SCIENTIFIC DATA MANAGEMENT PLAN

INFORMATION ABOUT THE PROJECT

Project name*: Global full-kinetic simulations with iPIC3D

Project objectives:

1. *provide first global full-pic simulations of a Mercury-like planetary magnetosphere*
2. *provide input for science operation of the BepiColombo mission*
3. *study plasma physics processes with feedback from small-scale (electrons) to large-scale (magnetosphere)*

Fundings:

1. ***ESA NPI program, OCA for funding the PhD***
2. ***CNES, UniPi, UFI for support missions and travels***
3. ***CINES, TGCC, Iscra for numerical resources***

Project coordinator and partners:

- ***Federico Lavorenti (PhD student)***
- *Pierre Henri (supervisor)*
- *Francesco Califano (supervisor)*
- *Johannes Benkhoff (ESA tutor)*

Contact:

- ***federico.lavorenti@oca.eu***
- ***federicolavorenti@gmail.com***
- ***pierre.henri@oca.eu***
- ***francesco.califano@unipi.it***

Project start date: 01/01/2021

Project duration: 3 years

DATA SETS DIRECTORIES NAMING RULES

Parent directory naming rules:

$PARENT = $WORKSPACE/<Initialization>/<Radius>/<IMF>-<boxsize>_<extra>/

- *$WORKSPACE is the path to the directory where your data are saved, e.g. $SCRATCH, $WORK etc.*
- *<Initialization> is the name of the initialization used to run the simulation, e.g. Mercury_SaeInit, Mercury_MarinerX, Mercury_BepiFB1 etc.*
- *<Radius> is a keyword defining the radius of the planet in the simulation from PR0 (small planet~2 di), to PR1 (medium planet~5 di) and PR2 (big planet~10 di)*

- *<IMF> direction of the IMF (can be Normal, minusBz, plus20deg, minus20deg etc.)*
- *<boxsize> dimension of the box, baseline box dimension is empty string and corresponds to 10x8x8 R. While newbox gets a factor 1.5, bigbox gets a factor 2.*
- *<extra> any extra information useful to characterize the run, e.g. more-ppc means larger number of particles per cell compared to baseline ppc=64, etc.*

Scratch Sub-Directories naming rules:
sub-directories in scratch correspond to different runs (run0, run1, run2…) that correspond to restarts of the same simulation. Moreover you can have runs named with a suffix _prec meaning that particles hitting the planet where collected in that run, and suffix _smth meaning that PrintCycle=1 for the fields. The raw data for each run are always stored in a directory called data, e.g. run0/data/.

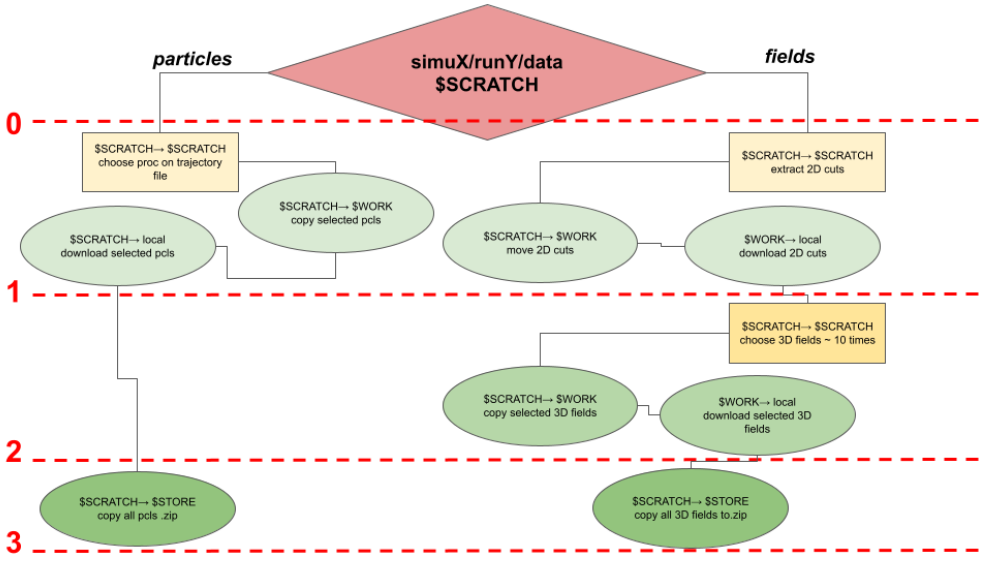
Work Sub-Directories naming rules:
Data process of level-1 shall be transferred from the scratch to the work space. The sub-directories in work are organized as follows:

- *texts contains all .txt files useful to interpret the simulation (metadata file SimulationData.txt, computational time, memory space etc.)*
- *images contains multiple plots showing quick-look of simulation output, computational time plots etc.*
- *data1 contains 2D cuts for all times of the fields (E,B,J,V,rho,T etc.)*
- *data2 contains 3D fields data at selected times (usually every 10 PrintCycle)*
- *dataP contains particles data for all restarts (proc0.hdf, proc1.hdf etc.) selected along given trajectory in the simulation box.*

DATA SETS CLASSIFICATION

List of data sets:

- ***level-0:** raw data, direct output of the simulations*
- ***level-1:** first processed data products, easy analysis to make data more user-friendly*
- ***level-2:** second processed data products, more refined analysis on level-1 data*



DATA SETS LEVEL-0

Data products:
*each simulation run stored in the scratch will have: files RUN_job.out, RUN_job.err with the output of the slurm job program, a file for slurm job submission named job_ipic3d.slurm, the executable iPIC3D, and the input file which usually is something with extension *.inp. On top of these, you have the raw data stored in the /data sub-directory organized as follows:*

- *<name>_B_<cycle>.vtk magnetic field 3D vector*
- *<name>_E_<cycle>.vtk electric field 3D vector*
- *<name>_Je_<cycle>.vtk total electron current 3D vector field*
- *<name>_Ji_<cycle>.vtk total ion current 3D vector field*
- *<name>_P<index><species>_<cycle>.vtk 2nd order moment of the distribution function (times the charge) of given species, 3D scalar field*
- *<name>_rho<species>_<cycle>.vtk density of given species, 3D scalar field*
- *SimulationData.txt metadata of the simulation*
- *ConservedQuantities.txt time evolution of scalar fields averaged over the box*
- *restart<nproc>.hdf files with the particles used for restart (serial printing)*
- *settings.hdf more metadata of the simulation (used for restart)*
- *RemovedParticles.txt list of particles hitting the surface of the planet and removed from the simulation box (produced in run*_prec using old type of internal BC)*

- *NB: usually <name>=Dipole3D, <species>=e0,i1 (can be longer for some specific simulations), and <index>=XX,XY,XZ,YY,YZ,ZZ.*

Purpose and relation to the objectives of the project:
These data are not intended to be totally stored on the long term given their very high storage needs. The procedure is to leave this data in the scratch after the simulation is completed, then via level-1 and level-2 processing these data are transformed in usable products and transferred to the work space. For long-term usage we keep one run (the best one at steady state) saved in compressed format (.zip) in the store workspace (particles and fields).*

Data types:
the fields are stored using parallel vtk writing method. The particles uses serial HDF5. The metadata and the simulation output and input files are written in text formats.

Data production methods:
These level-0 data are produced using the code iPIC3D (https://github.com/flavorenti/iPIC3D-OCA). The input and job file are stored in the level-0 directory as well.

Storage facility:
scratch space on national french super-computing facilities (usually TGCC/Irene) with usual quota threshold at 100 Tb (http://www-hpc.cea.fr/fr/complexe/tgcc-systeme-stockage.htm). One run is later transferred to the store space in compressed format.

*Expected size of the data**:***
For a typical high-quality simulation (4^3 ppc, 830^3 cells, 2 species) we can expect a total data volume of the order of 20-30 Tb. Most of this space is due to particles saved in the files

restart<nproc>.hdf that makes up around 5 Tb for each run (usually we have around 4-6 runs per simulation). The 3D fields occupy around 500 Gb per run, thus a total of 2-3 Tb.

General data policy
These data are not shared among the scientific community. They remain internally stored on the cluster where the simulations are run. Only people with an account on the machine and direct project collaborators can access these data.

Data property
Level-0 data belong to the numerical project team. The team is responsible for sharing and using these data.

DATA SETS LEVEL-1

Data products:
level-1 data products are produced from post-processing of the level-0 data and transferred to the work space sub-directories. The files are named as follows:

- *Texts:*
  - *computational_time.txt computational time of all the runs together, times taken from RUN_job.out in level-0 data, times are per cycle and split according to 6 different parts of the code (communication and computation for fields, particles and moments)*
  - *memory.txt output of command "du -h <file>" where <file> indicates (I) a 3D vector field, (ii) a 3D scalar field and (iii) one of the restart<nproc>.hdf files (using level-0 data).*
  - *output_dataP_<trajectory>.txt list of MPI processors containing the trajectory <trajectory> (if <trajectory> is empty string means that all trajectories are considered). This file is used to copy particles files from level-0 to level-1 in sub-directory dataP.*
  - *RUN_to2D_<nrun>.out output of the 2D cuts python routines launched using slurm job scheduler. Routines acting on level-0 data to create level-1 data stored in data1 sub-directory.*
  - *RUN_TnV.out output of python routine on the 2D cuts converting P,rho,J to T,n,V. Level-1 to level-1 data, all stored in data1.*
  - *RemovedParticles.txt (copy of level-0 data)*
  - *SimulationData.txt (copy of level-0 data)*
  - *ConservedQuantities<nrun>.txt (copy of level-0 data)*

- *Images:*
  - *plot_scalars.png plot of texts/ConservedQuantities<nrun>.txt*
  - *comp_time_percycle-plot.png plot of texts/computational_time.txt*
  - *comp_time_cumulative-plot.png cumulative plot texts/computational_time.txt*
  - *3d_plot_patches_<trajectory>.png 3d plot of texts/output_dataP_<trajectory>.txt*
  - *plot_2Dcuts_nBJ_<cycle>.png plot 2D cuts n (left), B (central), J (right) (6 plots, top dipole XZ plane, and bottom equatorial XY plane) per each cycle, data stored in data1 sub-directory*
  - *plot_2Dcuts_P<species>_<cycle>.png plot 2D cuts PXX (left), PYY (central), PZZ (right) (same format as nBJ), data stored in data1*

- *plot_2Dcuts_T<species>_<cycle>.png (same format as P), data stored in data1*

- *NB: usually <trajectory> = Bepi, Mariner (more can be added)*

- *data1:*
  - *<name>_B<cut>_<cycle>.vtk magnetic field 2D vector*
  - *<name>_E<cut>_<cycle>.vtk electric field 2D vector*
  - *<name>_Je<cut>_<cycle>.vtk total electron current 2D vector field*
  - *<name>_Ve<cut>_<cycle>.vtk total electron velocity 2D vector field*
  - *<name>_Ji<cut>_<cycle>.vtk total ion current 2D vector field*
  - *<name>_Vi<cut>_<cycle>.vtk total ion velocity 2D vector field*
  - *<name>_P<index><species><cut>_<cycle>.vtk 2nd order moment of the distribution function (times the charge) of given species, 2D scalar field*
  - *<name>_T<index><species><cut>_<cycle>.vtk total temperature (centered 2nd order moment of d.f.) of given species, 2D scalar field*
  - *<name>_rho<species><cut>_<cycle>.vtk density of given species, 2D scalar field*
  - *SimulationData.txt (copy of level-0 data)*

  - *NB: usually <cut>=dp, eq, tl (meaning dipolar, equatorial and tail cut respectively)*

- *data2:*
  - *<name>_*_<PrintCycle*10>.vtk (copy of level-0 data)*
  - *SimulationData.txt (copy of level-0 data)*

- *dataP:*
  - *restart<nproc>.hdf chosen particles files, <nproc> corresponds to list of numbers written in texts/output_dataP_<trajectory>.txt, these restart files are then used to extract the good particles written in text file output_good-pcls*.txt*
  - *SimulationData.txt (copy of level-0 data)*

Purpose and relation to the objectives of the project:
These data have multiple scopes: 1) they are used to check the quality of the simulation and its performances, 2) they are also intended to be used to carry out a first scientific analysis given the completeness of the data-set (all basic physical quantities in 2D at high frequency and in 3D at lower frequency), and 3) they are intended to be the input of the level-2 analysis that can be carried out in a second step to calculate more refined physical quantities. At this step the Goal.1 of the project is fulfilled, and a first step towards Goal.3 is also achieved.

Data types:
the fields (2D in data1 and 3D in data2) are stored using parallel vtk writing method. The particles (in dataP) uses serial HDF5. Other files have .txt format (in texts) and .png formats (in images).

Data production methods:
These level-1 data are produced starting from level-0 data and using the post-process python package available here (https://github.com/flavorenti/iPIC3D-OCA) in sub-directory post-process/pp-level-1 .

Storage facility:

work space on national french super-computing facilities (usually TGCC/Irene) with usual quota threshold at 5Tb (http://www-hpc.cea.fr/fr/complexe/tgcc-systeme-stockage.htm). Work space is backed up in the licallo machine based in Nice under the path … (TO BE ADDED)

Expected size of the data:
For a typical high-quality simulation (4^3 ppc, 830^3 cells, 2 species) we can expect a total data volume of the order of 500-1000 Gb. Texts and Images sub-directories occupy negligible fraction of space (around 100 Mb). On the other hand, data1 (~ 10 Gb), data2 (~550 Gb) and dataP (~250 Gb) sub-directories occupy the most space.

General data policy
These data can be shared among the scientific community, under unanimous decision of the numerical project team. These level-1 data are a first refined product that can be used for scientific studies and which is quite easily transferable to other clusters using the GEANT network (https://map.geant.org) at a transfer speed ~ 10 Gb/s, thus a transfer time of around 100 seconds (more realistically the transfer rate can be ~ 20 Mb/s, thus a transfer time of ~ 13 hours).

Data property
Level-1 data belong to the numerical project team. The team is responsible for sharing and using these data.

DATA SETS LEVEL-2 (TO BE DONE)

Data products:
level-2 data products are produced from post-processing of the level-1 data and transferred to the work space sub-directories. The files are named as follows:

- *<trajectory>:*
  - *output_good-pcls_<nrun>_<cycle>_<trajectory>.txt list of particles (x,y,z,u,v,w,q) around the trajectory <trajectory> in a given range dr. Particles extracted from the restart<nproc>.hdf files saved in dataP.*
  - *B_<trajectory>*
  - *E_<trajectory>*
  - *V, T, rho… (TO BE DONE)*

FAIR DATA - MAKING DATA FINDABLE

Standards and Metadata format*

The metadata of our simulations are stored in the file SimulationData.txt. This file is available at all levels of processing of the data (level-0,1,2) and contains the most important simulation parameters used in all runs. If a wider set of metadata is needed ask the team for level-0 data where more metadata can be read from the inputfile. To the latest version of the code metadata have this structure:

```
----------------------------
-  Simulation Parameters  -
----------------------------
Number of species = ...
qom[%s] = ...
----------------------------
x-Length               = ...
y-Length               = ...
z-Length               = ...
Number of cells (x) = ...
Number of cells (y) = ...
Number of cells (z) = ...
----------------------------
MPI procs (x)         = ...
MPI procs (y)         = ...
MPI procs (z)         = ...
----------------------------
x-center               = ...
y-center               = ...
z-center               = ...
Radius planet        = ...
Dipole Offset        = ...
----------------------------
SAL                    = ...
Nlayers_SAL        = ...
----------------------------
Time step              = ...
Number of cycles  = ...
----------------------------
rho init species %s = 1
rho inject species %s = 1
current sheet thickness  = 0.5
B0x                    = ...
B0y                    = ...
B0z                    = ...
----------------------------
v0x                    = ...
v0y                    = ...
v0z                    = ...
vth[%s]               = ...
----------------------------
Smooth                        = 0.5
SmoothNiter                = 2
GMRES error tolerance  = 0.001
CG error tolerance         = 0.001
Mover error tolerance     = 8
```

*Where %s is an integer defining the species (usually 0,1 for electrons and ions respectively). Fields with … means that they are usually variable while those with precise numbers are quite fixed among different simulations. Same naming conventions follows: qom= charge over mass ratio (negative per electrons and positive for ions), x,y,z lenghts are in units of the ion skin depth (di) computed from the density in the solar wind, same for all other length scales, SAL means Simple Absorbing Layer and defines the width of the external boundary conditions, B0 is the solar wind magnetic field in units of mi*wpi/e (where wpi is the plasma frequency in the solar wind), v0 is the fluid speed of the solar wind (in units of the speed of light c), vth is the thermal speed (still units of c).*

FAIR DATA - MAKING DATA ACCESSIBLE

*Data openly available**
We provide open access to level-1 data-sets of simulations used for scientific publications. Once the publication by the numerical team is accepted, level-1 data will be publicly available on the platform https://zenodo.org (sub-directories data2 and dataP are not provided given the limited storage of the platform).

*Tools to read or re-use data**
Python or paraview can be used to open and visualize the data stored on the platform.

ARCHIVING AND PRESERVATION

*Potential value of long term preservation**
Support future works and developments stemming from the publications ongoing during the project. Particularly relevant given the connection of this project with space missions that will extend and operate will after the end of this project.

Data at the end of the project
At the end of the PhD the 01/01/2014 we expect a level-1 and level-2 data volume of the order of ~10 Tb. These data shall be saved on the long term. Moreover level-0 compressed run can be saved in case more than 10 Tb of space are available for long-term storage.

*Recommended preservation duration**
10 years, this means until 01/01/2034. This will ensure coverage over BepiColombo science phase.

*Long term preservation storage**
Archiving data on the licallo local cluster in Nice seems the best option.

SCIENTIFIC PUBLICATIONS