



WP4: I/O & Data Flow

F.Iannone,A.Funel, ENEA

Report Project : Deliverable 4.2 – 08/04/2020

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

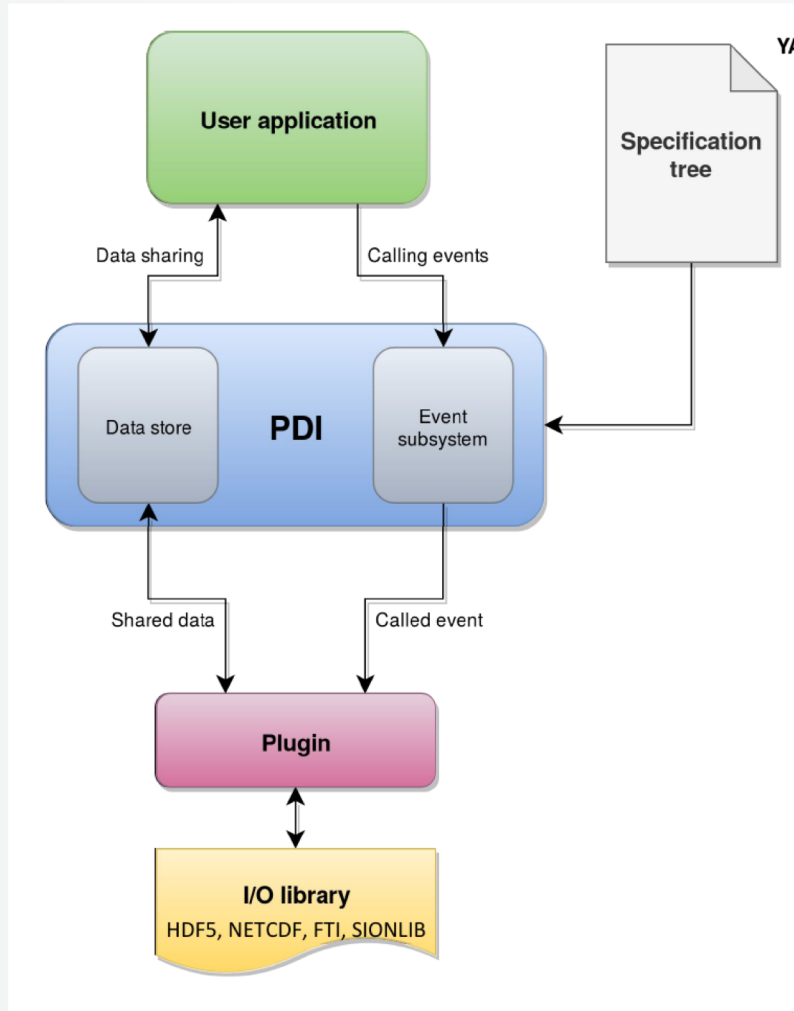


✓ ENEA activities:

- integrate PDI into the standard I/O benchmarking tool IOR waiting for GyselaX deployment (done)
- test performances of PDI integrated in IOR with POSIX interface (done)
- Gysela optimization:
 - Install Gysela code provided by CEA on CRESCO6 and Marconi SKL (done)
 - Test Gysela with old PDI integration developed in EoCoE I (done)
 - Test Gysela with new PDI integration (on going)

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

- ✓ PDI : Parallel Data Interface: API decoupling applications from IO concerns



YAML:

```
tree 1:
  array_1:
    - scalar_1
    - scalar_2
  array_2: [1,2,3]
tree_2: {subtree_1: scalar_1, subtree_2: scalar_2}
```

Data description:

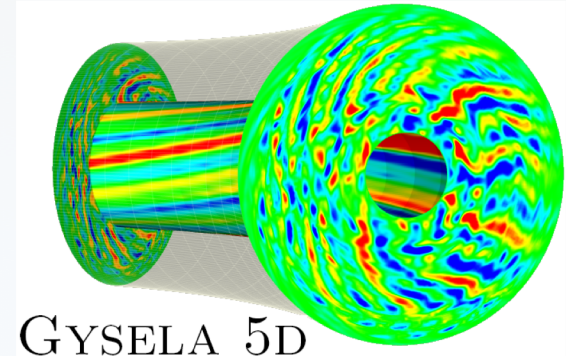
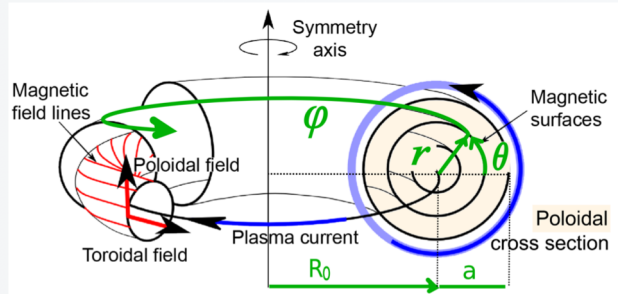
- scalar
- array
- record

Data types:

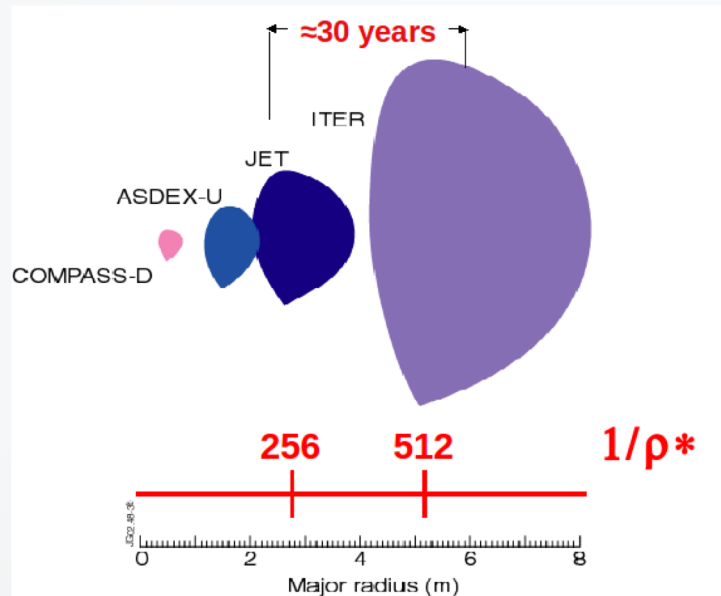
- C data types
- Fortran data types
- Custom

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

- ✓ GYSELA 5D: GYrokinetic SEmi-Lagrangian for plasma turbulence model in tokamak



$$nT\tau_E \geq 8 \times 10^{21} \text{ keV} \cdot \text{s} \cdot \text{m}^{-3}$$



$$\tau_E \propto (\rho_*)^{-2.8}$$

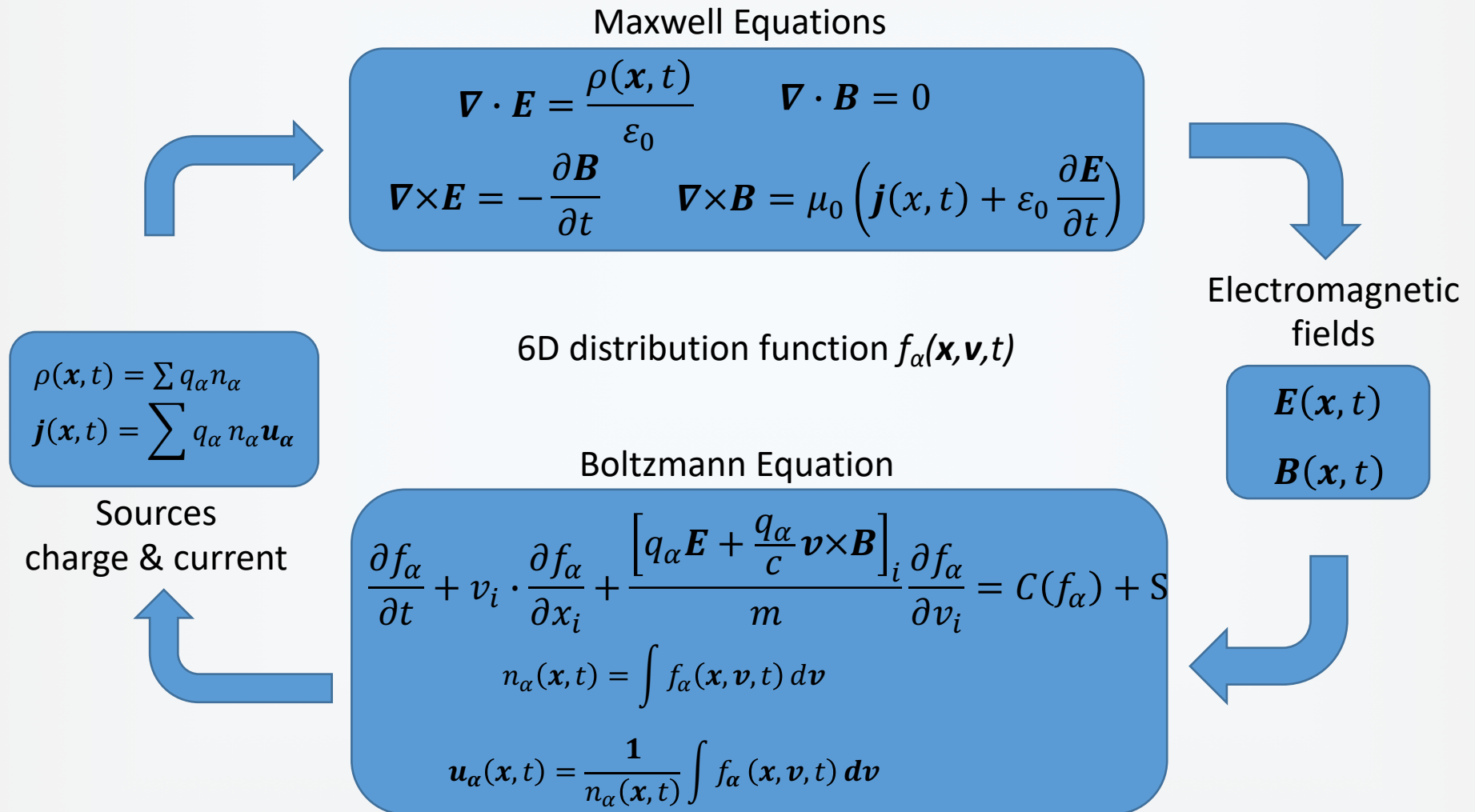
$$\rho_* = \frac{\rho_i}{a}$$

Ionic Gyroradius
nomalized to major radius

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

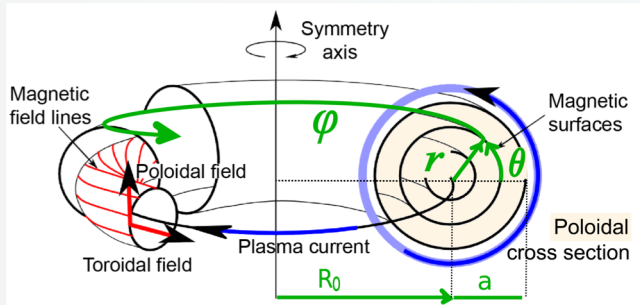


✓ *Kinetic model of plasma*

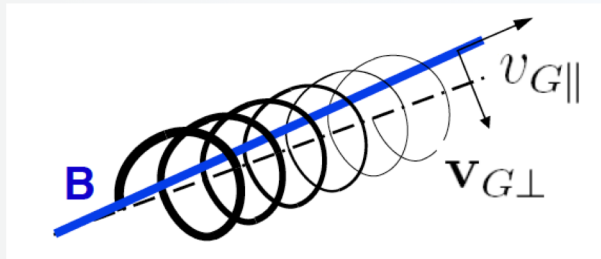


EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ *GyroKinetic model of plasma: from 6D to 5D*



$$f_{\alpha}(\mathbf{x}, \mathbf{v}, t) = f_{\alpha}(r, \theta, \varphi, v_{\parallel}, v_{\perp}, \phi, t)$$



plasma turbulence vs ion cyclotron frequency

$$\omega_{turb} \sim 10^5 s^{-1} \ll \omega_{ci} \sim 10^8 s^{-1}$$

Phase space reduction 6D to 5D: fast gyro-motion of ion is average out

Adiabatic invariant: magnetic momentum: $\mu = m_{\alpha} v_{\perp}^2 / (2B)$

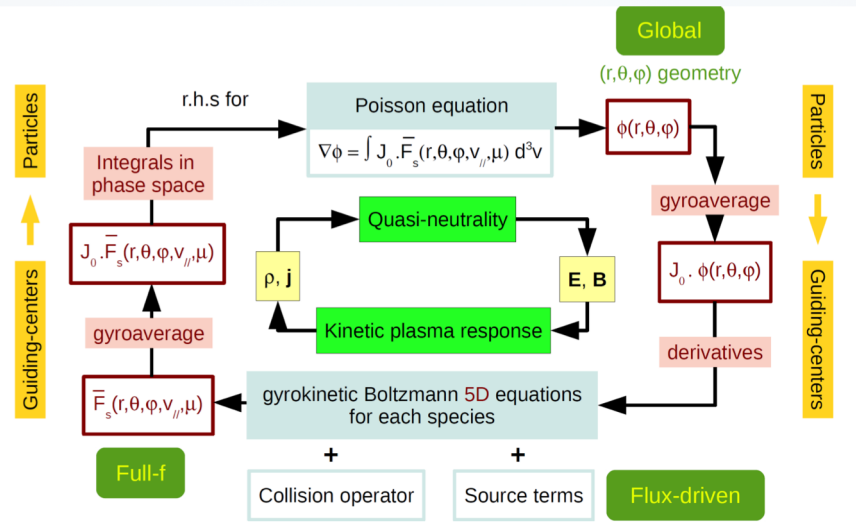
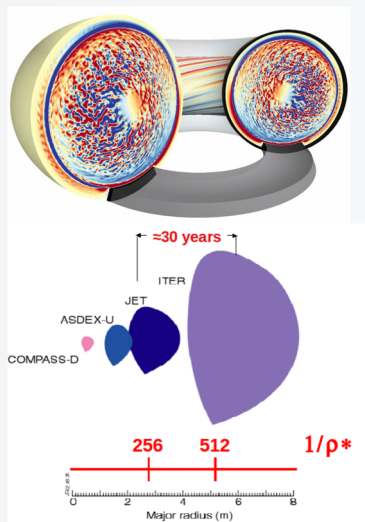
Velocity drifts of guiding-centres

GyroKinetic theory: 5D distribution function of guiding centres:

$$f_{\alpha}(r, \theta, \varphi, v_{\parallel}, \mu) \text{ with } \mu \text{ as parameter}$$

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

- ✓ *Gysela* code: developed at CEA-IRFM since 2001: Unique code based on a Semi-Lagrangian method (mix between PIC and Eulerian schemes)
 - *Global*: the entire tokamak space domain
 - *Full-f*: multi-scale physics in the space e time (simulate equilibrium and perturbations)
 - *Flux-driven*: heating power impacts on τ_E (included external additional power)
 - *Collisional*: simulate neoclassical and turbulent transports
 - *Kinetics electrons and multi-ion species*: impurity transport
 - *Circular magnetic configuration*: (EoCoE II: *Gysela-X* for X-point configuration)
 - *Electrostatic*: constant electromagnetic fields



$$\rho_*^{ITER} = \frac{\rho_i}{a} \sim 10^{-3}$$

Huge meshes for ITER

$$1024^3 \times 128 v_{\parallel} \times 16 \mu$$

Simulation time

$$\Delta t \sim 10^{-6} s \rightarrow \tau_E \sim 10 s$$

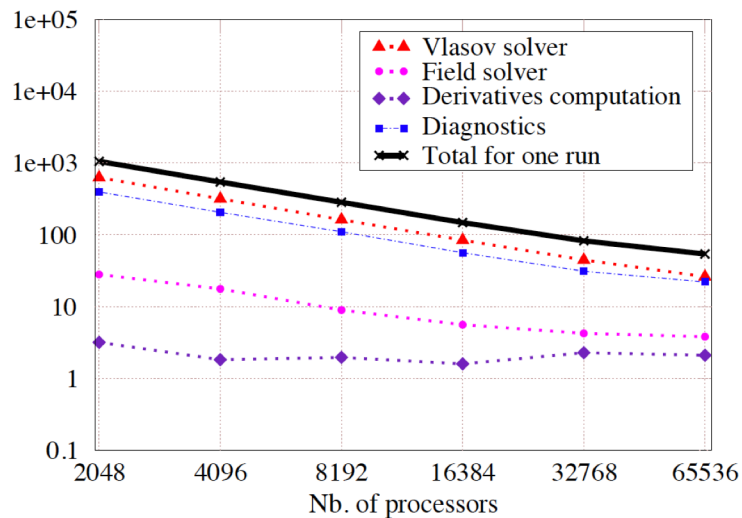
THE MONSTER

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

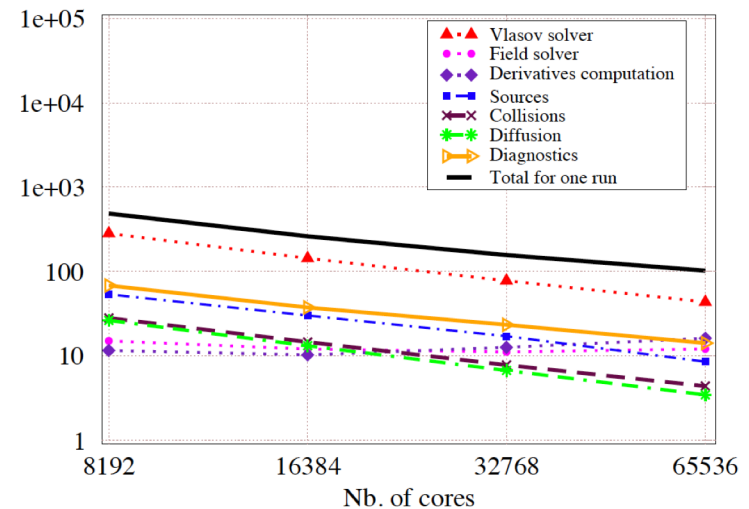
✓ Gysela benchmarks 2011

Strong scaling: $N_r = 512$, $N_\theta = 512$, $N_\varphi = 128$, $N_{v||} = 128$

$M_\mu = 32$, main data=1 TiB
Execution time, one run (Curie)



$M_\mu = 16$, main data=512 GiB
Execution time, one run (Turing)

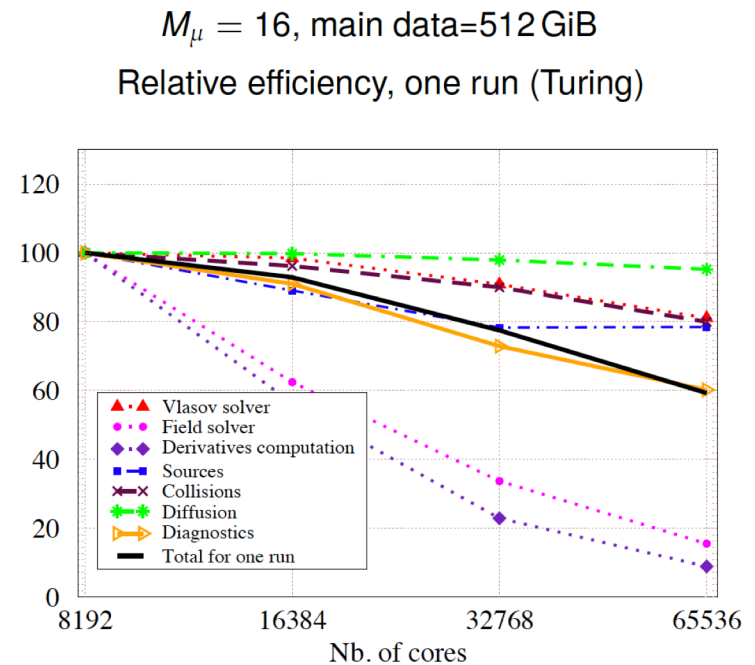
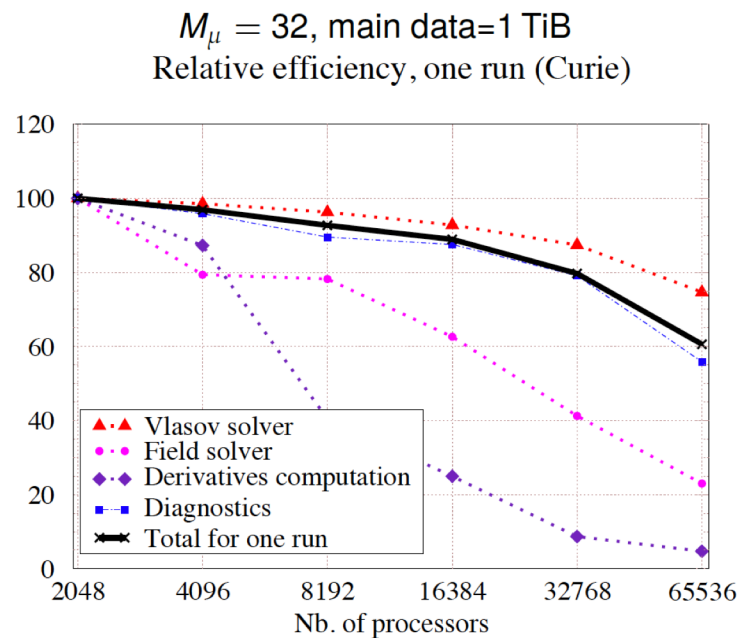


► time dominated by Vlasov solver

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2013

Strong scaling: $N_r = 512$, $N_\theta = 512$, $N_\varphi = 128$, $N_{v||} = 128$



≈ 60% efficiency at 64 ki core on both

► time dominated by Vlasov solver

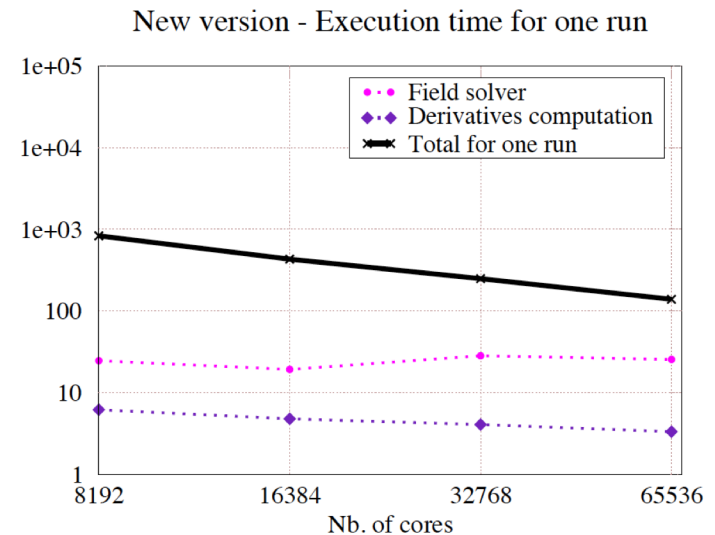
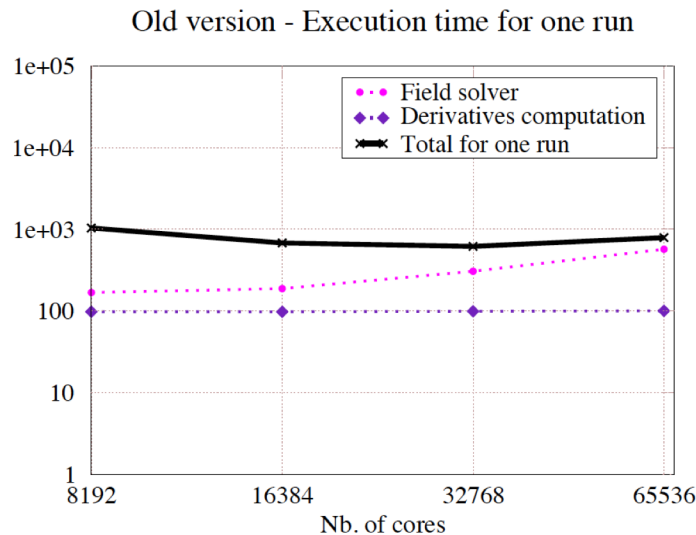
► scaling bottleneck: Poisson solver

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2011

Poisson solver communication problem

- ▶ Reduce communications to a minimum
 - ▶ Specifically tailored comm. scheme vs. broadcast
 - ▶ Prefer redundant computations



Strong scaling (Jaguar @ Oak Ridge 2011)

main data = 1 TB: $N_r = 512$, $N_\theta = 1024$, $N_\varphi = 128$, $N_{v_{||}} = 64$, $N_\mu = 32$

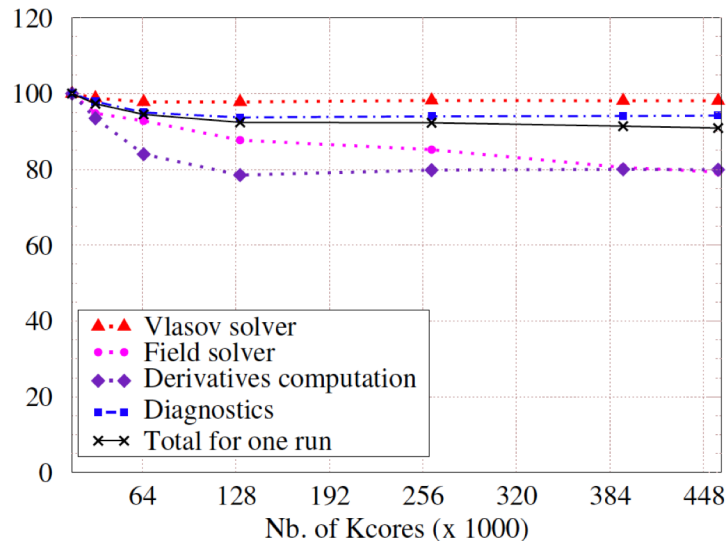
EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2013

Weak scaling: $N_r = 512$, $N_\theta = 1024$, $N_\varphi = 128$, $N_{v||} = 128$, $N_\mu = [2 - 56]$

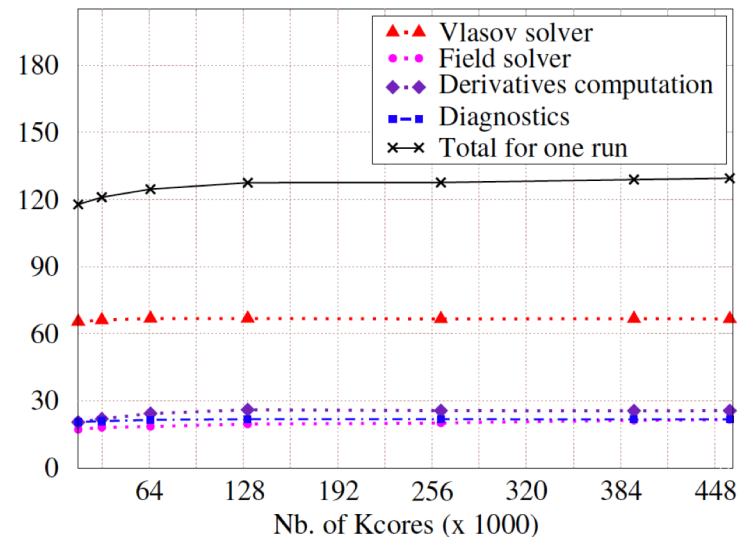
- ▶ 16 → 448 ki cores, 65 → 1 835 k threads (**Full Juqueen !**)
- ▶ No comm compression included

Relative efficiency, one run (Weak scaling - Juqueen)



- ▶ **91% efficiency at 458 752 cores**

Execution time, one Gysela (Weak Scaling - Juqueen)

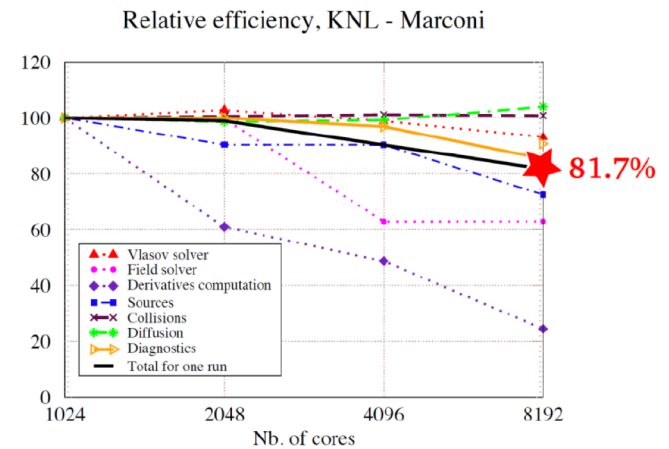
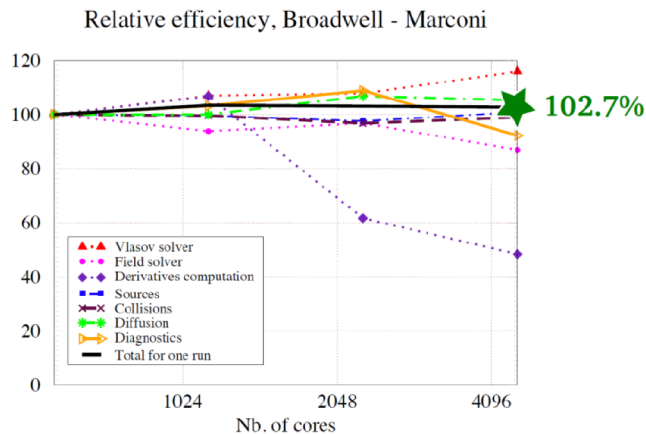


- ▶ Time still dominated by Vlasov
- ▶ Current bottleneck: still Poisson

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

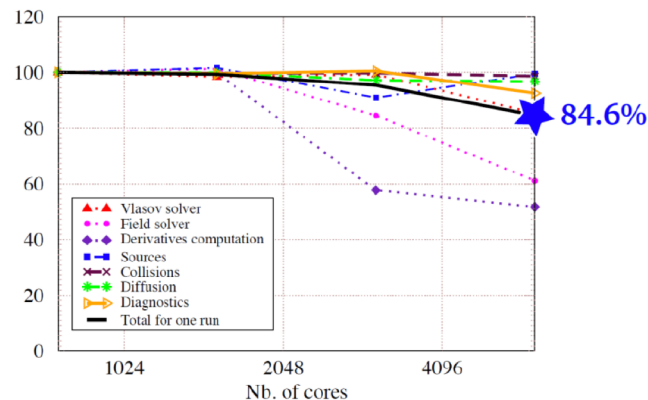
✓ Gysela benchmarks 2018

- From 16 up to 128 nodes. Domain size $512 \times 256 \times 128 \times 128 \times 16$ - Medium case



- *Good efficiencies*
- *Pb to scale:*
 - Field solver
 - Derivatives comp.

Relative efficiency, Skylake - Marconi



G. LATU & AL.

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2018

■ GYSELA is now ported on KNL machine

[EoCoE european Project + CVT GENCI
+ HLST IPP Garching + Atos-France]

Benchmark on one node Broadwell / KNL / Skylake (Marconi machine)

Adding of vectorisation + Lagrange	Steps \ Hardware	Broadwell	KNL	Skylake
	advec1D in vpar	12.7 (-78%)	12.2 (-85%)	6.4 (-86%)
	advec2D (r,theta)	16.3 (-60%)	24.7 (-43%)	8.9 (-70%)
	comm. transpose	31.2 (-25%)	12.9 (-48%)	15.5 (-53%)
	heat source	5.6 (-50%)	7.9 (-64%)	3.2 (-60%)
	...			
	Total	139 (-45%)	124 (-55%)	86 (-58%)

Table: Breakdown of timing (in s) for a small run. In parentheses, improvement compared to **initial version**.

[G. Latu, 2017]

■ CPU time on one KNL node comparable with one Broadwell node

■ Improvement of vectorisation essential for KNL

→ positive impact on Broadwell and Skylake machine

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2021

Disponibile su CRESCO6 come modulo (Intel19-OpenMPI)

Benchmark eseguiti con Intel19-IMPI sia su CRESCO6 che su Marconi SKL

512 x 1024 x 128 x 64 x 48 output size 1.5 TiB	192 nodes - 384 MPI-9216 cores		
	total time	total time (without init & diag)	time for restart file writing
IRENE-SKL	1129	888	17
CRESCO6 (/gporq3)	3587	1905	1098
CRESCO6 (/gporq2)	2735	1924	264
MARCONI SKL	2261	1342	290

512 x 1024 x 128 x 128 x 48 output size 3 TiB	192 nodes - 384 MPI-9216 cores		
	total time	total time (without init & diag)	time for restart file writing
IRENE-SKL	1707	1329	26
CRESCO6 (/gporq3)	5120	3331	1032
CRESCO6 (/gporq2)	4730	3336	613
MARCONI SKL	3261	2244	372

8 step run
Checkpointing
IO con HDF5

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2021

MARCONI SKL (512x1024x128x64x32) : 1 TiB					
node	core	parallel	total time [s]	total time (without init & diag) [s]	restart file writing [s]
128	6144	(nr=2,nteta=4,nmu=32)	332.44	111.41	18.96
256	12288	(nr=4,nteta=4,nmu=32)	305.64	77.8	27.97

CRESCO6 (512x1024x128x64x32) : 1 TiB					
node	core	parallel	total time [s]	total time (without init & diag) [s]	restart file writing [s]
128	6144	(nr=2,nteta=4,nmu=32)	646	140	206
256	12288	(nr=4,nteta=4,nmu=32)	410.31	78	90

MARCONI SKL (512x1024x128x64x48) 1.5 TiB					
node	core	parallel	total time [s]	total time (without init & diag) [s]	restart file writing [s]
96	4608	(nr=2,nteta=2,nmu=48)	NaN	NaN	NaN
192	9216	(nr=2,nteta=4,nmu=48)	406.75	120.5	35.3

CRESCO6 (512x1024x128x64x48) 1.5 TiB					
node	core	parallel	total time [s]	total time (without init & diag) [s]	restart file writing [s]
96	4608	(nr=2,nteta=2,nmu=48)	NaN	NaN	NaN
192	9216	(nr=2,nteta=4,nmu=48)	843.11	156	296

1 step run
Checkpointing
IO con HDF5

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2021

Benchmarks: Intel19-IMPI in CRESCO6 and Marconi SKL

Darshan application level IO characterization tool

512 x 1024 x 128 x 128 x 48 output size 3 TiB	192 nodes - 384 MPI-9216 cores			
	total time	total time (without init & diag)	time for restart file writing	time for initialization
IRENE-SKL	1707	1329	26	115
CRESCO6 (/gporq2)	2736	1924	265	299
MARCONI SKL (no DARSHAN)	2362	1667	53	176
MARCONI SKL 9441477	2476	1756	51	186
MARCONI SKL 9436808	2494	1749	51	199
MARCONI SKL 9431143	2503	1772	52	192
MARCONI SKL 9443364	2629	1685	51	427
MARCONI SKL 9443284	2421	1663	114	180
MARCONI SKL 9443526	2484	1659	183	179
MARCONI SKL 9430577	2549	1747	123	194
MARCONI SKL 9433585	2610	1770	156	193
MARCONI SKL 9433325	2644	1758	195	203
MARCONI SKL 9427542	2716	1758	95	367
MARCONI SKL (no DARSHAN)	3261	2244	372	183
MARCONI SKL (no DARSHAN)	3311	2273	369	203

8 step run
Checkpointing
IO con HDF5

Fast IO

Slow IO

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2021

DARSHAN application level I/O characterization

✓ *DARSHAN I/O module of 384 MPI tasks analyzed for 3TiB test case*

- *Darshan Diagnostic modules intercepted*

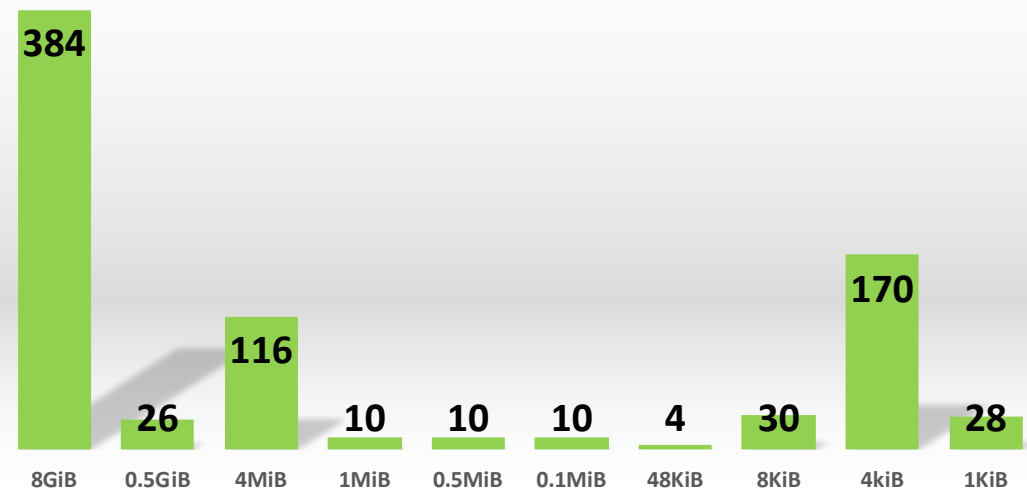
H5F File object

H5D Dataset object (HDF5 API: H5Dread/HD5write)

POSIX object (C IO functions: read/write)

H5Fcreate (H5Fopen)	create (open) File
H5ScreateSimple/H5Screate	create fileSpace(not used)
H5Dcreate (H5Dopen)	create (open) Dataset
H5Sselect_hyperslab	select subsections of data
H5Dread, H5Dwrite	access Dataset (posix IO)
H5Dclose	close Dataset
H5Sclose	close fileSpace
H5Fclose	close File

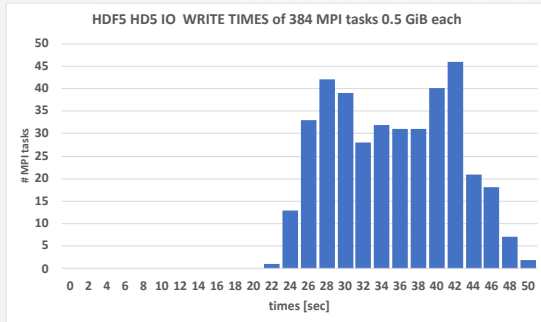
I/O files size distribution in Gysela 3TB use case



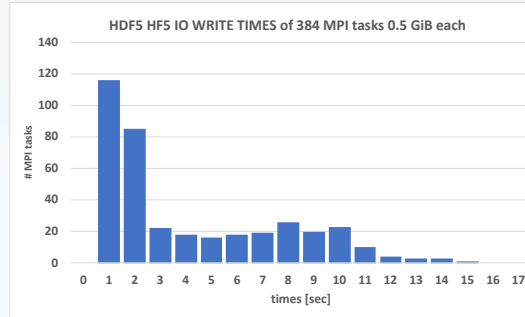
EoCoE²: Energy Oriented Center of Excellence : toward exascale for energy

✓ Gysela benchmarks 2021 DARSHAN application level I/O characterization

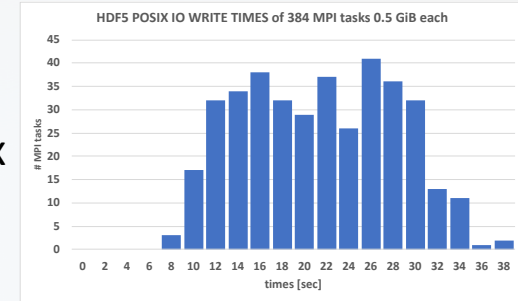
H5D



H5F

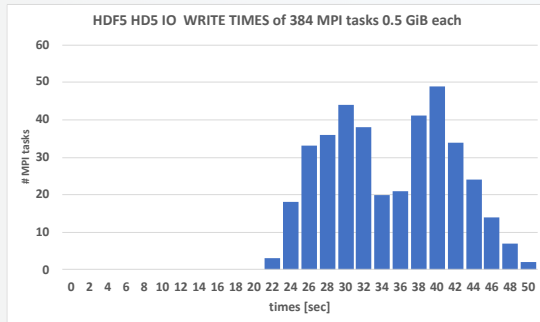


Posix

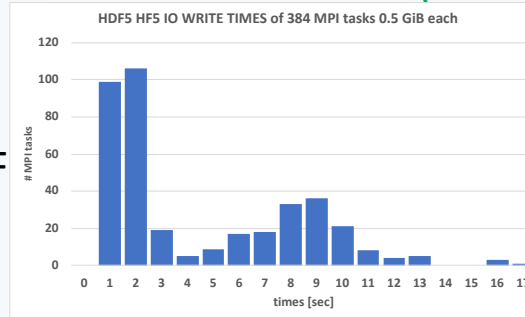


9441477 (2476 sec – 51 sec)

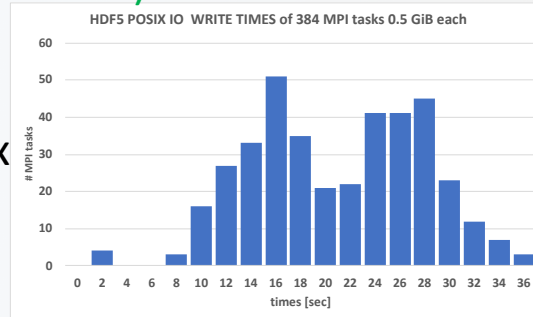
H5D



H5F

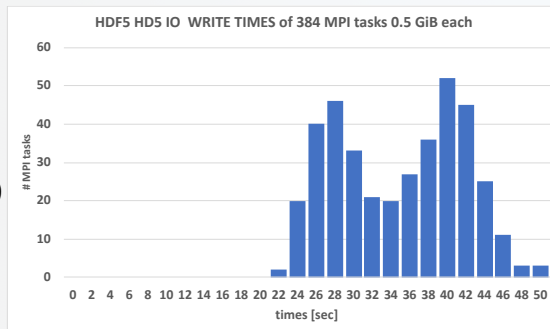


Posix

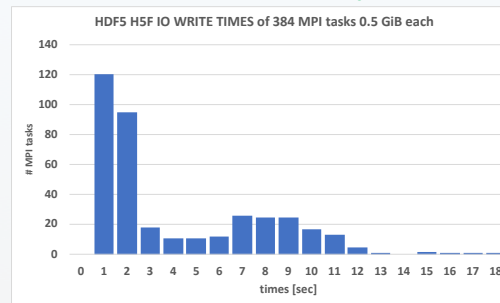


9431143 (2503 sec – 52 sec)

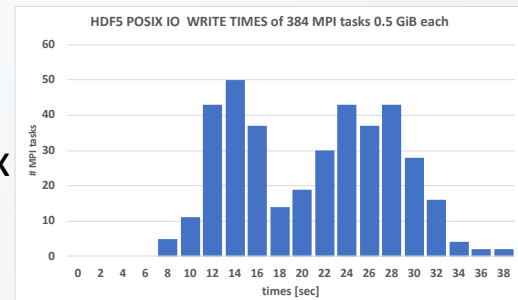
H5D



H5F



Posix

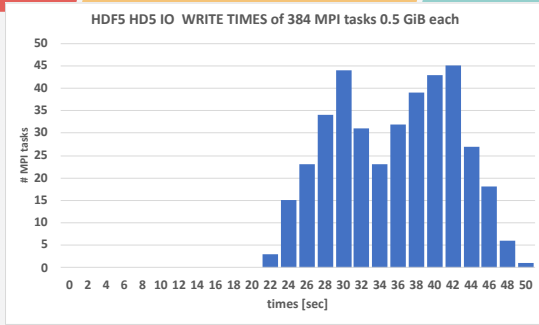


9436808 (2494 sec – 51 sec)

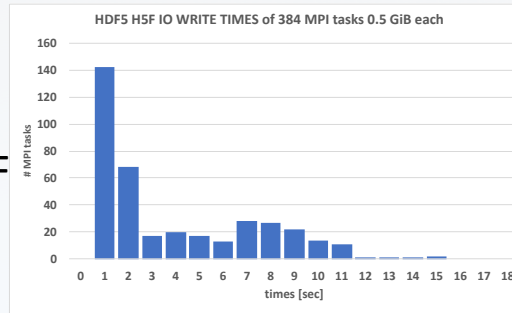
EoCoE²: Energy Oriented Center of Excellence : toward exascale for energy

✓ Gysela benchmarks 2021 DARSHAN application level I/O characterization

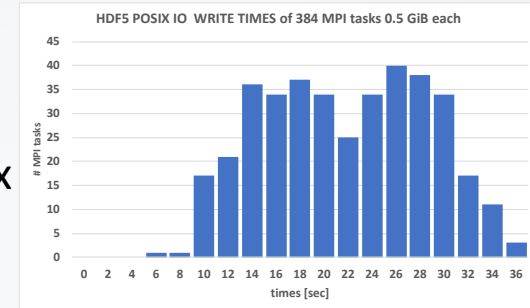
H5D



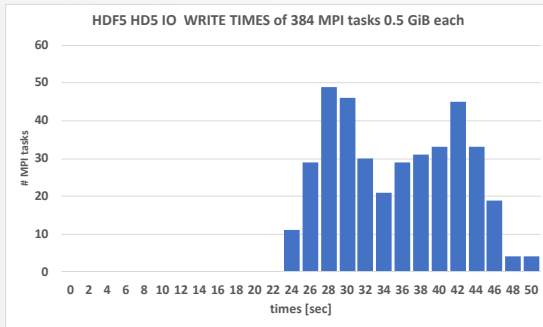
H5F



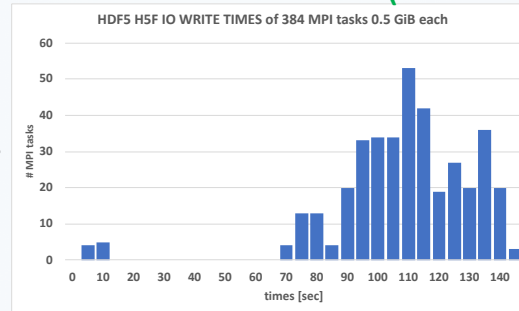
Posix



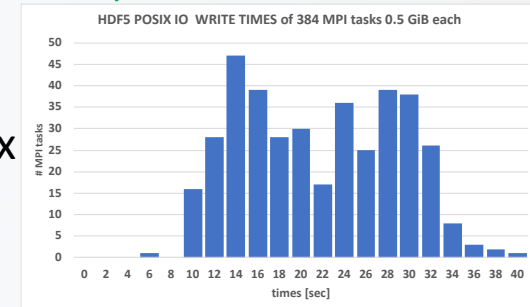
H5D



H5F



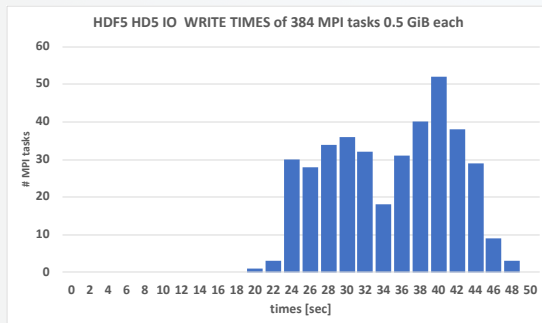
Posix



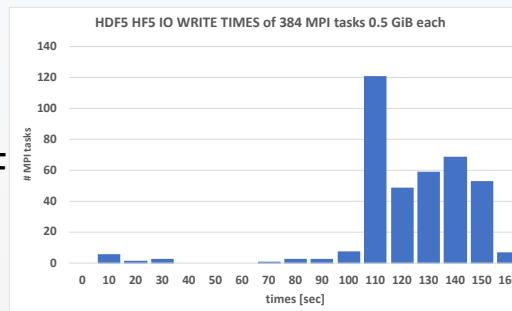
9443364 (2629 sec – 51 sec)

9443526 (2484 sec – 183 sec)

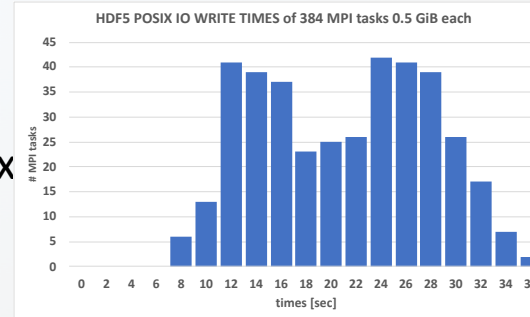
H5D



H5F



Posix

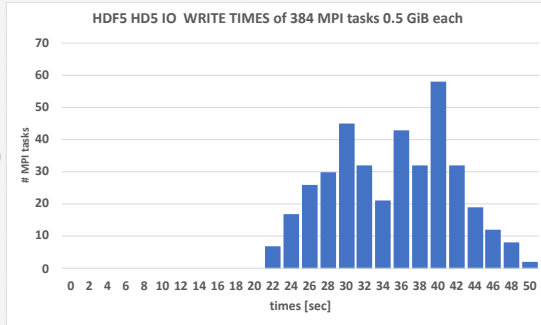


9433325 (2644 sec – 195 sec)

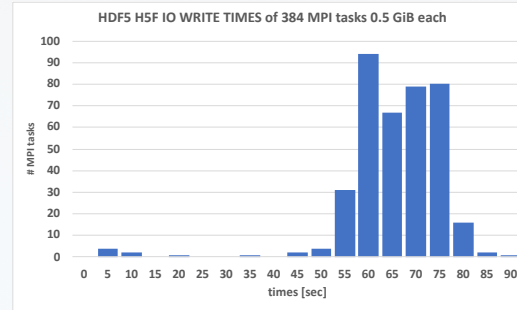
EoCoE²: Energy Oriented Center of Excellence : toward exascale for energy

✓ Gysela benchmarks 2021 DARSHAN application level I/O characterization

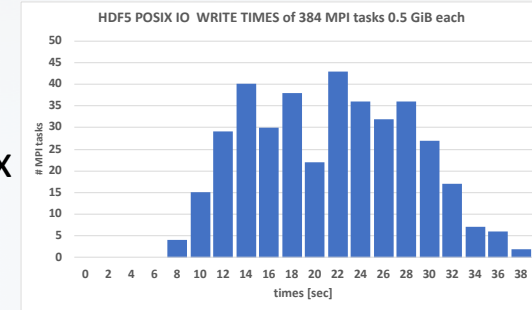
H5D



H5F

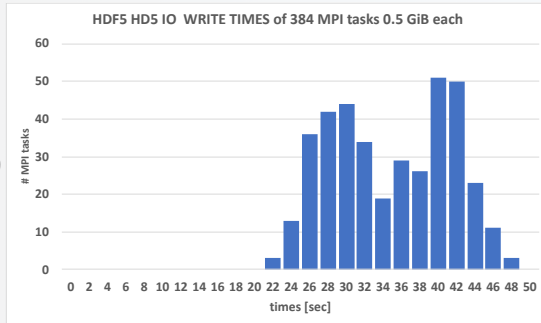


Posix

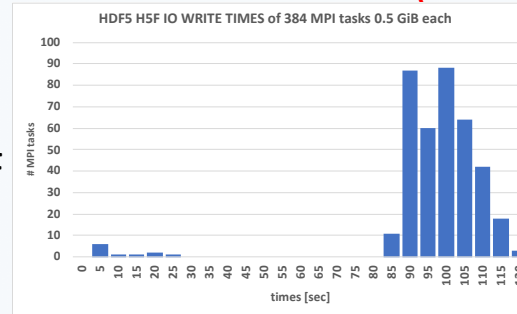


9430577 (2459 sec – 123 sec)

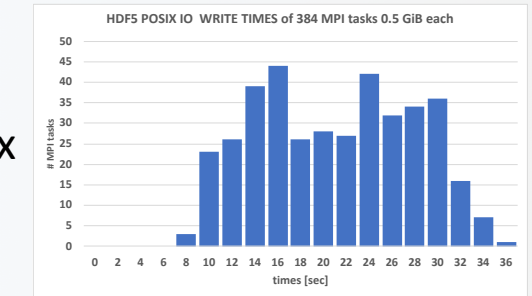
H5D



H5F

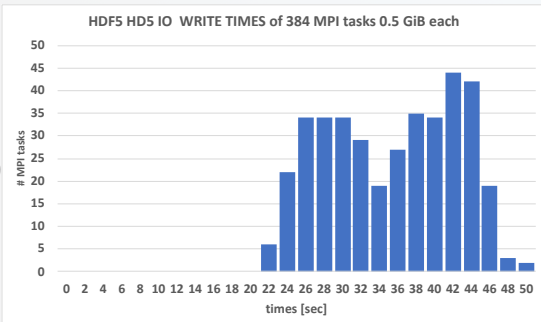


Posix

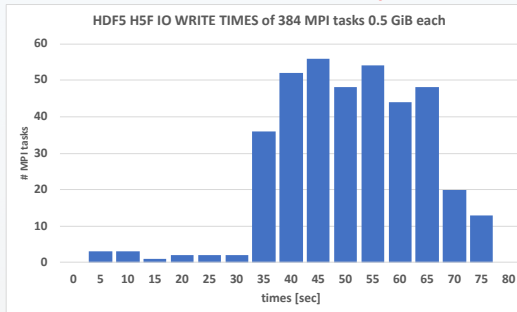


9433585 (2610 sec – 156 sec)

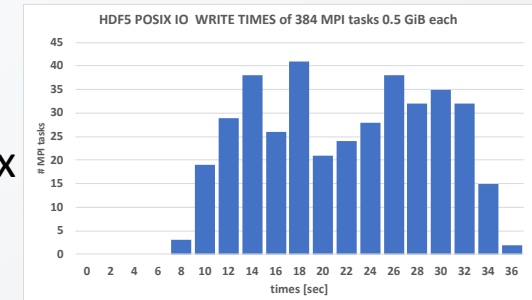
H5D



H5F



Posix

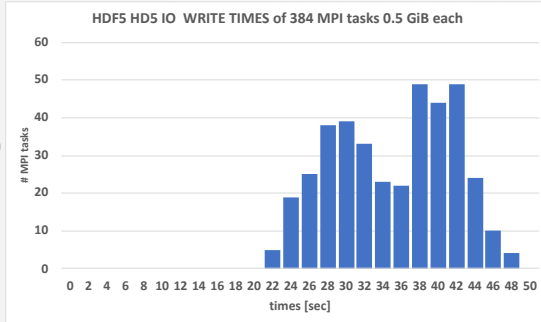


9443284 (2421 sec – 114 sec)

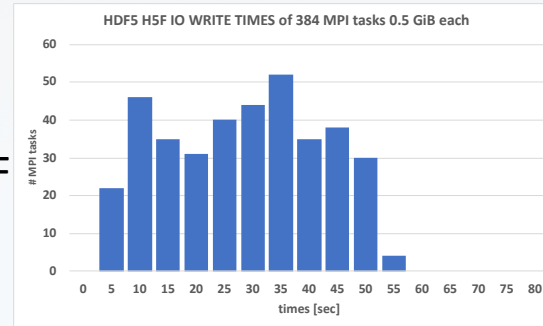
EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ Gysela benchmarks 2021 *DARSHAN application level I/O characterization*

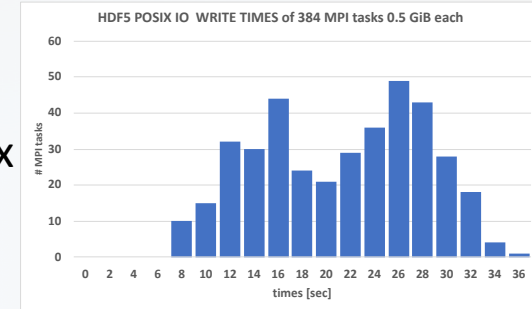
H5D



H5F



Posix



9427542 (2716 sec – 95 sec)

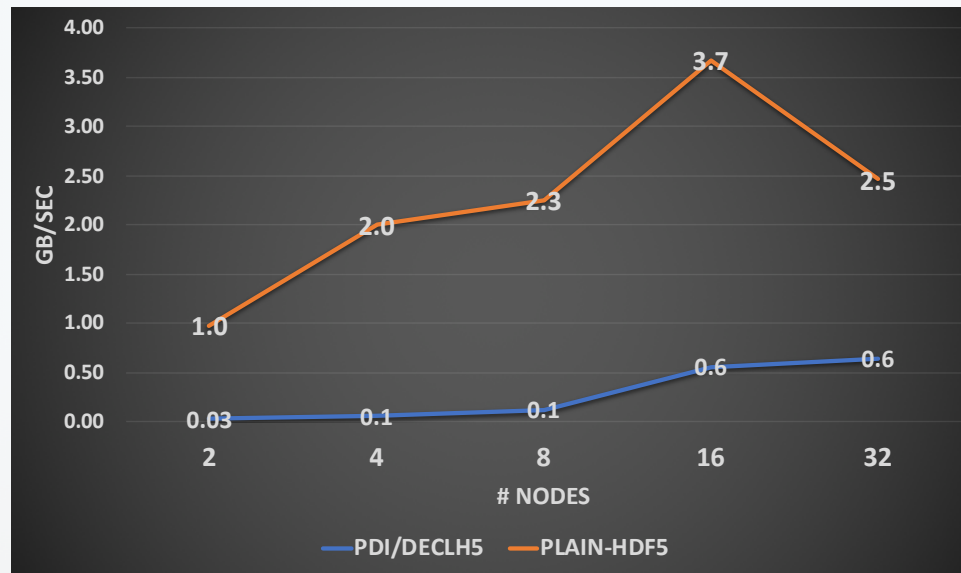
Conclusions:

- Posix function within 40 sec. for all cases (fast and slow IO)
- H5D module within 50 sec. for all cases
- H5F module is blocking for slow cases (maybe H5Dwrite is blocking)

EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

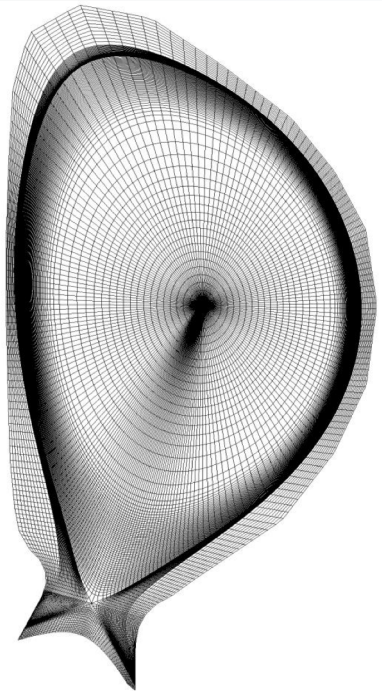
- ✓ *Gysela* benchmarks old release 30 (EoCoE I) with PDI integration (use case 4D) waiting for new PDI integration release

PLAIN-HDF5					PDI/DECLH5				
(512 x 1024 x 128 x128)					(512 x 1024 x 128 x128)				
# nodes	#cores	GB	sec	GB/sec	# nodes	#cores	GB	sec	GB/sec
2	96	66	67.7	1.0	2	96	66	2476	0.03
4	192	131	65.6	2.0	4	192	131	2352	0.1
8	384	131	58.2	2.3	8	384	131	1099	0.1
16	768	131	35.7	3.7	16	768	131	235.3	0.6
32	1536	261	106	2.5	32	1536	261	406	0.6



EoCoE²: *Energy Oriented Center of Excellence : toward exascale for energy*

✓ *Gysela-X: Development of a new code*



$$\tau_{E,Th}^{ELMy} \propto \tau_B \rho_*^{-0.83} \beta^{-0.50} v_*^{-0.10} M^{0.97} q^{-2.52} \epsilon^{-0.55} k^{2.72}$$

Elongation k and Triagularity δ impact on confinement time

Numerical Challenges

- MESHES:
 1. Mapped grid
 2. Aligned Mapped Grid
 3. Cartesian Grid
- Boundary Conditions
- Kinetic electrons and electromagnetic effects

Next step for ENEA in EOCoe II WP4

- New PDI integration in Gysela by CEA is ready
- Rebuild the latest Gysela version with PDI integration
- New benchmarks session
- Need manpower