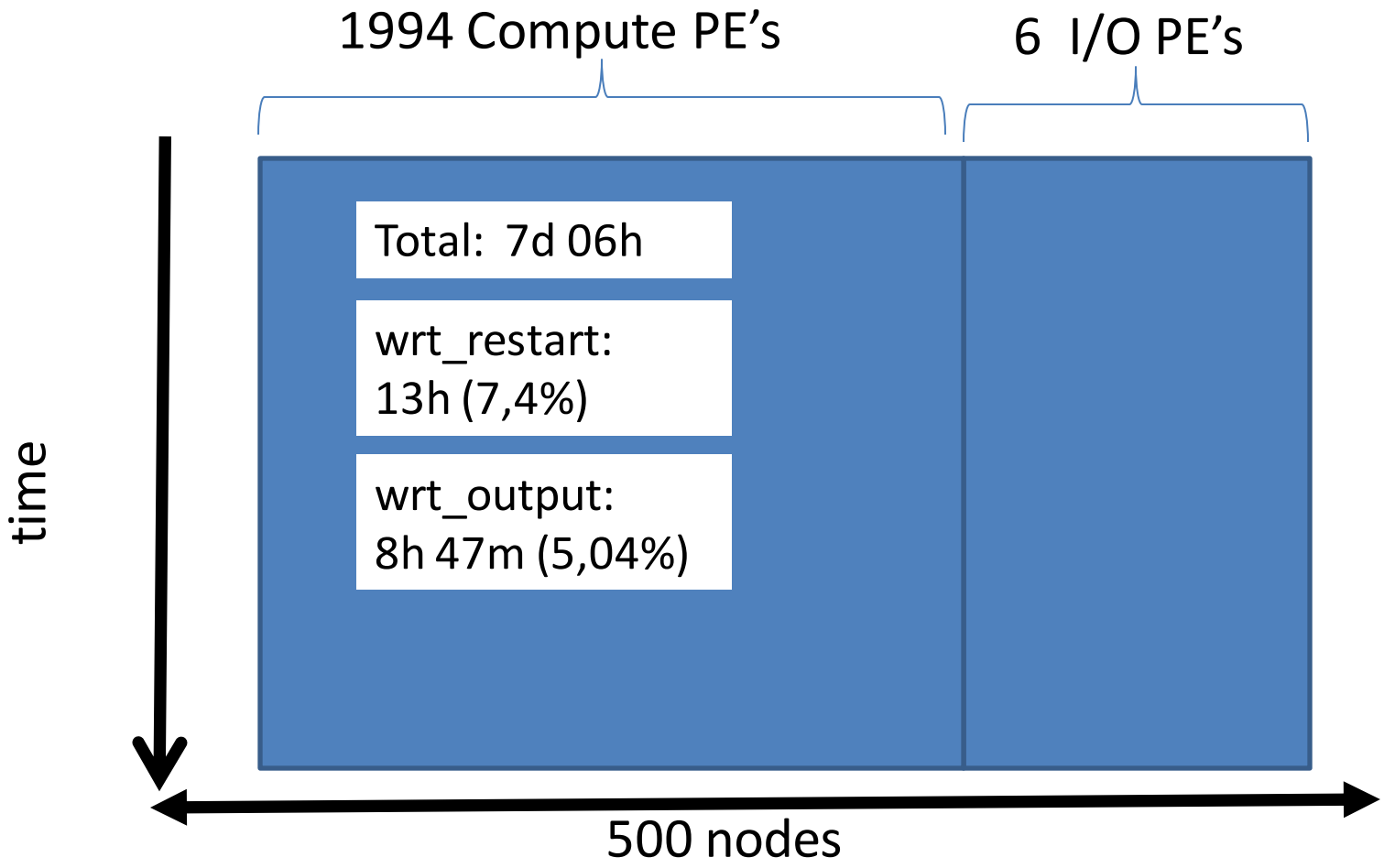
Asynchronous („Parallel“) I/O in ICON



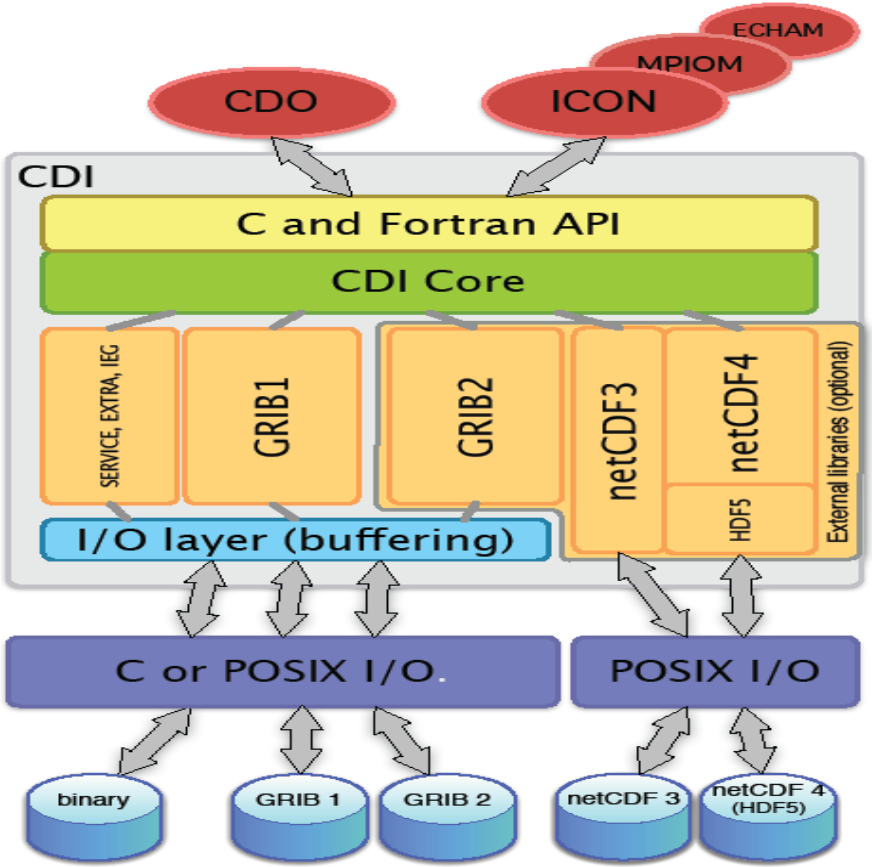
HD(CP)^2 Phase-I Final Experiment

- 3 Domains (625m, 312m, 156m)
- Output of 169 variables (2D/3D) at different intervals (9s, 10s,5min,15min,30min,1hour)
- 1 model day on 500 mistral nodes
  - with 4 MPI Processes x 6 OpenMP threads per node
  - Wallclock : 7days 6hours
  - Total size of Data: 48 TB

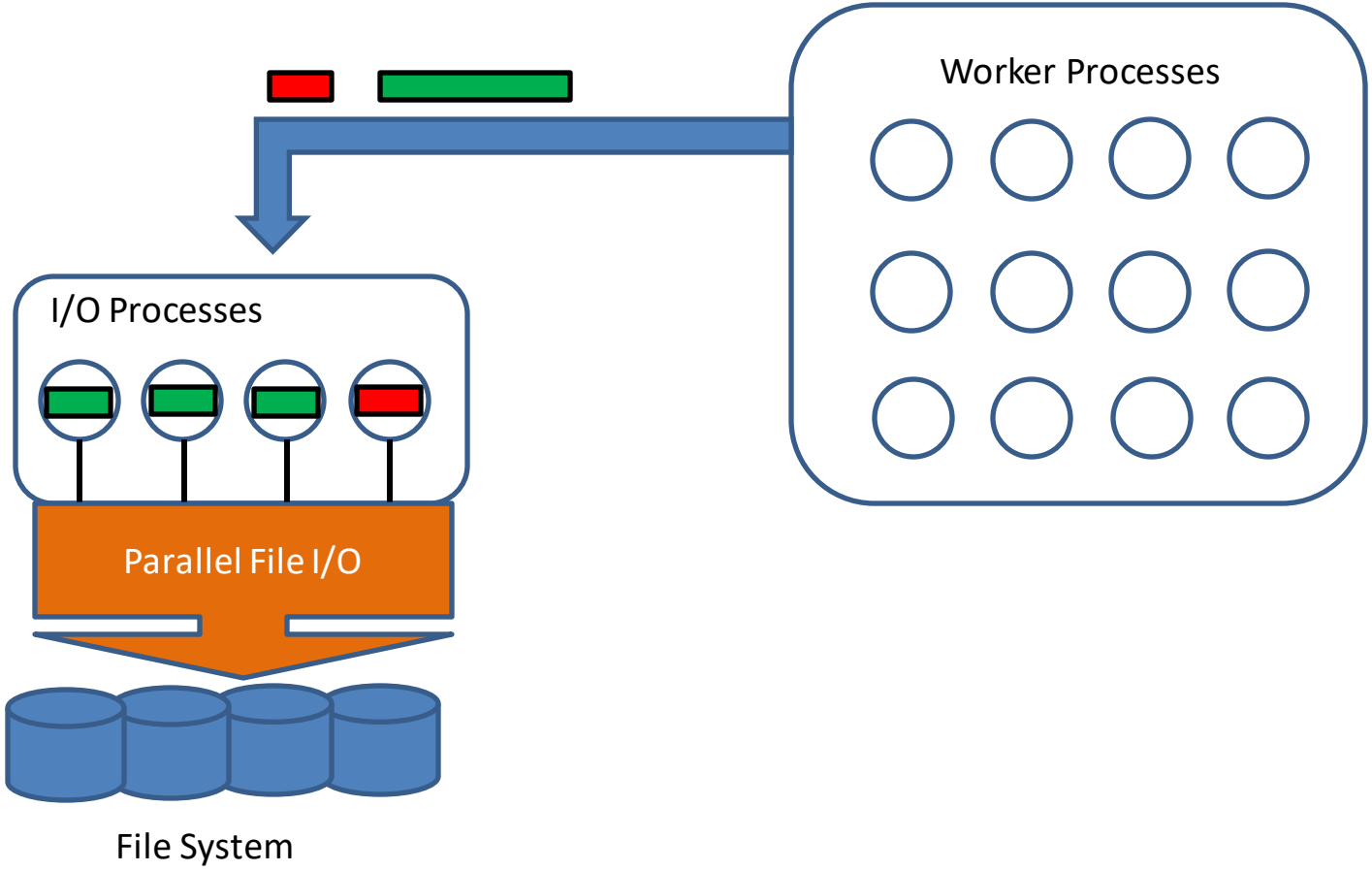
ICON-Parallel Asynchronous Output



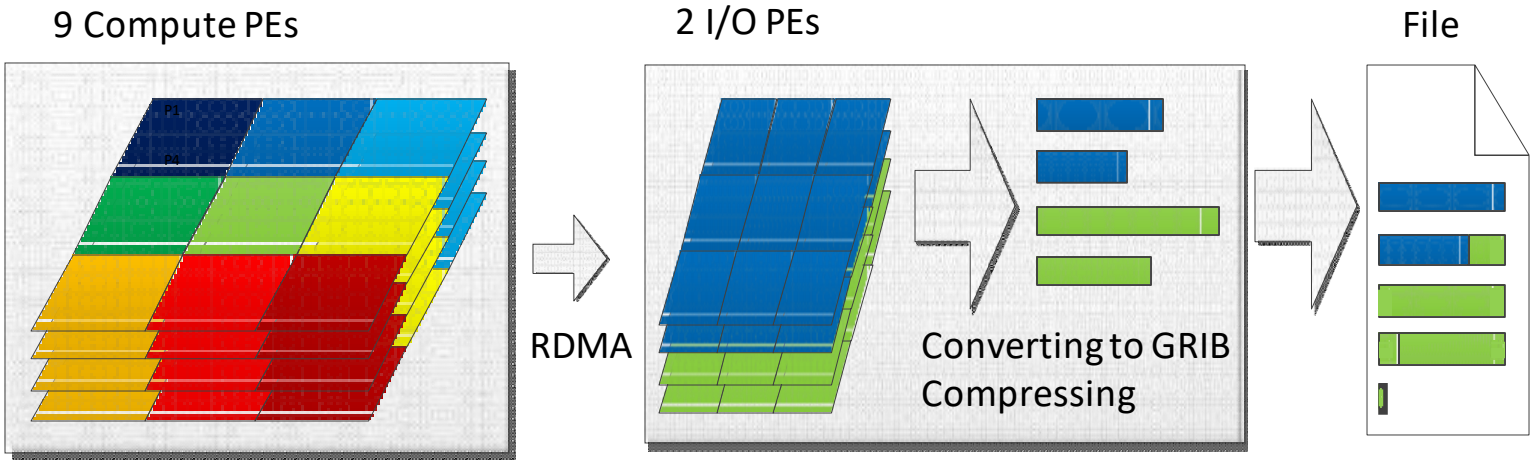
CDI-PIO


ECHAM
MPIOM
CDO
ICON
CDI
C and Fortran API
CDI Core
SERVICE, EXTRA, IEG
GRIB1
GRIB2
netCDF3
netCDF4
HDF5
External libraries (optional)
I/O layer (buffering)
C or POSIX I/O.
POSIX I/O
binary
GRIB 1
GRIB 2
netCDF 3
netCDF 4 (HDF5)

Parallel I/O in ICON

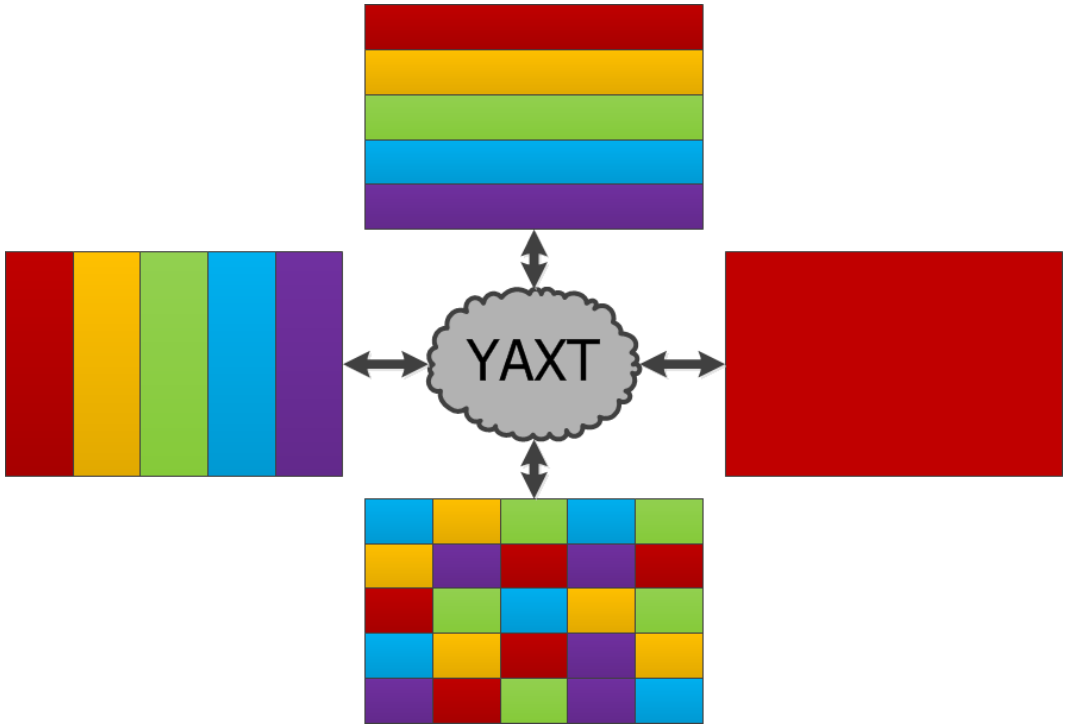


Example with 9 Compute PEs and 2 I/O PEs



YAXT : Yet Another eXchange Tool

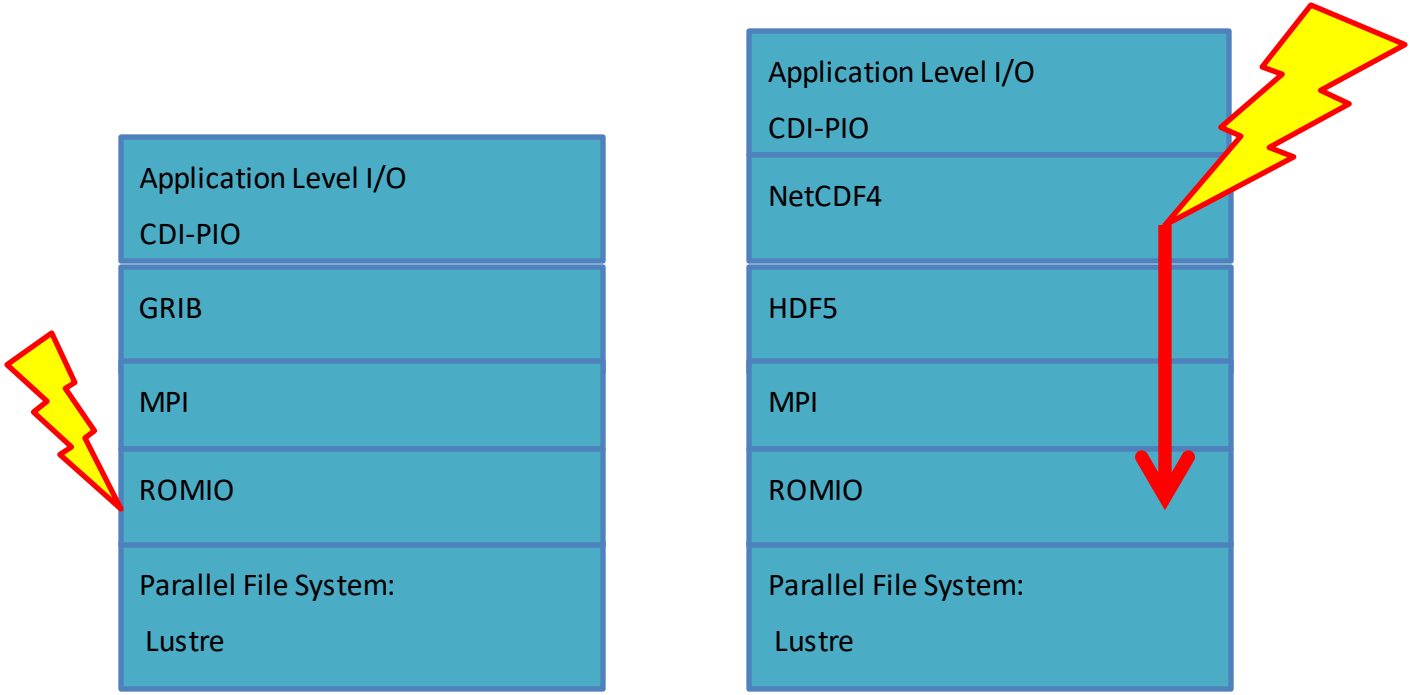
- Redistribution of data between two sets of processes



Tuning of CDI-PIO

- Max single stream performance = 1 GB/s
- Performance in CDI-PIO
  - of parallel GRIB output < 1GB/s
  - of parallel NetCDF4 < 1GB/s

Tuning CDI-PIO through ROMIO-Hints



Tuning of CDI-PIO via ROMIO-Hints

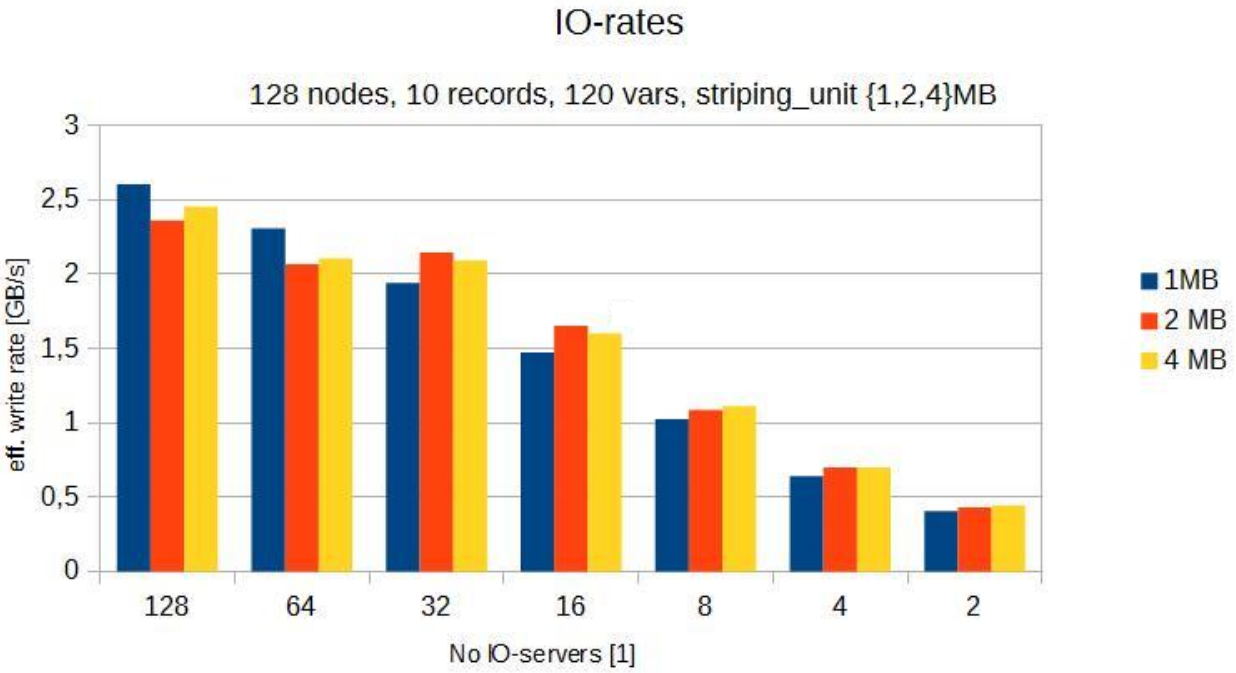
- **striping_factor:** determines on how many OST's (object storage targets) a file will be distributed
- **striping_unit:** size of stripes in Byte
- **Performance = f(IO_PES/node, striping_factor, striping_size,……..)**
  - Striping_factor = g (nHosts, maxStripes)
  - Striping_unit = k * (1024*1024 Byte)

GRIB Output Benchmark

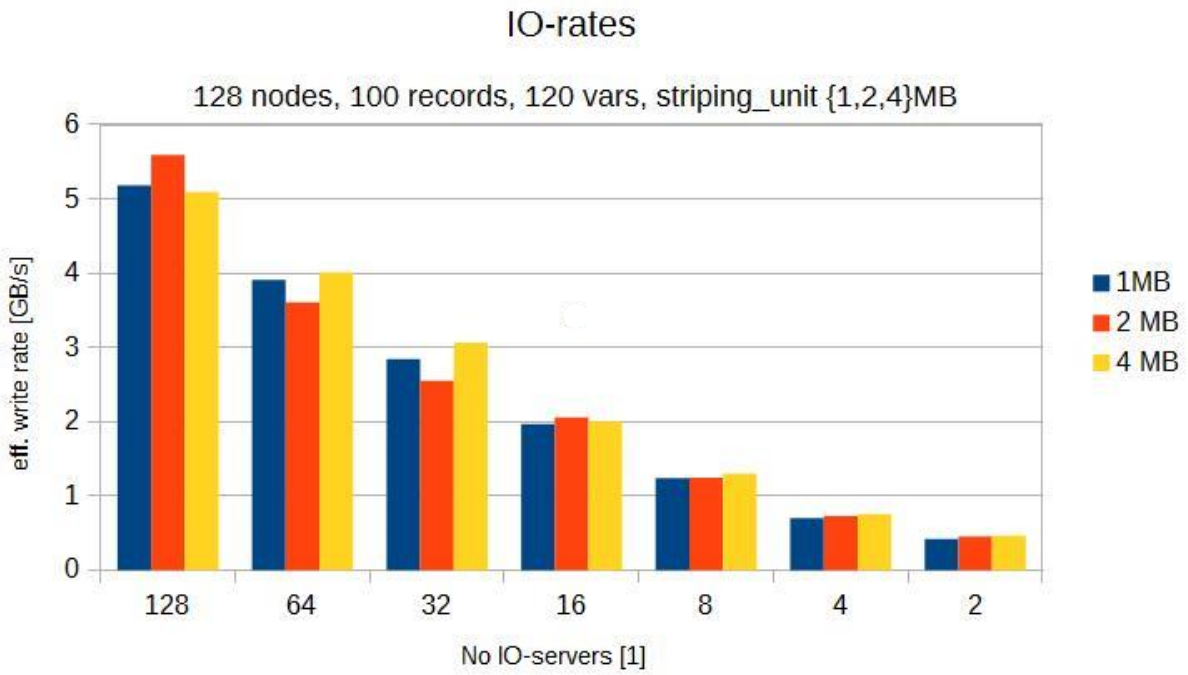
- 120 variables a 768x384x95 and 100 timesteps
- 256 nodes Mistral

IO-PES/node	Striping_unit	Striping_factor	Time [s]	Size[GB]	Rate [GB/s]
4	4194304	1	910,07	676	0,74
1	4194304	64	163,22	676	4,14
2	4194304	64	125,52	676	5,39
4	4194304	64	111,81	676	**6,05**

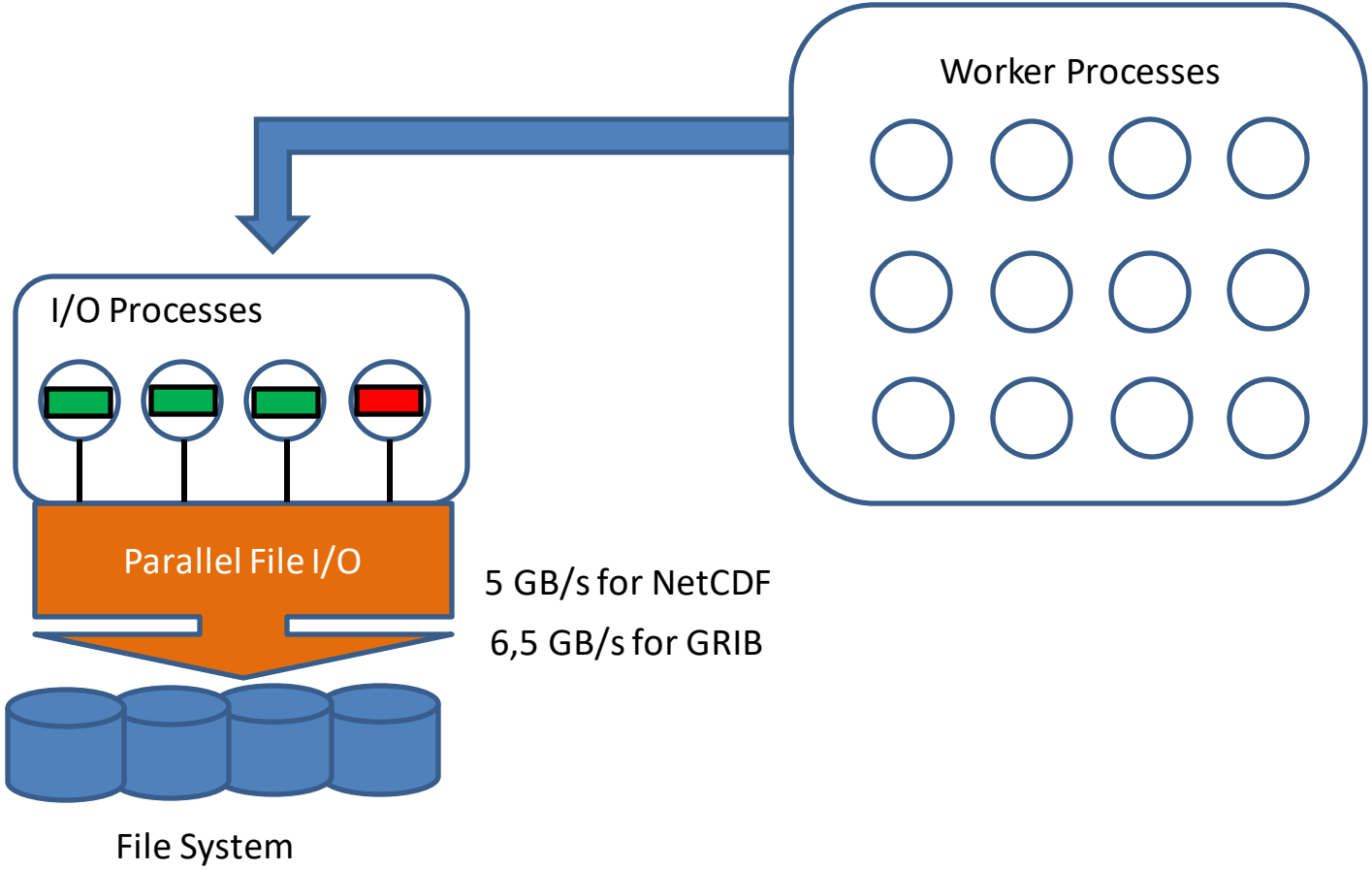
CDI-PIO NetCDF4(HDF5)



CDI-PIO NetCDF4(HDF5)



Parallel I/O in ICON



JUWELS Booster Cluster

R2B09 (5km) Experiment using 3 Output Files and 20 I/O Server

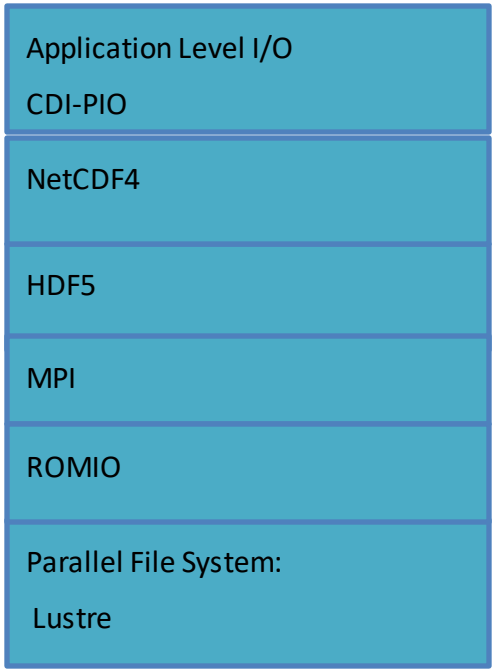
CDI-PIO output	duration[sec]	Percent in total	**Classic async. I/O**	duration[sec]	Percent in total
total	952.7	–	total	3183.2	–
wrt_output	262.6	27.6%	wrt_output	2484.6	78%

R2B09 (5km) Experiment using 16 Output Files und 20 I/O Server

CDI-PIO output	duration[sec]	Percent in total	**Classic async. I/O**	duration[sec]	Percent in total
total	1036.0	–	total	699.15	–
wrt_output	340.8	33%	wrt_output	4.79	0.69%

Tuning CDI-PIO: data traffic among IO-Server

Different types of Domain Decomposition affect the data traffic among IO-Server at different levels (CDI-PIO/YAXT, HDF5, MPI, ROMIO)

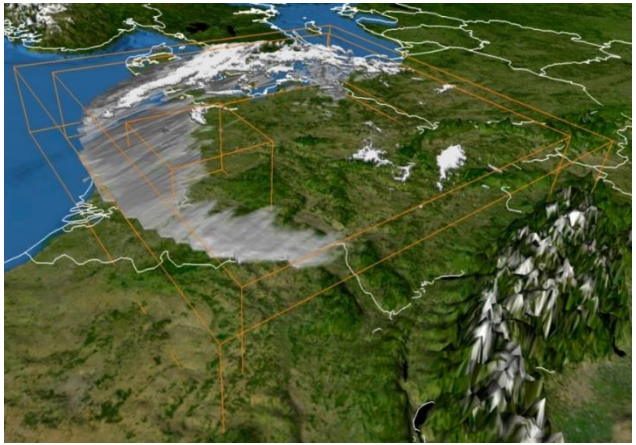


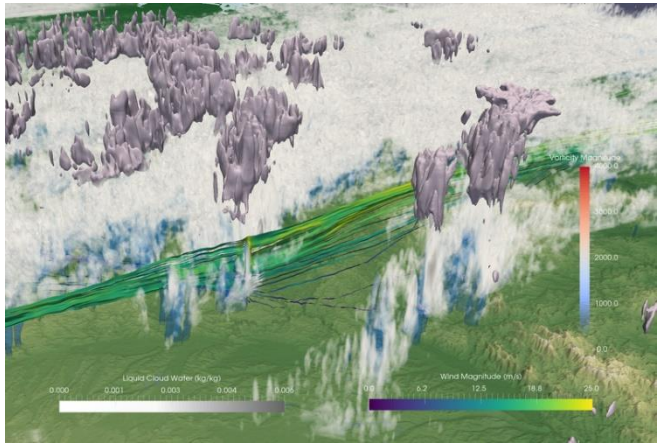
CDI generates with YAXT Z-decomposition

HDF5 uses decomposition in x,y,z-direction

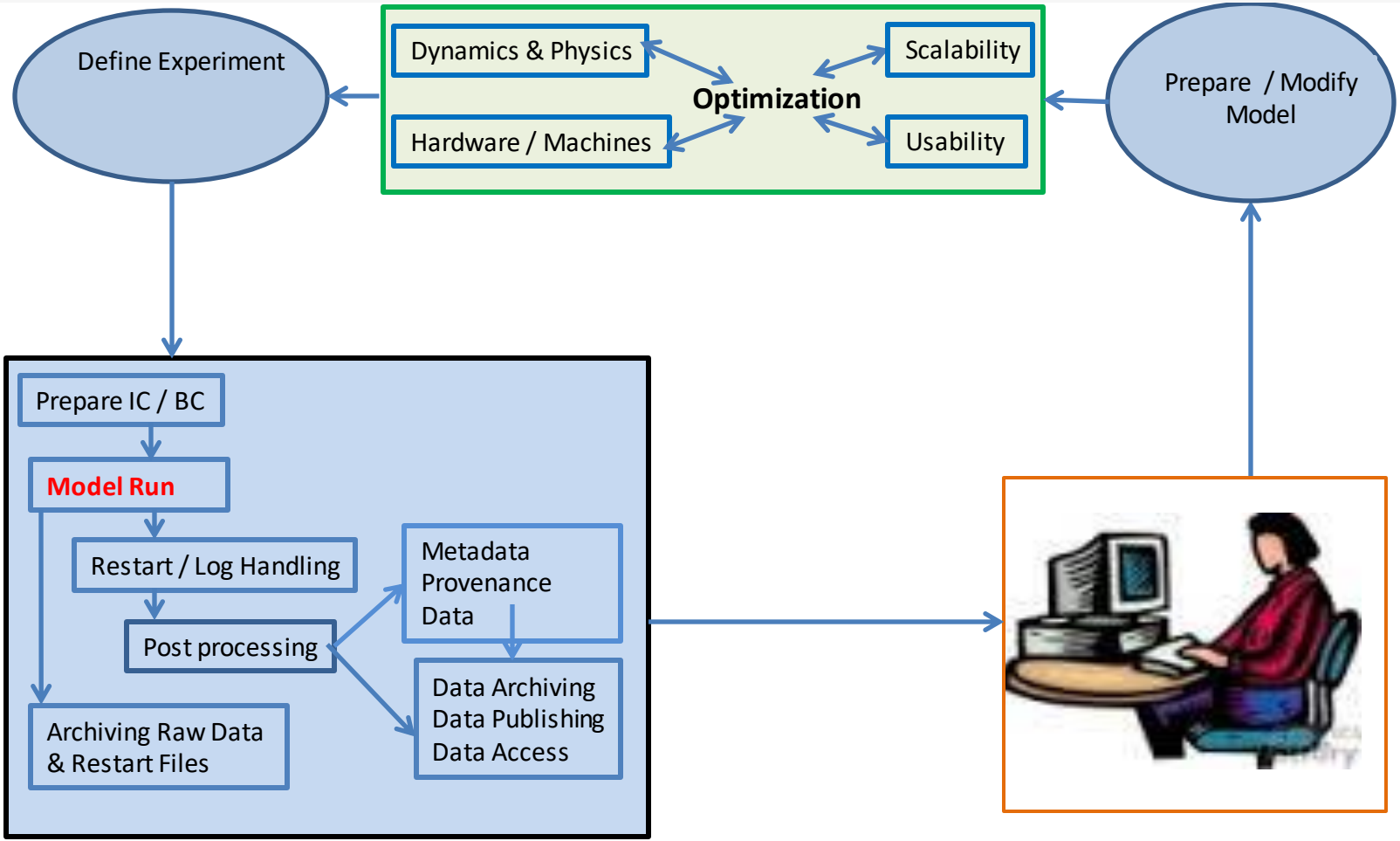
ROMIO might change the decomposition again for its own needs

Time to Solution

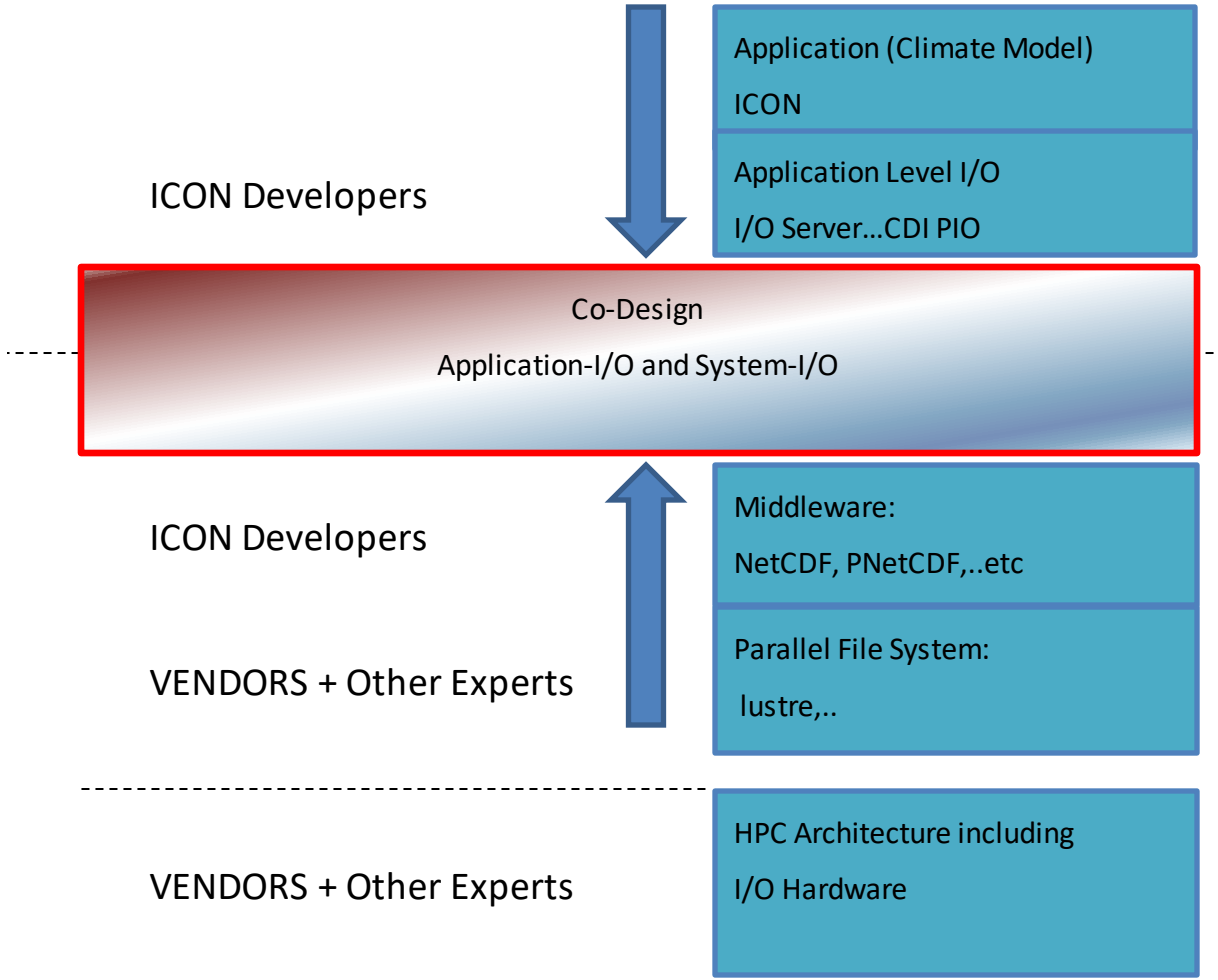




Time to Solution



Conclusion : Co-Design



Thank you for your attention