The ECHO code for astrophysical relativistic plasmas: GPU acceleration, turbulence, and Pulsar Wind Nebulae

Luca Del Zanna

Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze

INAF - Osservatorio Astrofisico di Arcetri

INFN - Sezione di Firenze

ICSC - Spoke 2: Fundamental Research & Space Economy

luca.delzanna@unifi.it

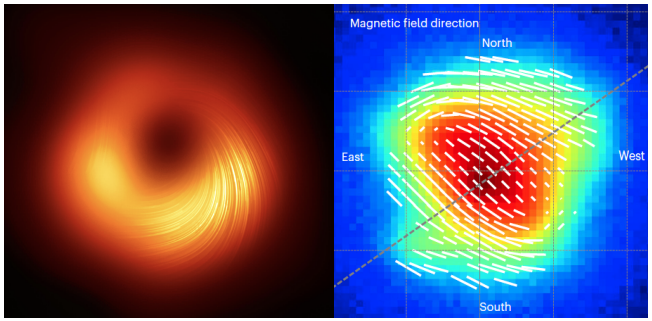
GPUs and first case study: relativistic MHD turbulence in PWNe

HPC motivation

- Digitalization and HPC encouraged and financed worldwide
- Pre-exascale era of supercomputing, Leonardo (top 10) at CINECA in Italy
- Porting of codes to GPUs necessary
- Turbulence studies require large resources

Astrophysical motivation

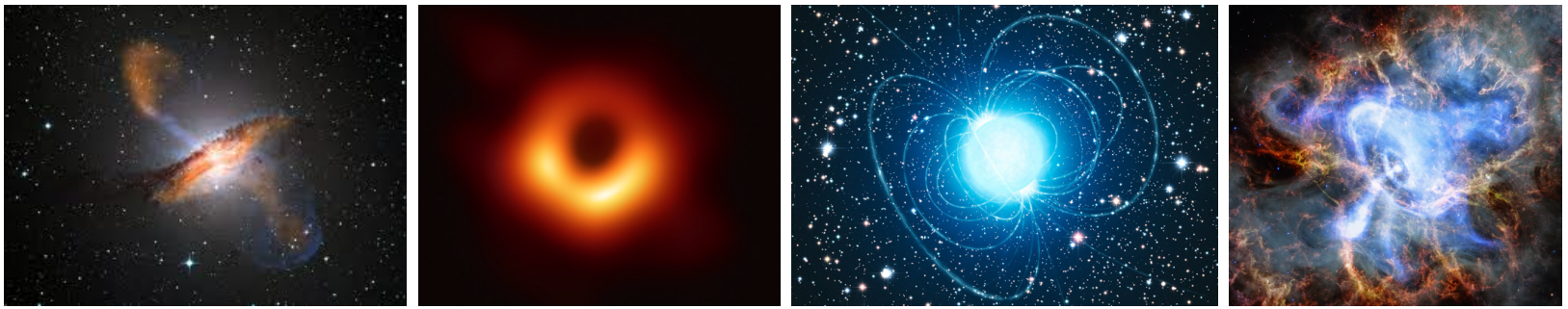
- Synchrotron polarization observed in many high-energy astrophysical sources, degree of turbulence can be estimated
- Pulsar Wind Nebulae (PWNe) are the best laboratories for relativistic MHD dynamics, turbulence and non-thermal emission



Relativistic plasmas in High Energy Astrophysics

Relativistic plasmas are ubiquitous in High Energy Astrophysics sources, such as:

- AGN jets
- Accretion tori around black holes
- Pulsars and magnetars
- Pulsar Wind Nebulae



The magnetic field is always a crucial ingredient, often providing an ordered structure to the object, a source of energy and of particle acceleration (reconnection), and leading to characteristic non-thermal emission (synchrotron, often highly linearly polarized).

Relativistic MHD: physical conditions

Astrophysical plasmas are often described by the single-fluid closure of magneto-hydrodynamics (MHD), especially if global large scales are treated.

The use of (general) relativistic MHD is necessary to treat extreme conditions, like:

- Fast bulk flow velocities:
$$\gamma \gg 1 \Rightarrow v \lesssim c$$
- Fast kinetic velocities:
$$p \gtrsim \rho c^2 \Rightarrow c_s \lesssim c$$
- Strong magnetic fields:
$$B^2 \gtrsim \rho c^2 \Rightarrow c_a \lesssim c$$
- Strong gravity in the vicinity of compact objets (GRMHD in this case):
$$r \gtrsim 2GM/c^2 \Rightarrow v_e \lesssim c$$

The (ideal) relativistic MHD energy-momentum tensor under these assumptions is:

$$\boxed{T_{\mu\nu} = (\rho h + b^2) u_\mu u_\nu - b_\mu b_\nu + (p + b^2/2) g_{\mu\nu}}$$

where h is the specific relativistic enthalpy and b^μ the magnetic field in the fluid frame (while $e^\mu = 0$). The system of ideal relativistic MHD equations in covariant form is then:

$$\boxed{\nabla_\mu(\rho u^\mu) = 0, \qquad \nabla_\mu T^{\mu\nu} = 0, \qquad \nabla_\mu(u^\mu b^\nu - b^\mu u^\nu) = 0}$$

Relativistic MHD: flat metric equations

Given the extreme conditions, shocks and nonlinearities are often encountered, so a numerical approach is needed, especially in multi-D problems.

Modern GRMHD codes rely on the so-called $3+1$ split of the metric and quantities: the curvature of metric (i.e. gravity) enters as modified quantities and extra source terms.

In special relativistic MHD gravity is neglected. The equations in flat metric are:

$$\begin{aligned}
&\partial_t D + \boldsymbol{\nabla}\cdot(\rho\Gamma\boldsymbol{v}) = 0,\\
&\partial_t \boldsymbol{S} + \boldsymbol{\nabla}\cdot(\rho h\Gamma^2\boldsymbol{vv} - \boldsymbol{EE} - \boldsymbol{BB} + (p + \tfrac{1}{2}(E^2+B^2))\mathcal{I}) = 0,\\
&\partial_t U + \boldsymbol{\nabla}\cdot(\rho h\Gamma^2\boldsymbol{v} + \boldsymbol{E}\times\boldsymbol{B}) = 0,\\
&\partial_t \boldsymbol{B} + \boldsymbol{\nabla}\times\boldsymbol{E} = 0,
\end{aligned}$$

where Γ is the flow Lorentz factor, and the conserved quantities are

$$\begin{aligned}
D =& \rho\Gamma,\\
\boldsymbol{S} =& \rho h\Gamma^2\boldsymbol{v} + \boldsymbol{E}\times\boldsymbol{B},\\
U =& \rho h\Gamma^2 - p + \tfrac{1}{2}(E^2+B^2),
\end{aligned}$$

with $\boldsymbol{E} = -\boldsymbol{v}\times\boldsymbol{B}$ (ideal MHD) and the addition of the solenoidal constraint $\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0$.

Relativistic MHD turbulence is often studied in a local parcel of plasma, modeled as a (periodical) numerical box, where the above equations apply.

The ECHO code

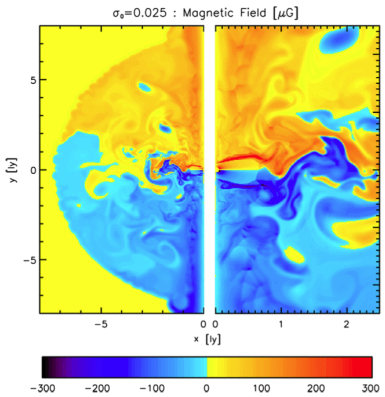
The Eulerian Conservative High Order code is a *home-made* shock-capturing code for astrophysical MHD (classic and relativistic, ideal and resistive/dynamo), featuring:

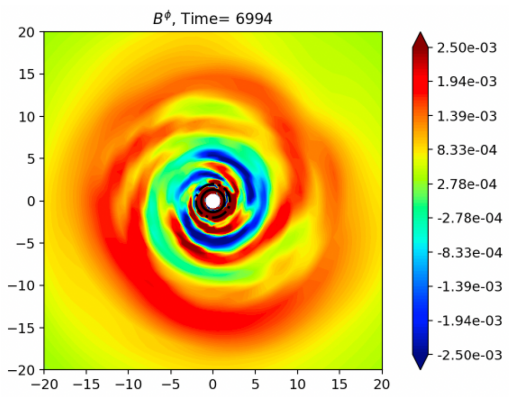
- conservative approach (even in 3 + 1 GRMHD), finite-differences (much easier!)
- High Order reconstruction of point-value primitives, HLL fluxes, HO flux derivatives
- Upwind Constrained Transport for $\nabla \cdot \boldsymbol{B} = 0$ (UCT: Londrillo & Del Zanna 2000, 2004)

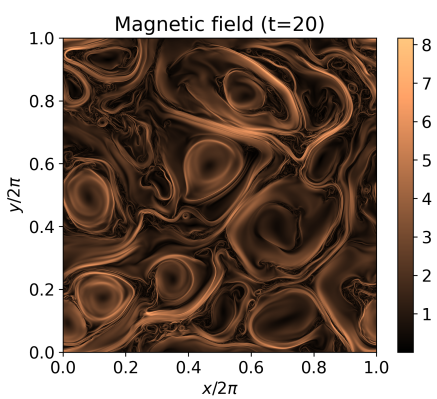
ECHO

ECHO is not public, it is continuously developed and maintained at the University of Florence (Italy) since 2000, the main reference is: Del Zanna et al., A&A 473, 11, 2007.

Examples below are: Pulsar Wind Nebulae (left panel), accretion disks around black holes (central panel), and high-resolution decaying turbulence (right panel):







Discretization and parallelization strategy in ECHO

The MHD and GRMHD equations in ECHO are spatially discretized using point values at different locations, in semi-discrete form the evolution equations are:

$$\frac{d}{dt}[\mathcal{U}_i]_c = -\sum_j \frac{1}{h_j}([\hat{\mathcal{F}}_i^j]_{S_j^+} - [\hat{\mathcal{F}}_i^j]_{S_j^-}) + [\mathcal{S}_i]_c,$$

$$\frac{d}{dt}[\mathcal{B}^i]_{S_i^+} = -\sum_{j,k}[ijk]\frac{1}{h_j}([\hat{\mathcal{E}}_k]_{L_k^+} - [\hat{\mathcal{E}}_k]_{L_k^-}),$$

where the first set are standard conservative equations (plus source terms) and the second one is the discretized form of the curl-like induction equation (UCT method).

Best strategy for parallelization and acceleration of ECHO: MPI domain decomposition. Then, for any task and for each time (sub-) iteration, 3D loops are needed for:

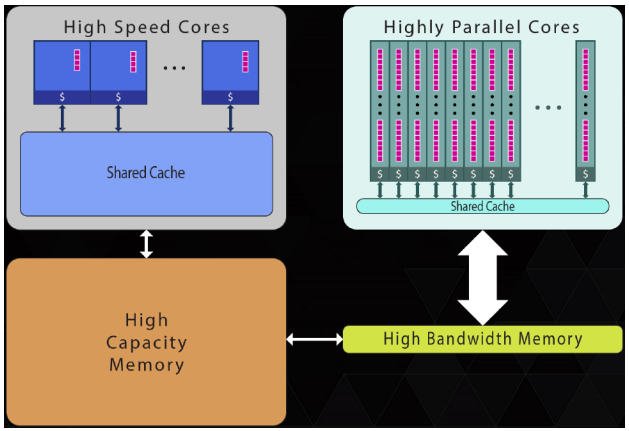
- REC step: reconstructed primitives and numerical fluxes $\mathcal{F}_i^j$ (and $\mathcal{E}_k$ for UCT)
- DER step: their higher-order corrections (HO, hatted quantities) and derivatives
- primitive variables from conservative ones
- the local and then global timestep
- the update in time according to the above equations (Runge-Kutta methods)

where the above loops must perform just local and atomic computations.

From CPUs to GPUs

High Performing Computing in the pre-exascale era is drifting from hardware based on (multicore) CPUs towards accelerated devices, GPUs, characterized by:

- lower clock rate, that is lower speed cores (BAD)
- less energy consuming units (GOOD)
- highly parallel cores (a lot of them!), strong multithreading (GOOD)



Problem: how to move data between CPUs and GPUs? How to accelerate loops?

Astrophysicists need to employ IT and software engineers to do that, then use either:

- new languages or extensions (CUDA)
- meta-programming libraries (SYCL, DPC++, KOKKOS)
- directives (OpenACC)

or . . . nothing at all: Standard Language Parallelism, just a programmer's dream?

Technical blogs and papers (e.g. J. Larkin, R. Caplan) suggest that Modern Fortran (DC loops), a few OpenACC directives, and the NVIDIA Unified Memory paradigm are enough for acceleration! It seemed the best option for the porting of ECHO on GPUs...

A first command assigns a single GPU for each MPI task:

```
!$acc set device_num(mod(rank,ngpu))
```

Fortran **`do concurrent`** (DC) constructs must be used for the main 3D loops, e.g.:

```
do concurrent (iz=iz1:iz2,iy=iy1:iy2,ix=ix0:ix2) local(g,s)
```

where locality must be explicitly declared for any array (`s`) or structure (`g`) used inside.

Internal routines and functions run concurrently only if declared **`pure`** (no side effects):

```
pure function holib_rec(s) result(rec)
```

followed by the directive `!$acc routine` to offload them on the accelerated device.

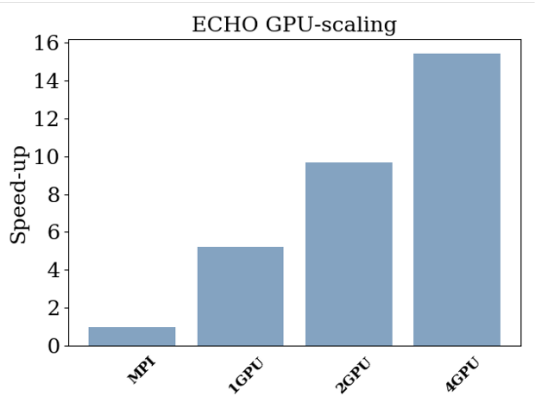
Finally, the NVIDIA Fortran compiler is invoked simply with

```
nvfortran -stdpar=gpu
```

or via the corresponding MPI command for parallel runs.

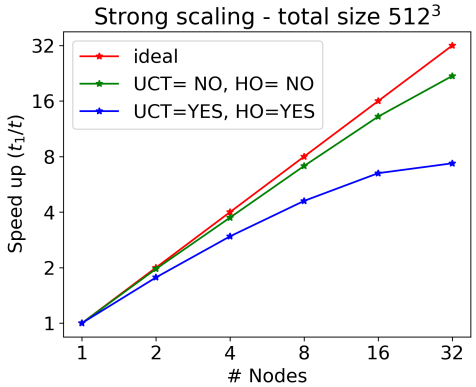
Speed-up and strong scaling

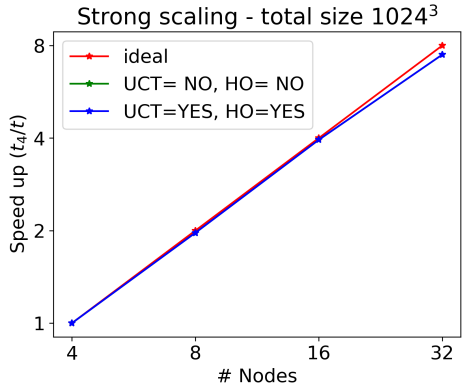
NVIDIA-CINECA hackathon in June 2022, speed-up on a single node (Marconi100):



32 tasks on CPUs ($\times$4 hyper-threading) vs 1 $-$ 4 Volta V100 GPUs: $\times$16 speed-up

Strong scaling tests on the top-10 cluster Leonardo (Ampere A100, 4 GPUs per node):



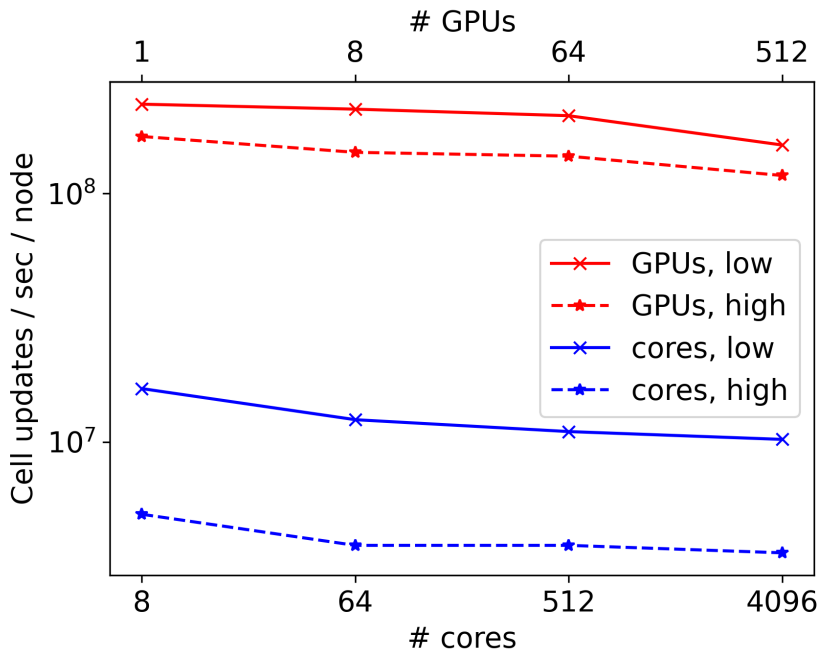


Scaling improves by minimizing communications (MPI `sendrecv` directives), or if the number of cells per node is increased. Full details in:

Del Zanna et al., *Fluids* 9, 26, 2024

Efficiency and weak scaling

We measured efficiency and weak scaling using 200^3 per CPU core vs 400^3 per GPU (one GPU is equivalent to 8 cores) on Leonardo, up to 128 nodes and 3200^3 cells:



Speed up of GPUs over CPU cores ranges from $\times 15$ for low resolution runs (REC 2nd order, RK2, no UCT) up to $\times 38$ for high resolution runs (REC 5th order, RK3, UCT).

All efficiency and scaling tests refer to a relativistic Alfvén wave propagating along the diagonal of a cubic periodic domain, evolving 9 variables.

Results are equivalent, if not better, than AthenaK (Kokkos) or Pluto (OpenACC).

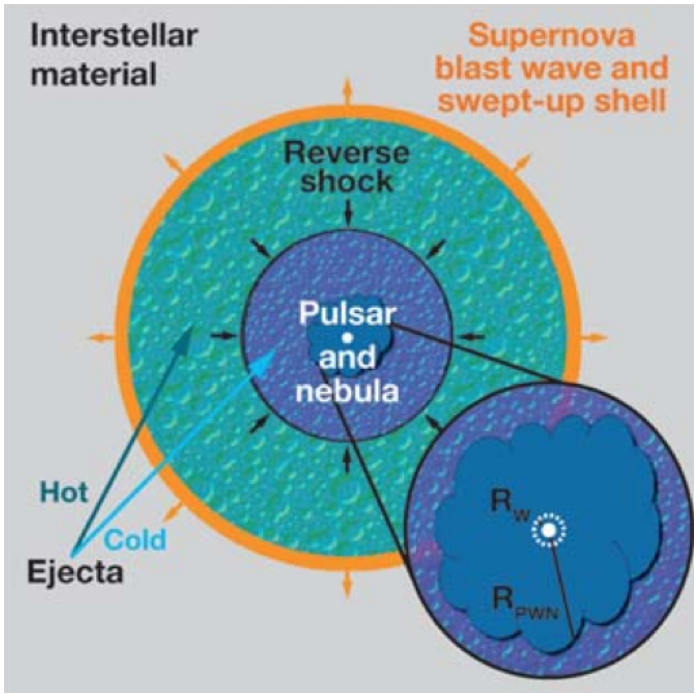
Pulsar Wind Nebulae: overview

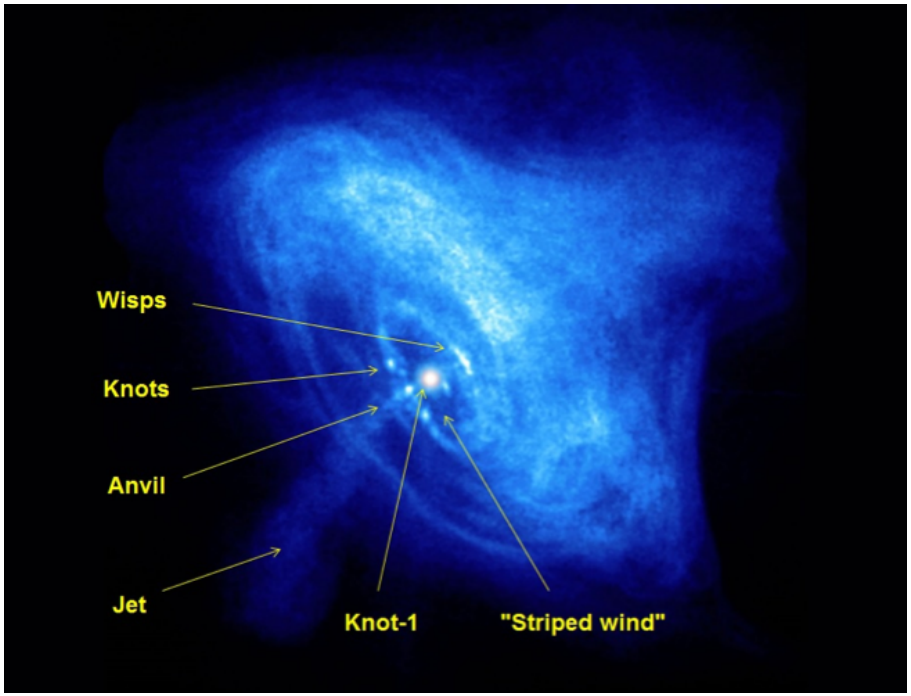
The Crab Nebula is powered by the pulsar spin-down luminosity (Pacini 1967!)

$$L_0 = -I\Omega\dot{\Omega} \approx 5 \times 10^{38}\ \mathrm{erg/s}$$

in the form of a magnetized relativistic wind. Interaction with the slowly expanding SN ejecta creates the hot plasma bubble (Rees & Gunn 1974; Kennel & Coroniti 1984).

The Crab Nebula is young (1000 years), it is a PWN in free expansion in the cold SNR ejecta, before the reverberation phase.



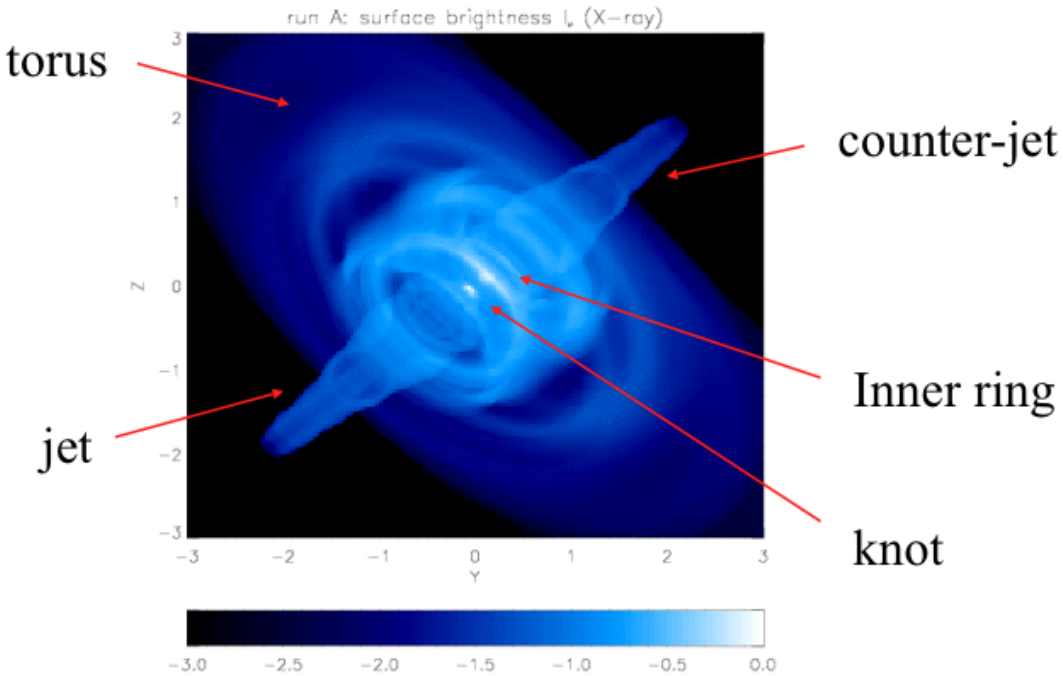


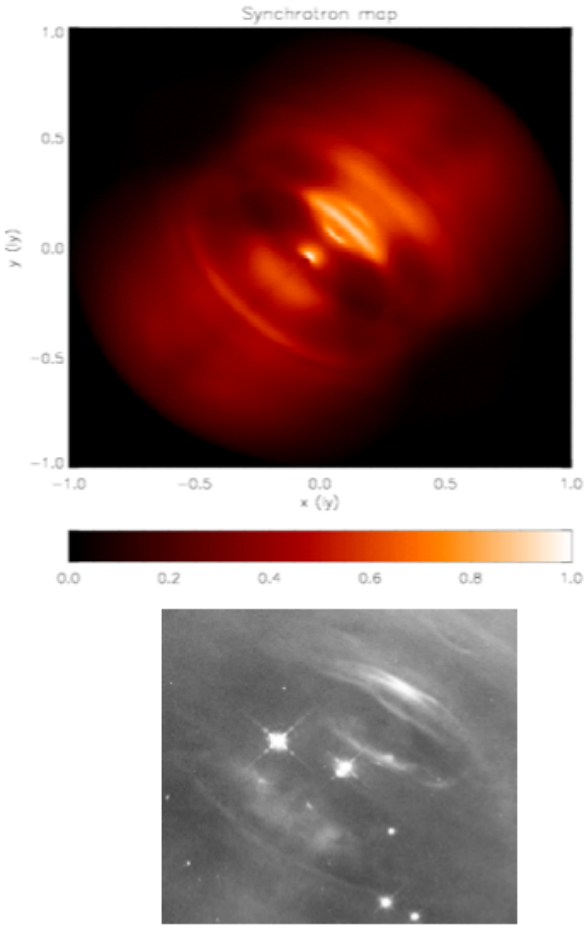
Chandra X-rays data (2000): can we model the jet-torus structure and fine details?!?

Simulated synchrotron maps from 2D axisymmetric relativistic MHD computations (here from our ECHO code, celebrating 20 years): success!

Simulated maps vs images:

(Del Zanna et al. 2006)

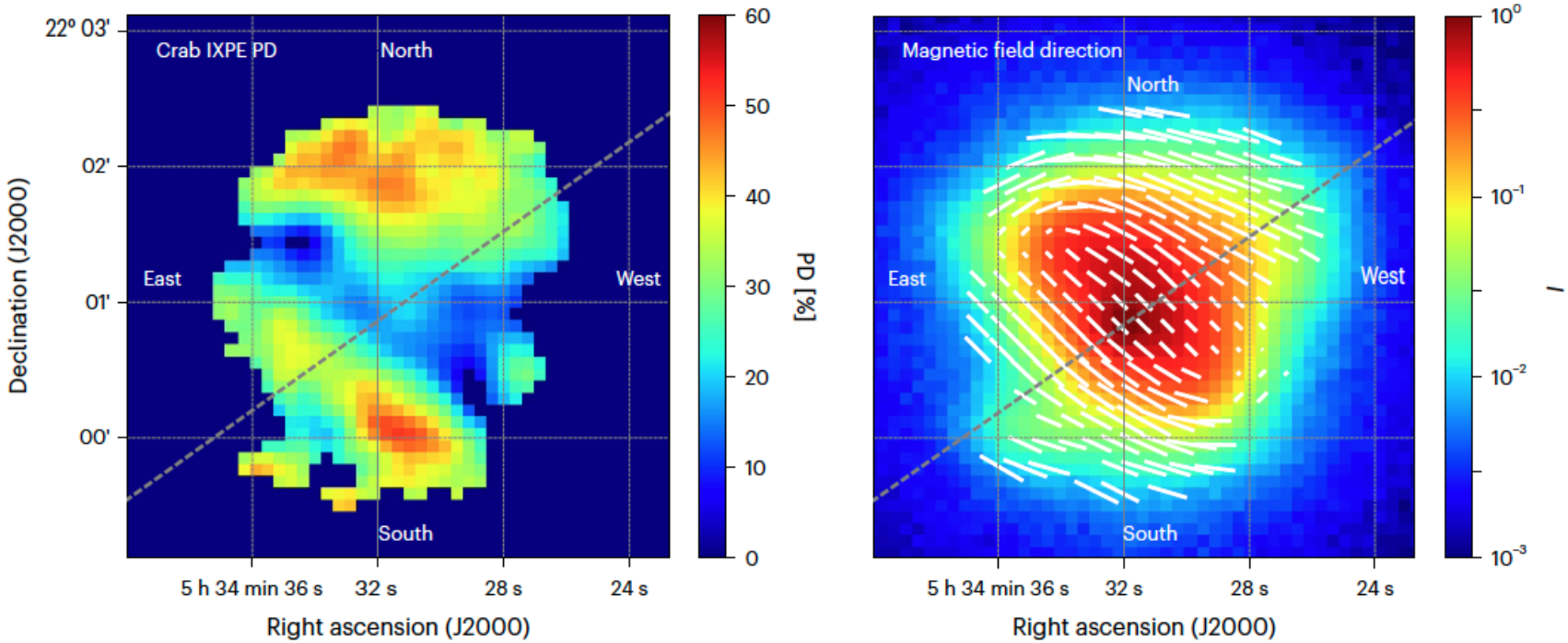




Recipes, including boosting and polarization, in (Del Zanna et al. 2004, 2006; Volpi 2008).
Motion of wisps also reproduced (Camus et al. 2009, Olmi et al. 2015).

New data: X-ray synchrotron polarization with IXPE

Further information on the nebular magnetic field comes from soft X-ray synchrotron polarization mesures, made possible by IXPE (Bucciantini et al. 2023).



The Crab Nebula clearly shows a toroidal field, confirming RMHD models.

The patchy distribution of the Polarized Degree (PD, $\sim 45 - 50\%$ max) may indicate different levels of turbulence (PD $= 70\%$ is the theoretical maximum for a uniform field).

We need to investigate synchrotron emission by a turbulent magnetic field.

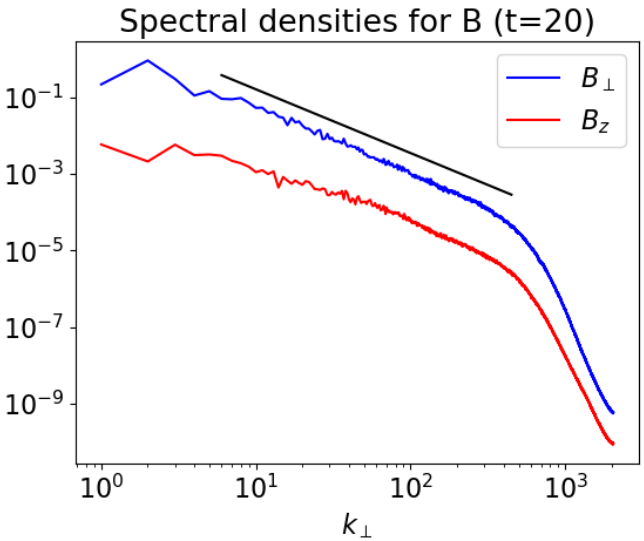
2D simulations: first test

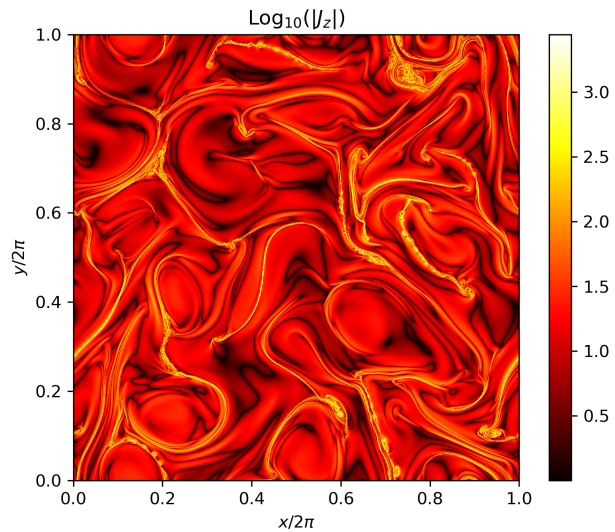
First test of the GPU version of ECHO: 2D 4096^2 decaying turbulence in the $\perp$ plane

$$\delta \boldsymbol{B} \propto B_0 \sum_{\boldsymbol{k}=-4}^{4} \frac{\boldsymbol{k}}{k} \cos(\boldsymbol{k} \cdot \boldsymbol{x} + \varphi) \times \boldsymbol{e}_z,$$

here $\boldsymbol{k} \equiv \boldsymbol{k}_\perp$, with a similar Alfvénic (incompressible) setup for $\delta \boldsymbol{v}$ in terms of c_a, with

$$\sigma = B_0^2/\rho = 100, \quad \beta = 2p/B_0^2 = 1, \quad \delta B_{\rm rms}/B_0 = 0.25$$





Clear Kolmogorov spectrum for more than two decades, for velocity fluctuations too; intermittent current sheets and plasmoids ubiquitous (Del Zanna et al. 2024, ideal MHD!).

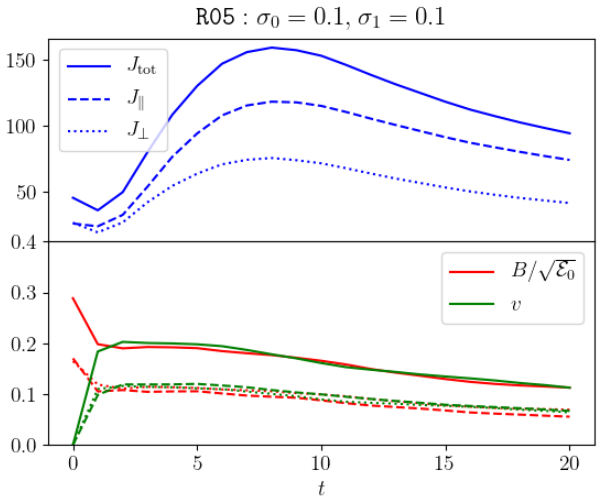
3D simulations: setup

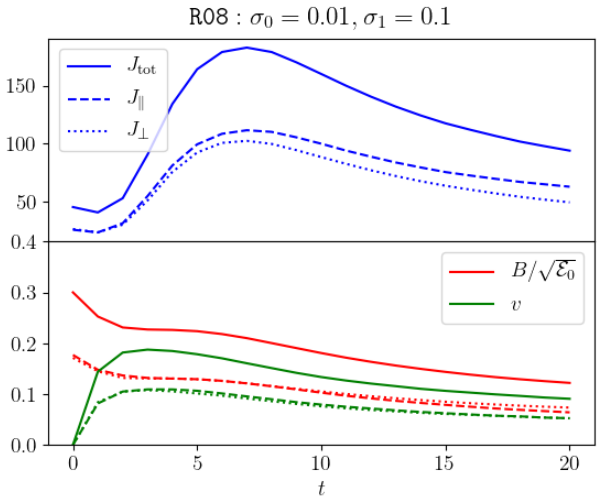
In full 3D we choose a different setup (with a view to applying to PWNe):

- hot plasma:
$$\rho = 1, \quad h = 1000 \Rightarrow p \gg \rho$$
- 3D magnetic fluctuations and $\boldsymbol{k}$, along B_0 too (φ and ϑ random, for each $\boldsymbol{k}$):
$$\delta\boldsymbol{B} \propto B_0 \sum_{\boldsymbol{k}=-4}^{4} \cos(\boldsymbol{k}\cdot\boldsymbol{x} + \varphi)(\cos\vartheta\,\boldsymbol{e}_1 + \sin\vartheta\,\boldsymbol{e}_2)$$
where
$$\boldsymbol{e}_1 = \frac{\boldsymbol{k}\times\boldsymbol{e}_z}{|\boldsymbol{k}\times\boldsymbol{e}_z|}, \qquad \boldsymbol{e}_2 = \frac{\boldsymbol{k}\times\boldsymbol{e}_1}{|\boldsymbol{k}|},$$
- no initial velocity fluctuations (energy transfer expected)
- 12 runs, from small to substantial mean and *rms* magnetizations (*hot* definition):
$$\sigma_0 = \frac{B_0^2}{\rho h} = 0.001,\ 0.01,\ 0.1,\ 1; \qquad \sigma_1 = \frac{\delta B_{\rm rms}^2}{\rho h} = 0.05,\ 0.1,\ 0.2$$
- corresponding to moderate to large initial amplitudes (from $\simeq 0.22$ to $\simeq 14$):
$$\delta B_{\rm rms}/B_0 = \sqrt{\sigma_1/\sigma_0}$$
- resolution 512^3: 15 minutes to run (8 GPUs), much more to download data!

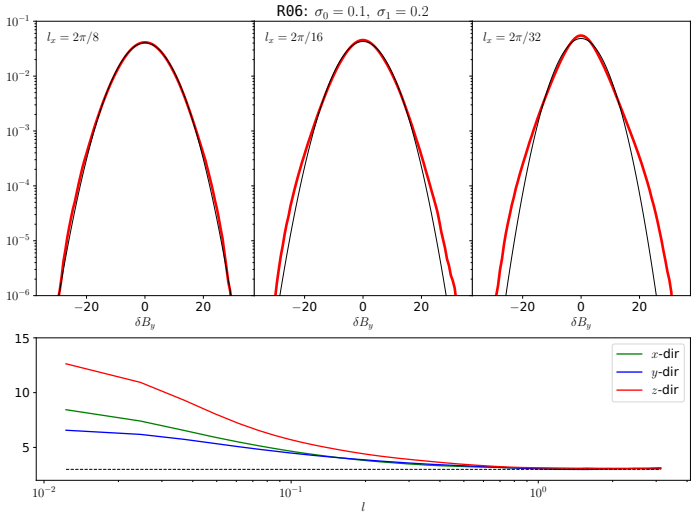
Analysis of 3D turbulence

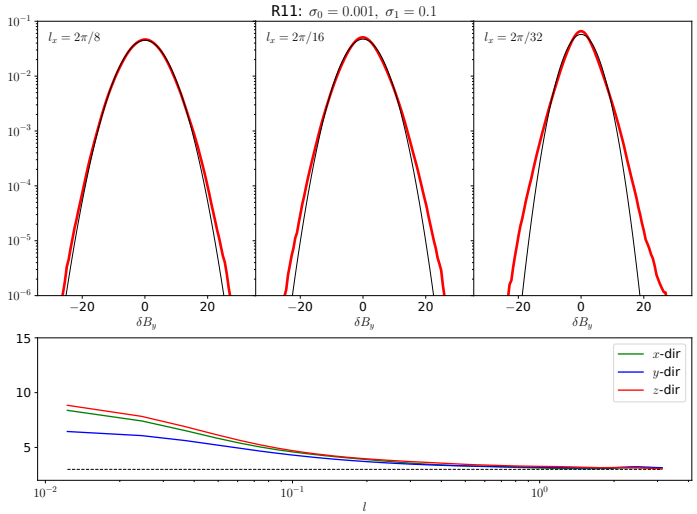
Time series of *rms* quantities, peak of turbulence and Alfvénic-type equilibrium:





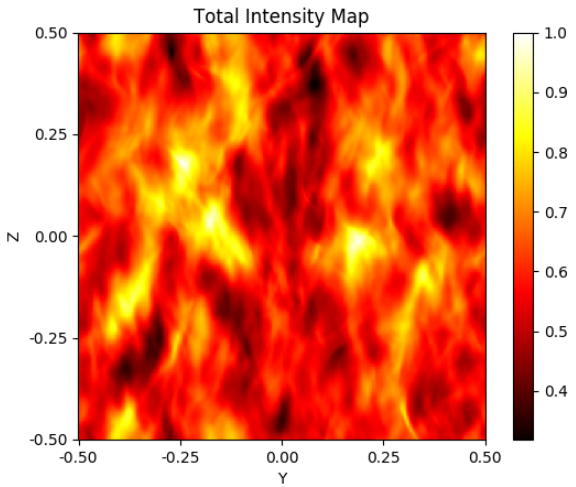
Well behaved turbulence: intermittency at small scales (here B_y across x scales):

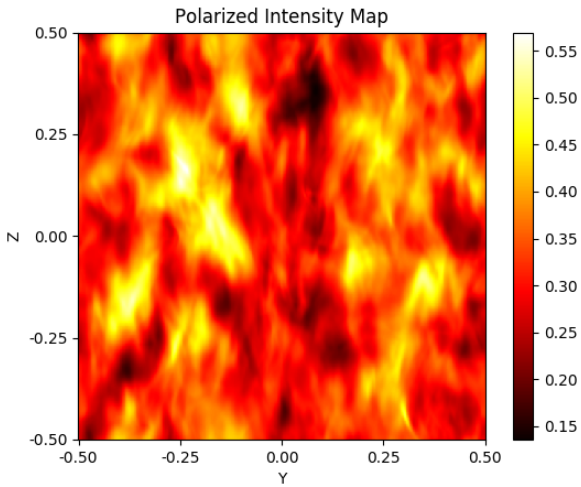




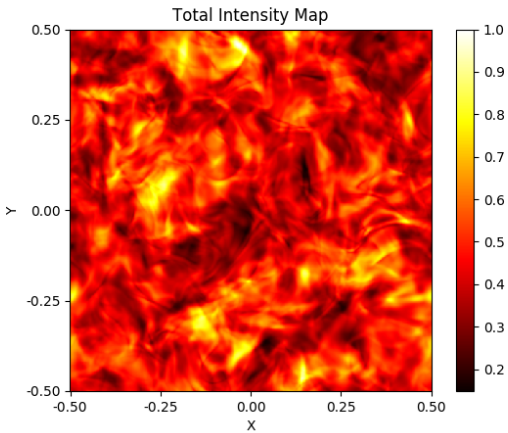
Synchrotron maps from turbulence simulations ($\sigma_0 = 0.1, \sigma_1 = 0.1$)

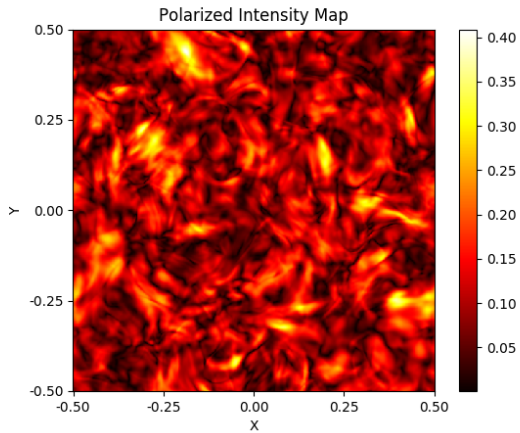
Integrated synchrotron maps along x, that is with LOS perpendicular to $\mathbf{B_0}$:





Integrated synchrotron maps along z, that is with LOS parallel to $\mathbf{B_0}$:

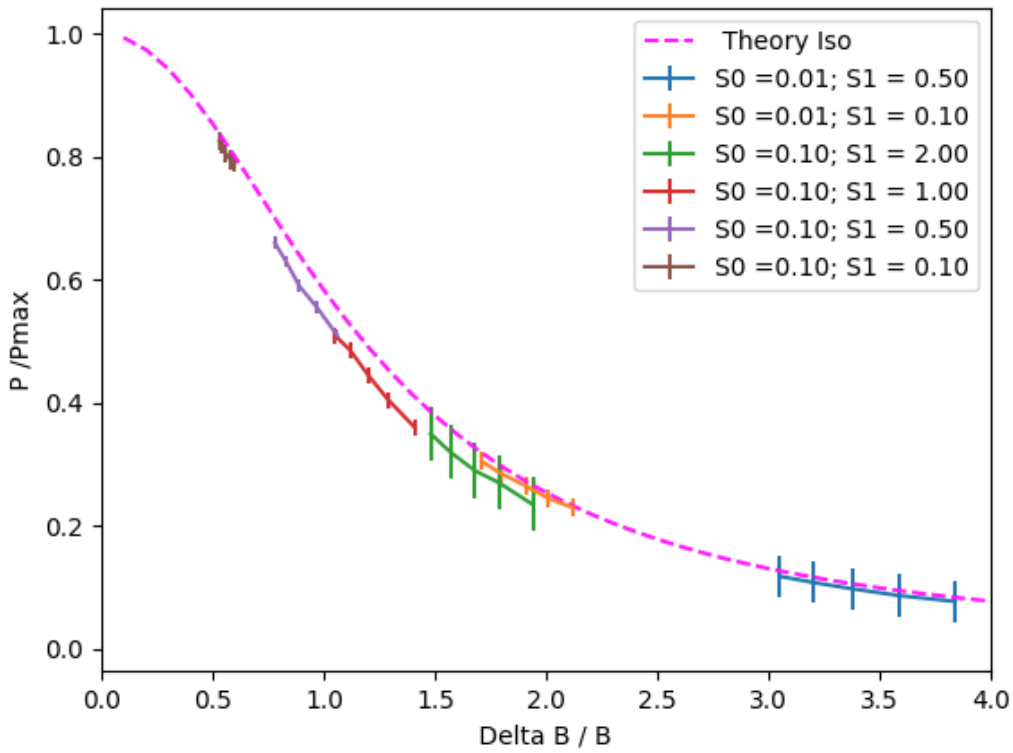




The polarized intensity is higher in the first case (center of the torus), as expected.

Polarized fraction and magnetic fluctuations (preliminary!)

In the explored range, the PD seems to follow a simple relation with $\delta B_{\rm rms}/B_0$ (which decreases in time after the turbulence peak):



Agreement with analytical predictions for Gaussian fluctuations (Bandiera & Petruk 2016):

$$\frac{\Pi}{\Pi_{\rm max}} = \frac{5+s}{8}\,\frac{{}_1F_1\left((3-s)/4,\,3,\,-x\right)}{{}_1F_1\left(-(1+s)/4,\,1,\,-x\right)}\,x; \qquad x = \frac{3}{2}\left(\frac{\delta B_{\rm rms}}{B_0}\right)^{-2}$$

Each tick is a 4.6 GB 512^3 data cube...

Summary

We briefly summarize our presentation and the results obtain:

- We have successfully ported ECHO on GPUs using standard Fortran constructs
- The code is up to 38 times faster and scales well up to 256 GPUs (400^3 per GPU)
- Ideal 2D and 3D RMHD turbulence: well behaved turbulence and dissipation
- Kolmogorov spectra for $\delta\mathbf{B}$, intermittency at small scales (plasmoids)
- Synchrotron polarization maps with LOS parallel or perpendicular to $\mathbf{B_0}$
- Gaussian fluctuations at large scales: analytical predictions for PD confirmed
- Different turbulence conditions explain different levels of PD observed in PWNe

However magnetization and initial amplitudes have a role. . . work is in progress!

Relevant papers:

- Del Zanna et al, A&A 453, 621, 2006 (Synchrotron recipes for RMHD simulations of PWNe)
- Del Zanna et al, A&A 473, 11, 2007 (The ECHO code for GRMHD)
- Bandiera and Petruk, MNRAS 459, 178, 2016 (Polarized emission for turbulent fields in SNRs)
- Bucciantini et al, MNRAS 470, 4066, 2017 (Polarized emission for turbulent fields in PWNe)
- Bucciantini et al, Nature Astr. 7, 602, 2023 (IXPE polarimetry of the Crab pulsar and PWN)
- Del Zanna et al, Fluids 9, 16, 2024 (The ECHO code on GPUs with standard Fortran)
- Del Zanna, Bucciantini, Landi, *in prep* (Synchrotron from 3D turbulence in hot plasmas)