Leave a star
It helps packaging !

# Shamrock

Beyond a PhD Code: The Rise of the Shamrock

No. 101053020 (Dust2Planets)

# Last year !


Joint Franco-Australian
5th Phantom and MCFOST
Users Workshop
2024
Storms knock out power to 174,000 homes in Australia
Published: 14 Feb 2024 - 08:25 am | Last Updated: 14 Feb 2024 - 08:32 am
Pictures of last year talk 😬
Domain decomposition
The global simulation is divided
into multiple patches
(abstract decomposition)
We only have to communicate
the interface to itera
• >1G part in a simulation
Patches are scattered across the nodes
If patches are too large they split
If they are too small they merge
10
ou take a big cluster
Adastra mi250x
GPUs
100
1000
8.73e+09
6.76e+09
5.34e+09
1.38e+09
• 9G part / seconds
• 65.000.000.000 particles
( 4000³ )
• 7 sec / iterations
Global time steps :
(Estimate 10k steps at 1M part)
Isothermal disc : 1G part = 1 orbit/16h
Isothermal disc : 16G part = 1 orbit/4day
29


?
PHANTOMSPH

# Why ???????????????????????????

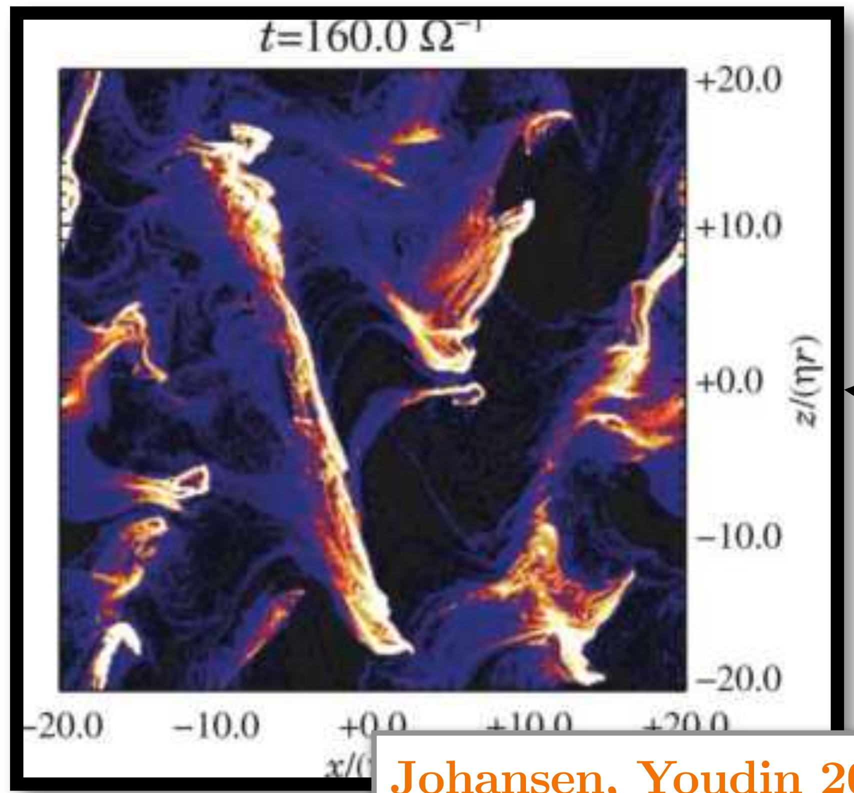


Why a new code ?

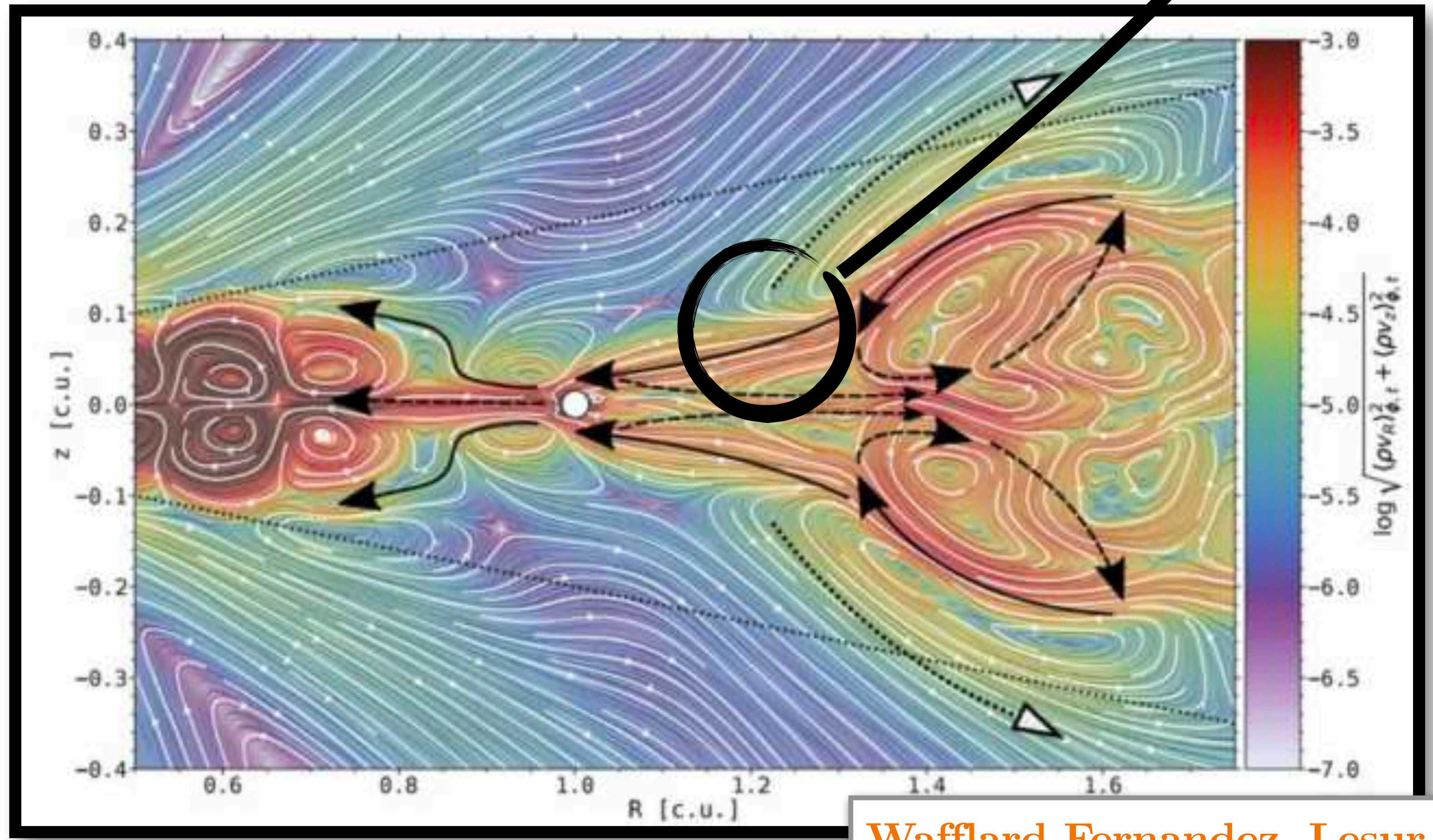
Sometimes we need a lot of resolution ...

# Few examples

Global instabilities ?
(1-10 Gpart)

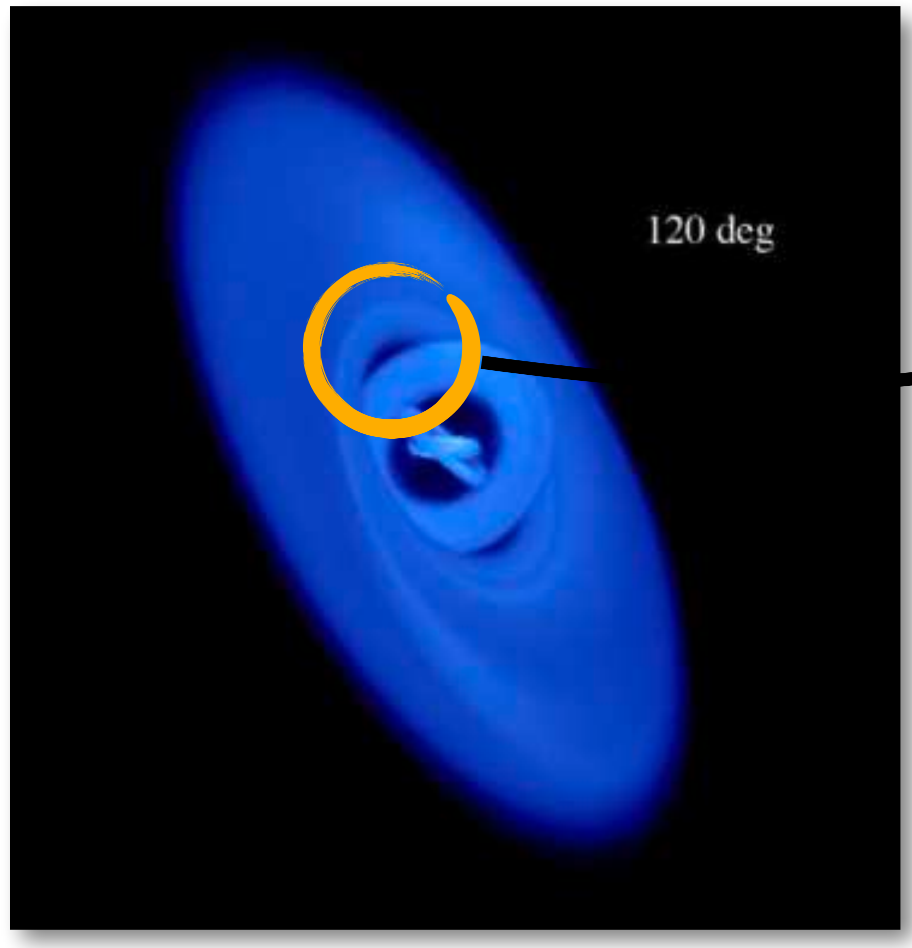


Johansen, Youdin 2007

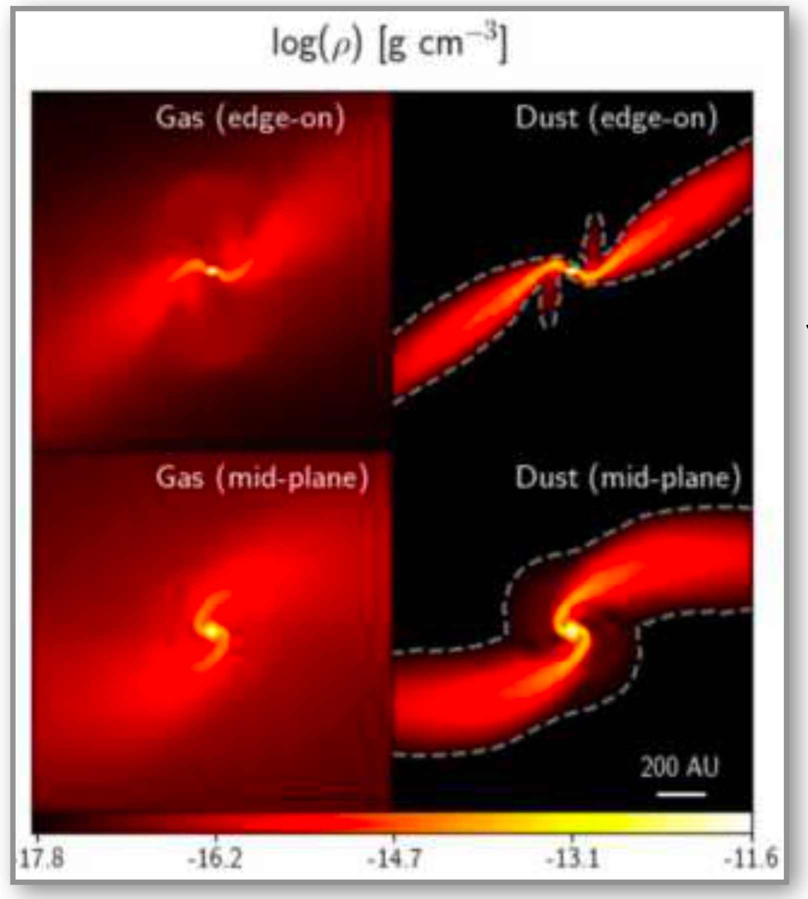


Wafflard-Fernandez, Lesur 2023

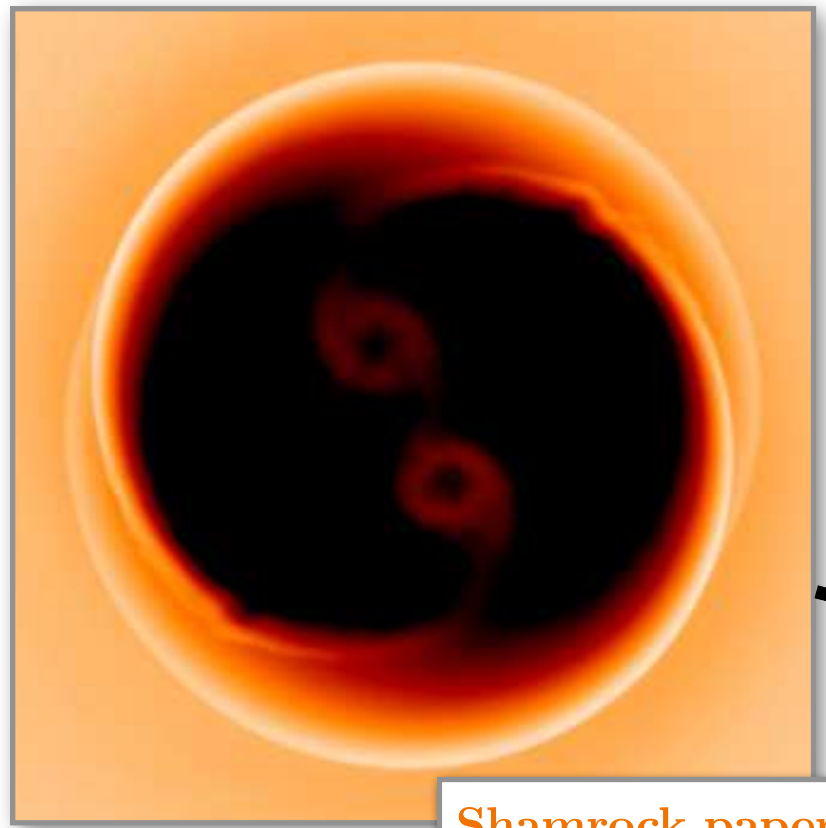
What happens here ?
(1 Gpart)




Envelope, disc, low $\alpha_{\rm num}$
(1 Gpart)




Circum-stellar accretion ?
(>1 Gpart)


Shamrock paper


**1Gpart** $\Leftrightarrow \alpha_{\rm num} \simeq 10^{-3}, 10^{-4}$

# Phantom on GPU ???




1 Exaflop

$= 10^{18}$ float operations per second

**Phantom SPH code :**

**Limitations :**

- **CPU only**
- **No MPI scalability (limited to few nodes)**
- **Soft limit around 10-100M particles**

**Phantom can not be ported to GPUs ⇒ New approach**

- Kd-tree build
- Domain decomposition
- Single loop approach

Shamrock

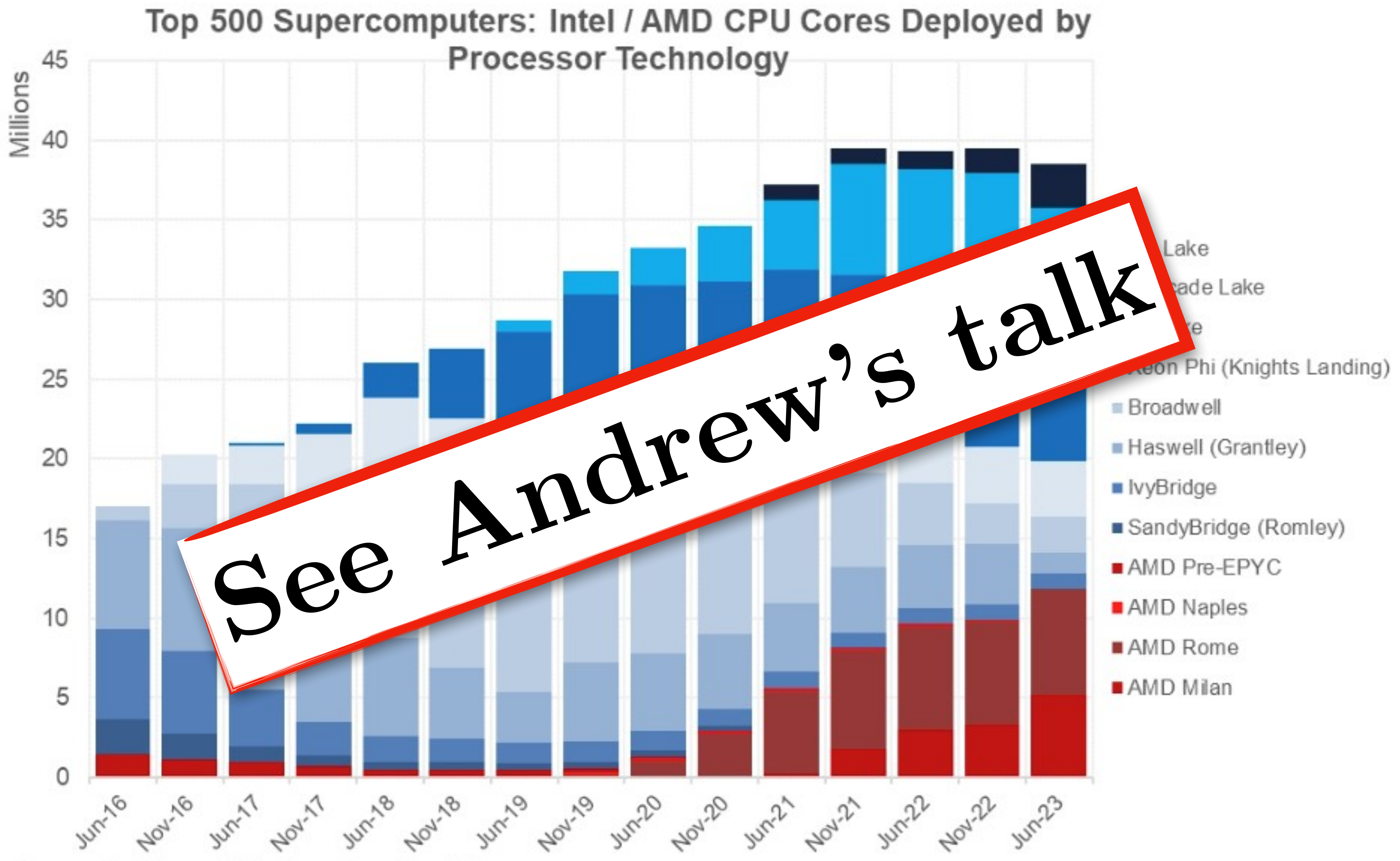
# Supercomputers



Top 500 list :

| Rank | System | Cores | Rmax (PFlop/s) | Rpeak (PFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States | 11,039,616 | 1,742.00 | 2,746.38 | 29,581 |
| 2 | Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States | 9,066,176 | 1,353.00 | 2,055.72 | 24,607 |
| 3 | Aurora - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States | 9,264,128 | 1,012.00 | 1,980.01 | 38,698 |
| 4 | Eagle - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States | 2,073,600 | 561.20 | 846.84 | |
| 5 | HPC6 - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy | 3,143,520 | 477.90 | 606.97 | 8,461 |

GPU GPU GPU GPU GPU



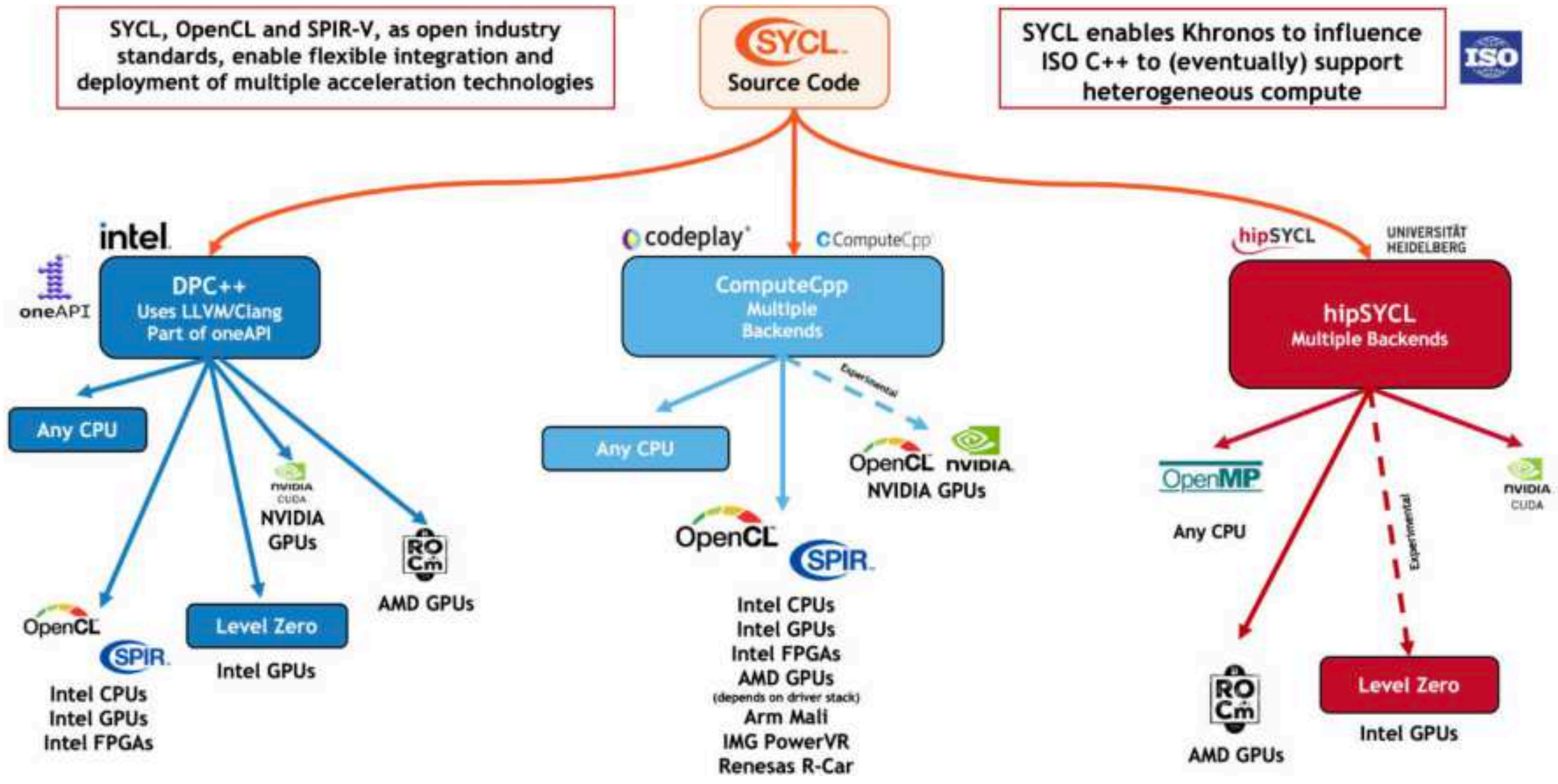

Source: Top500.org; Wells Fargo Securities, LLC

$\Rightarrow$ Supercomputers move to GPUs

**(Efficiency, density, AI, ...)**

**(But all vendors involved)**

Could also happen with NPUs soon-ish




**Our choice :** SYCL™

- C++17 extension
- Programming **standard**
- **Directly** compiles through CUDA, Hip, OpenMP, ...
- Supports **all major** CPU & GPUs (AMD, Nvidia, Intel, ARM)

# The summary

## The goal: Exascale


1

$= 10^{18}$ float o

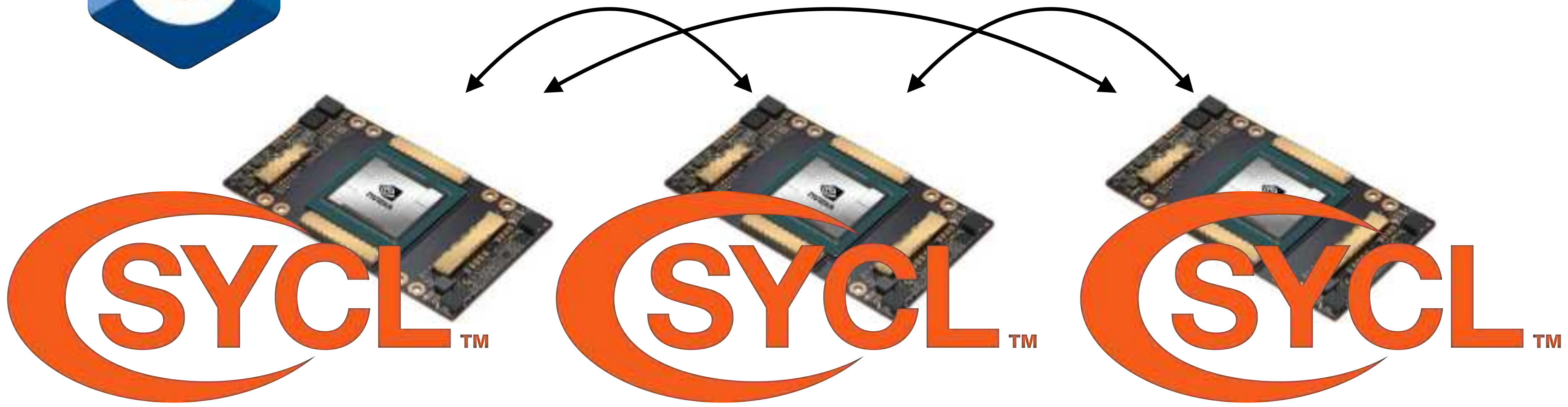

Technical talk at IWOCL

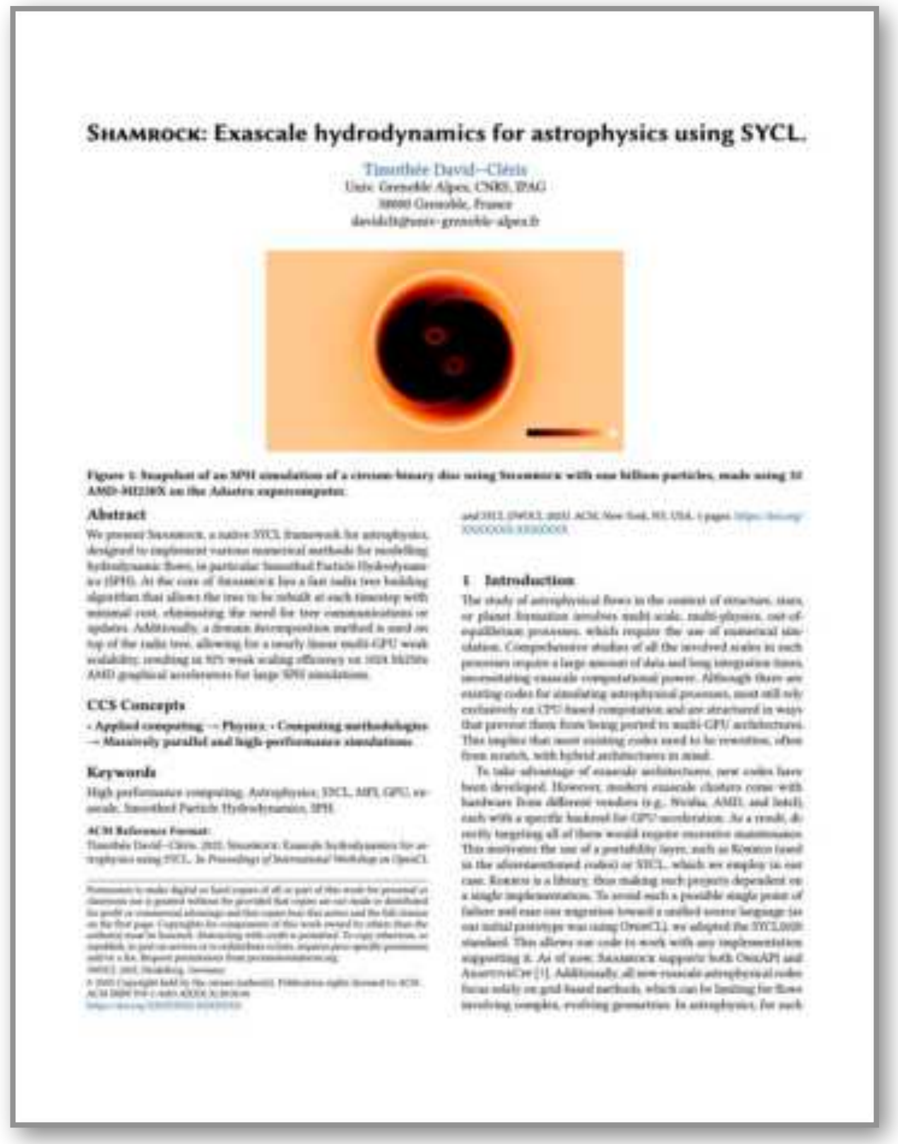
## How ?

Do everything on GPUs

17

MPI



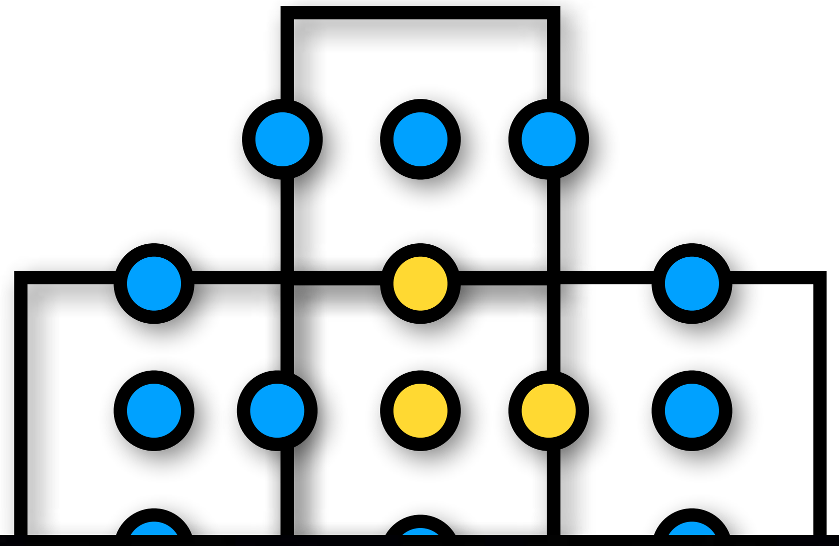

Accepted

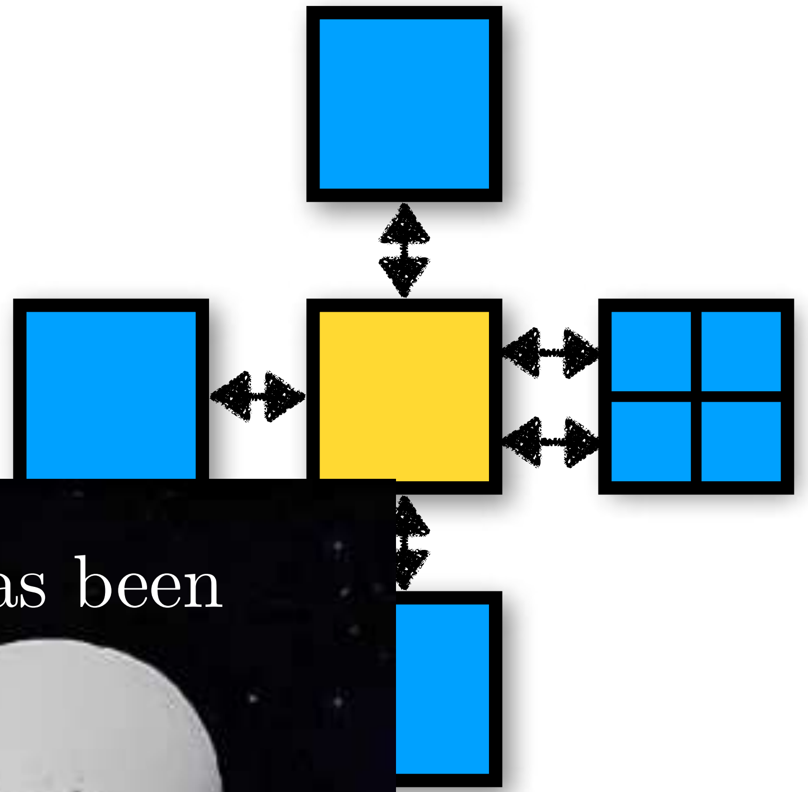

Published

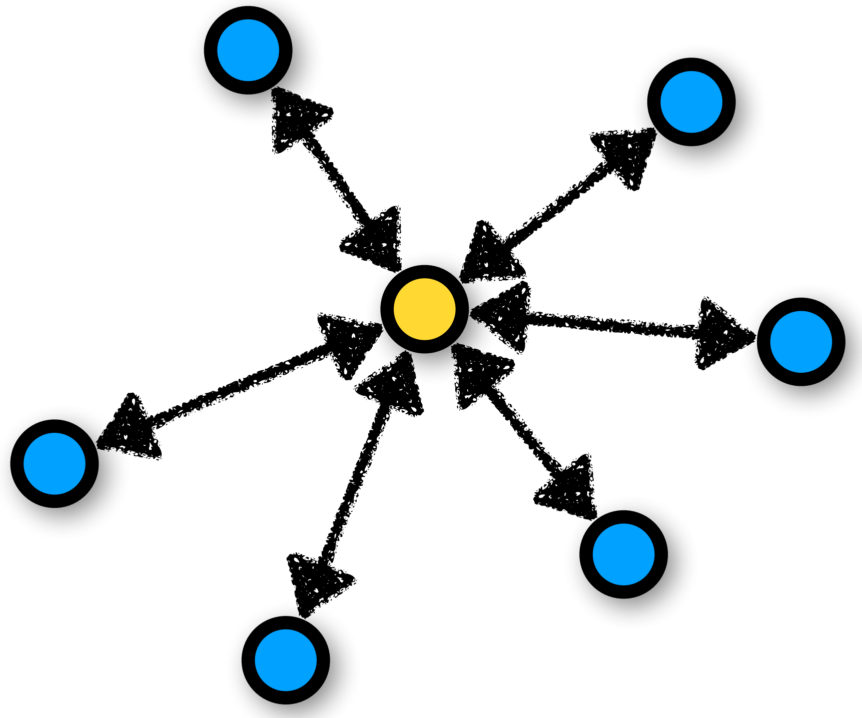
# 2-slide summary

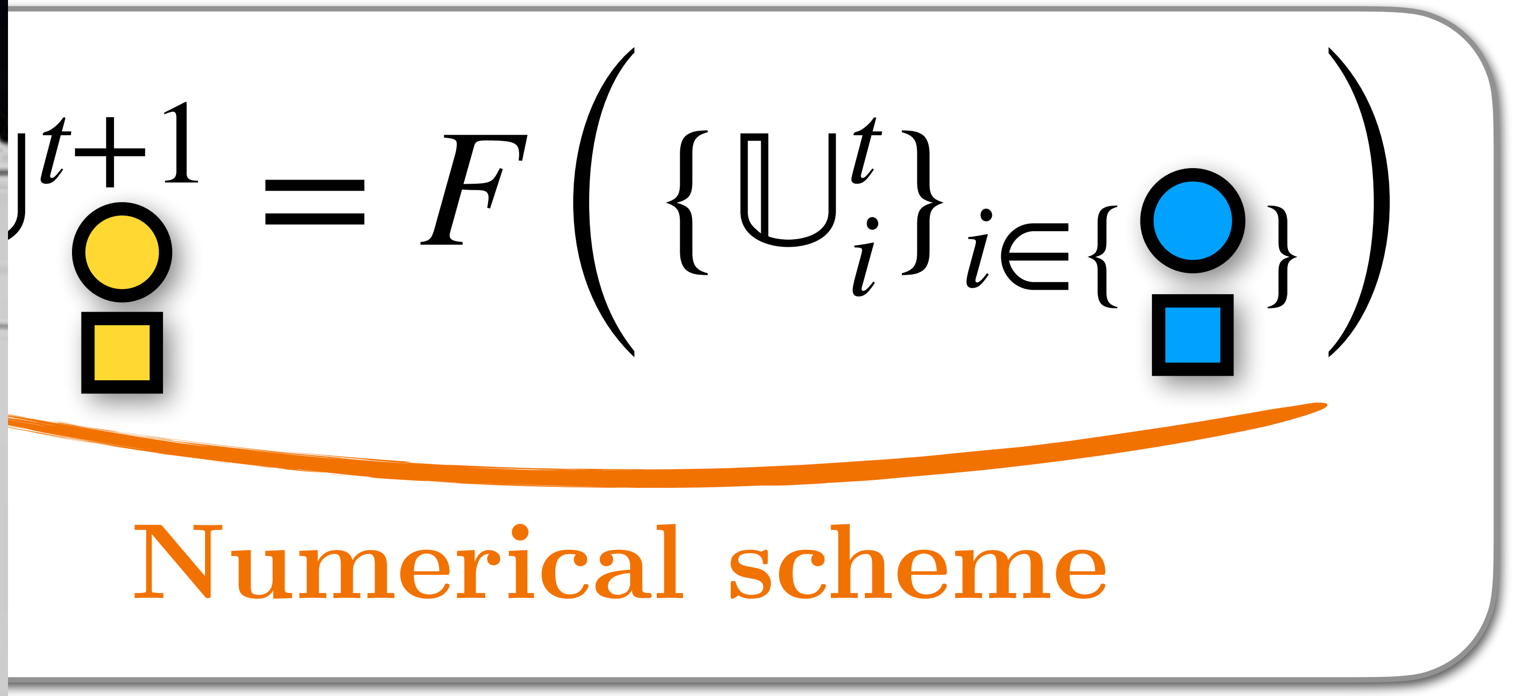
# All solvers are the "same"

**Finite elements** | **Finite Volumes** | **SPH**







# The goal in Shamrock



Shamrock

Criterion $\gamma$
Scheme $F$

Numerical method
+ Exascale scalability

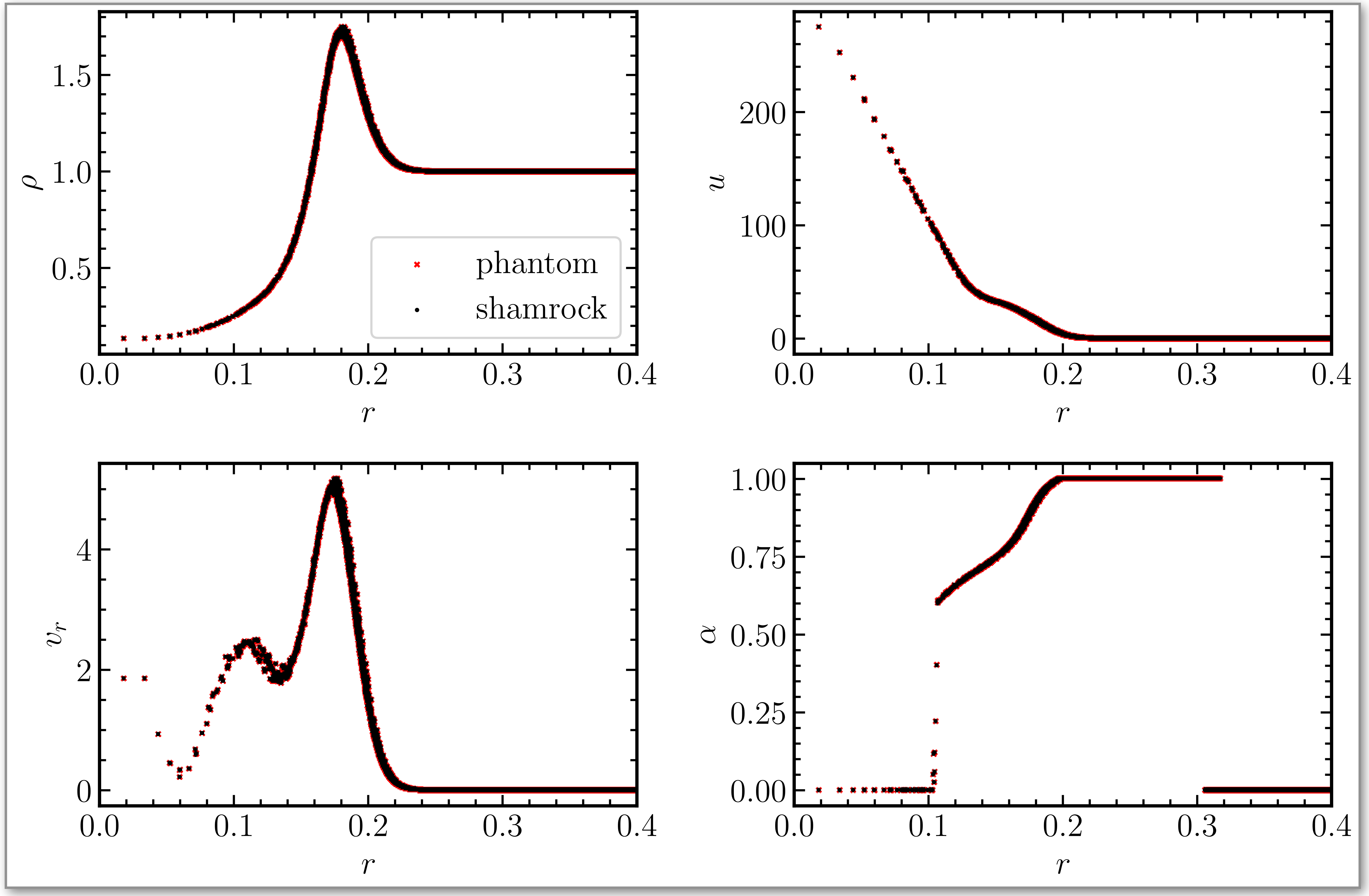
Generic modules

**Optimising SPH $\Leftrightarrow$ Optimising AMR**

Generic modules are coded **once** for all schemes

Reality is a bit more nuanced but that's the goal

## Sedov-blast wave:

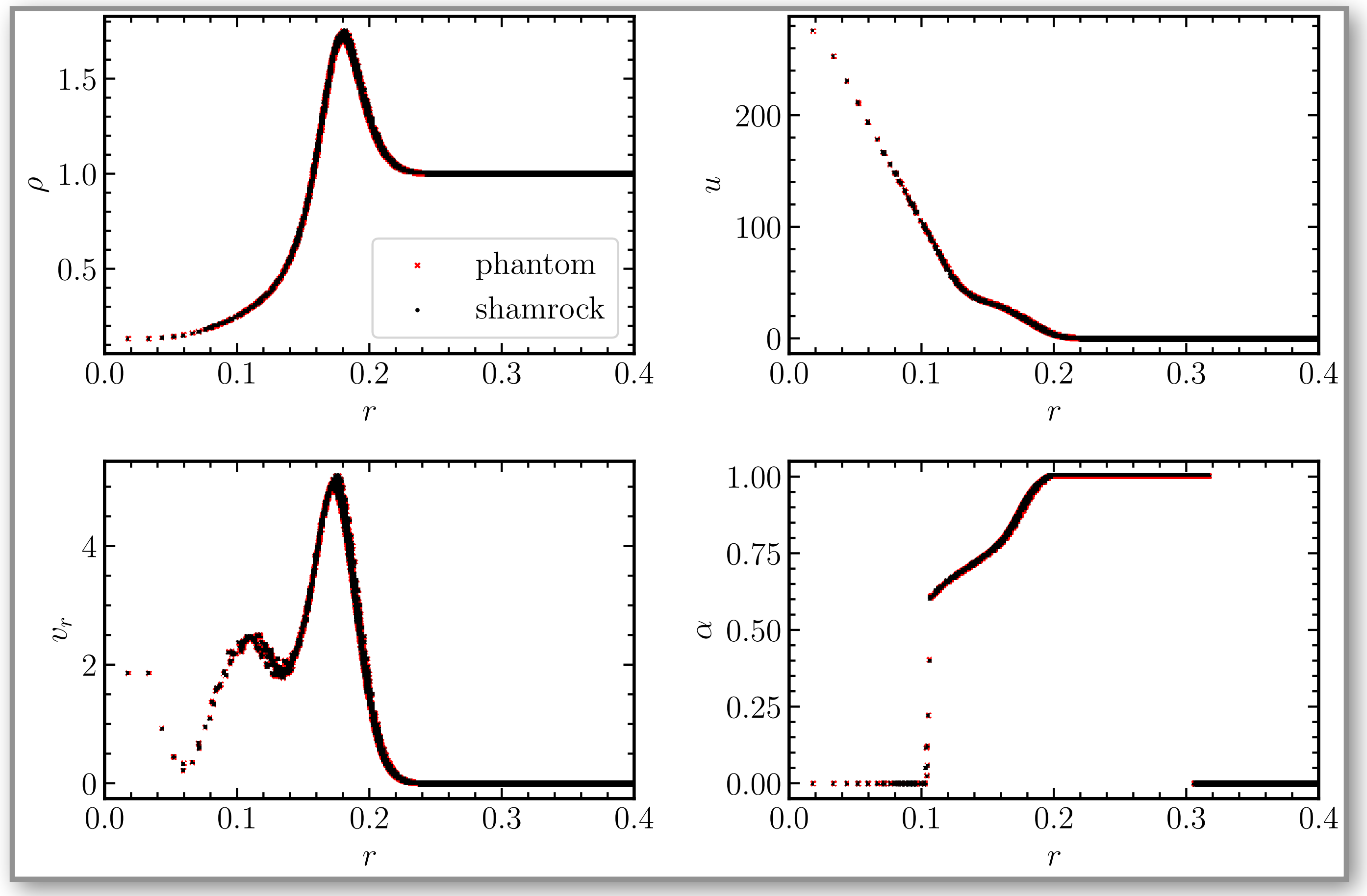



**Reproduce Phantom solver !**

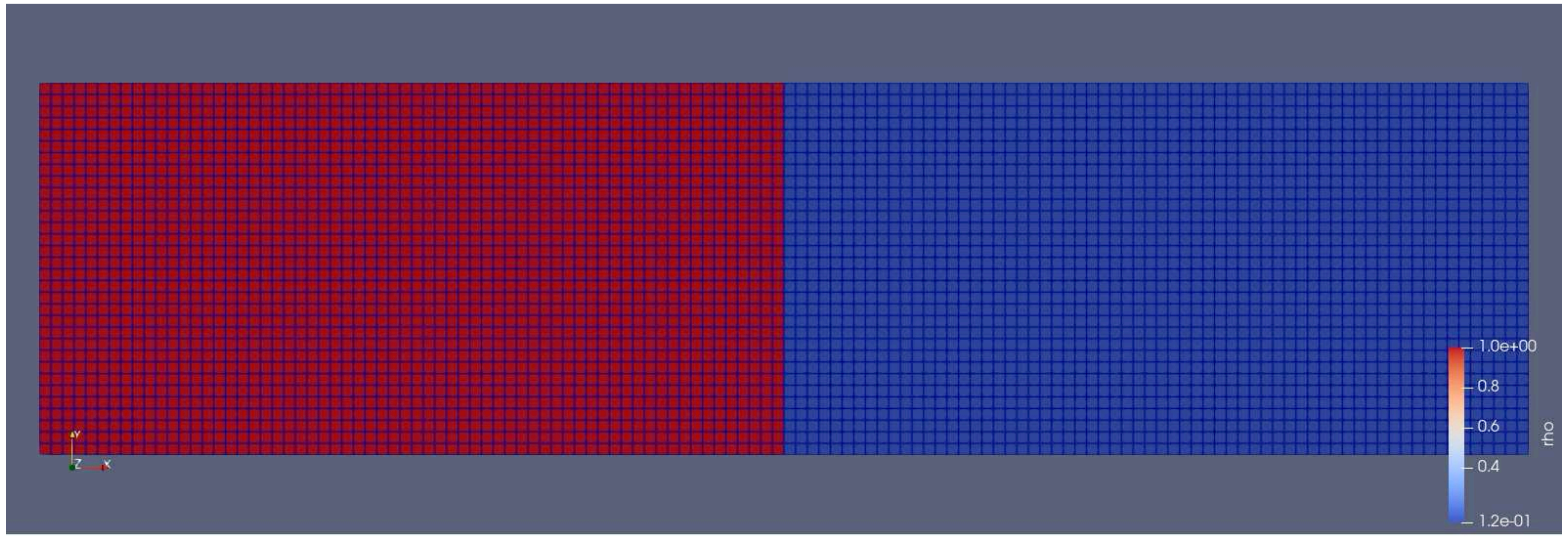
# But also AMR & Zeus



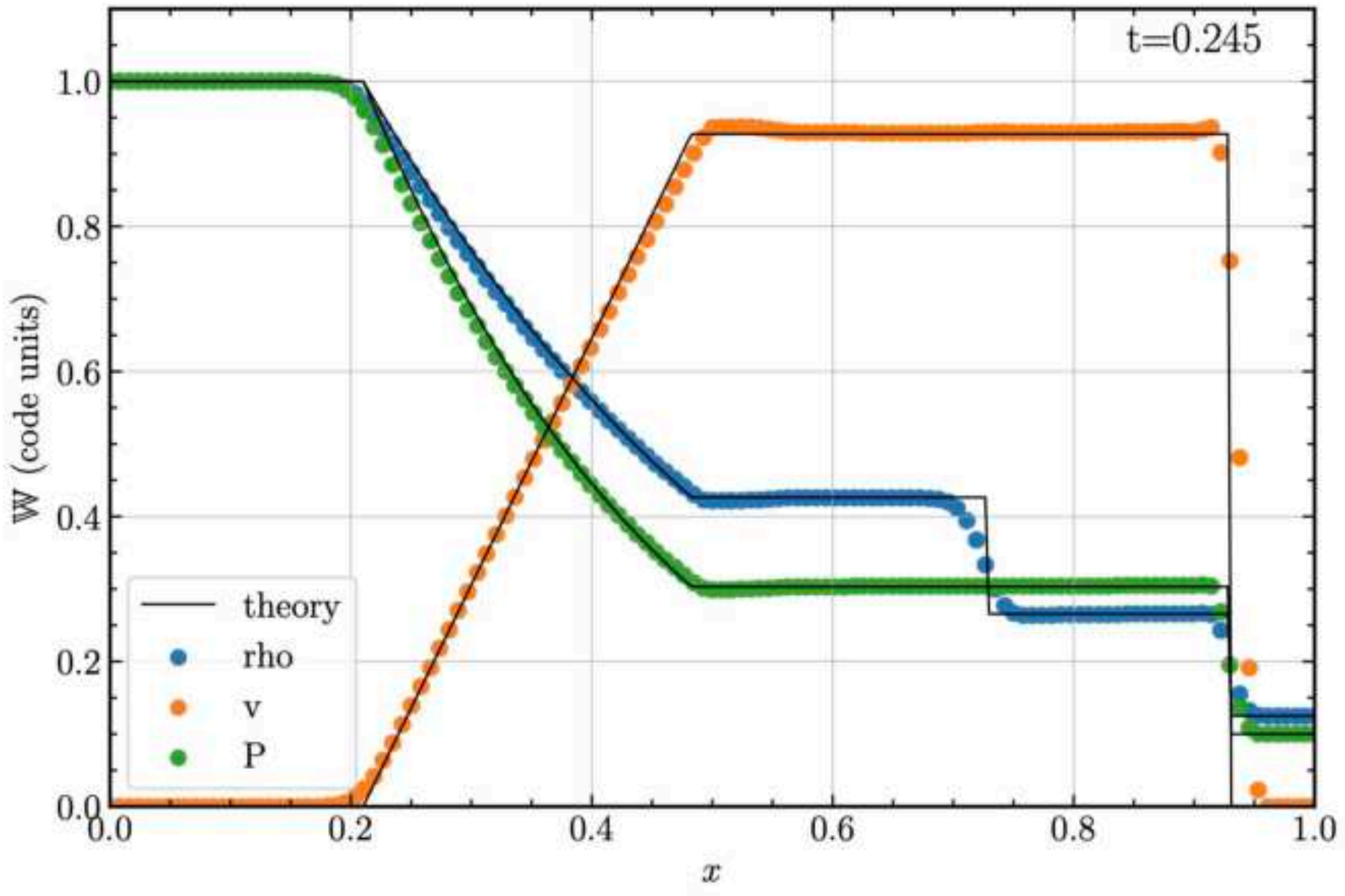
## SPH (Phantom kind)

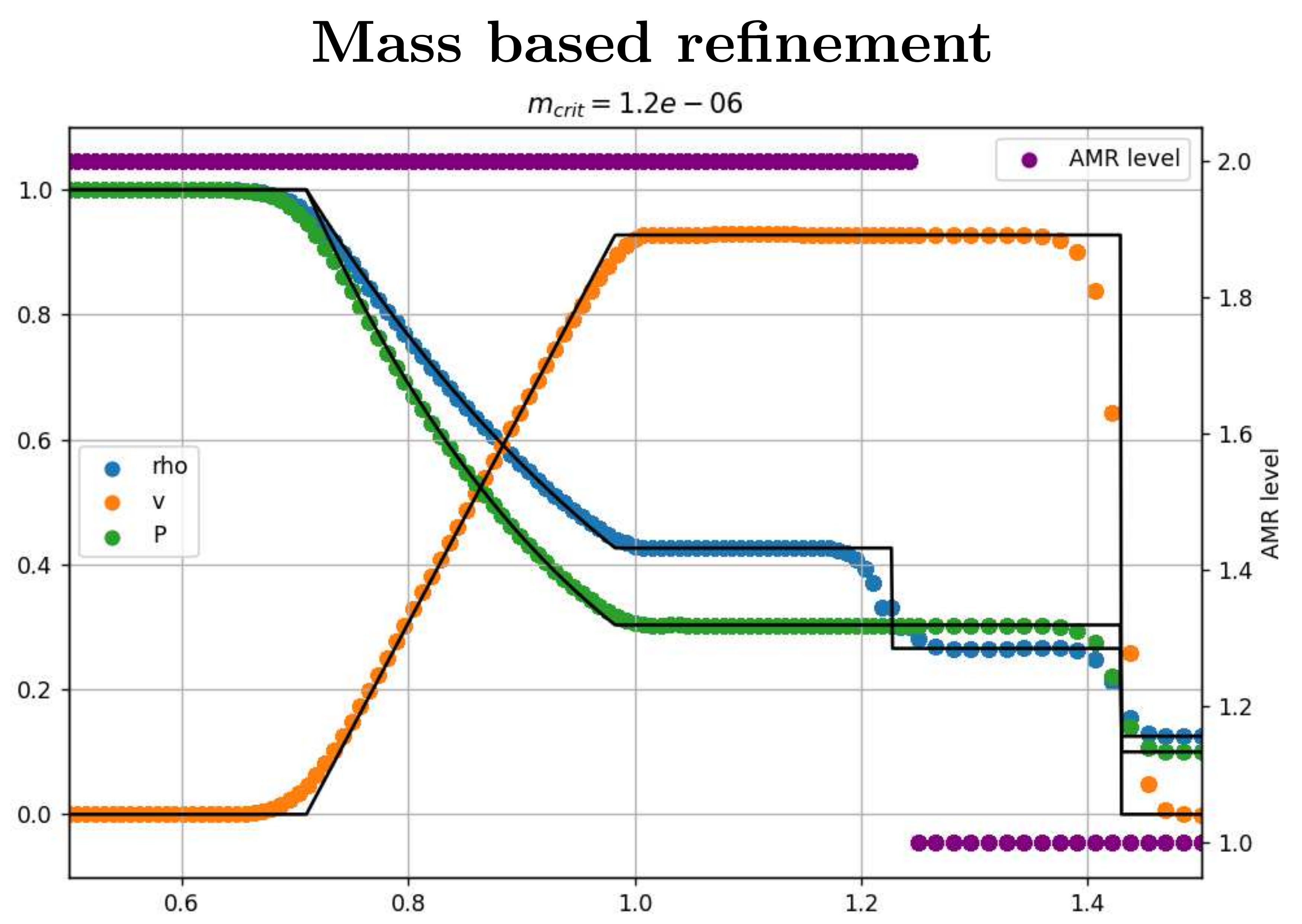



## Finite volume (RAMSES kind)




## Finite elements (Zeus kind)



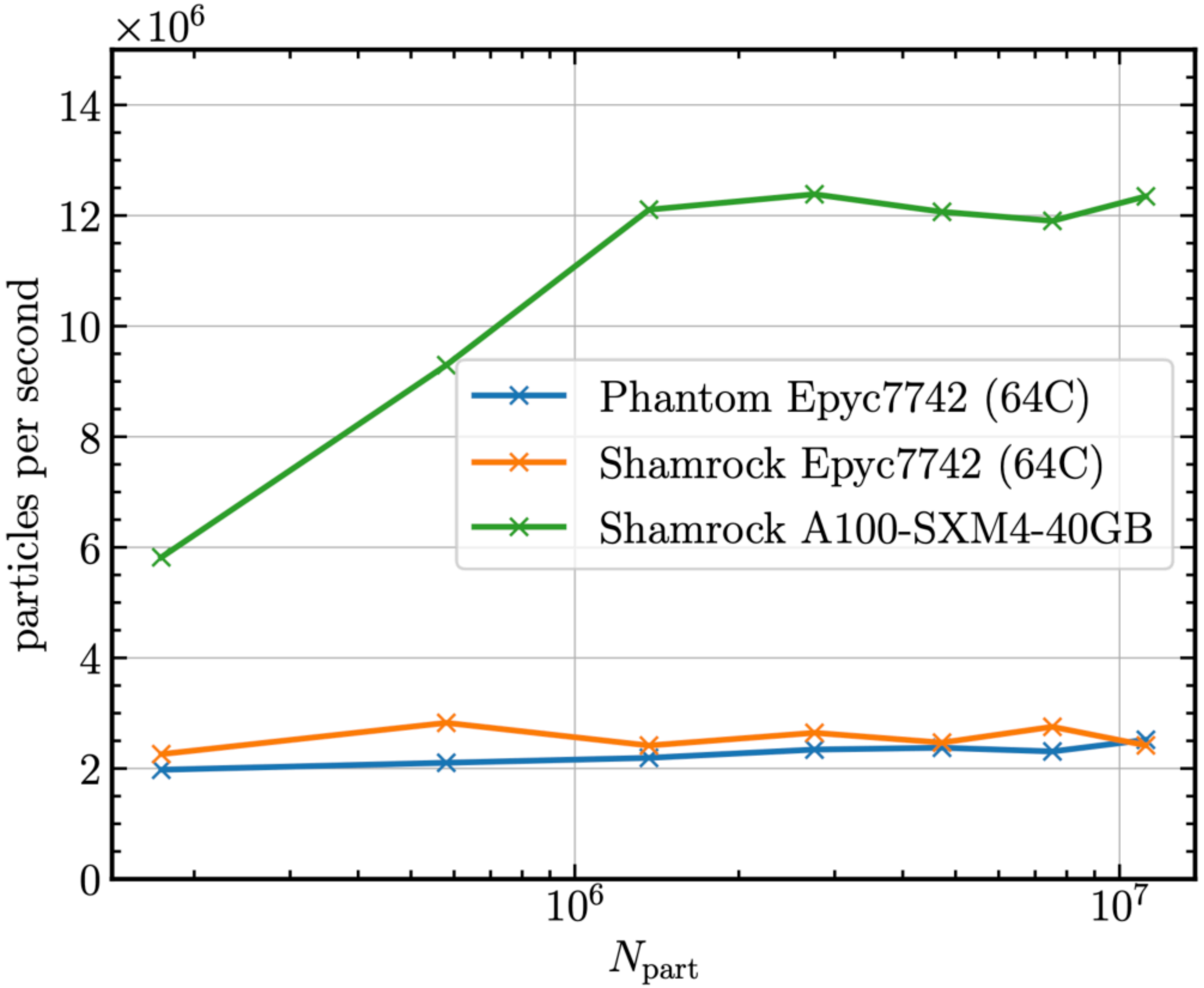



Performance !!!!

(Measured in part/sec)

# Single GPU/CPU perf

**Sedov-blast wave:**




**Latest benchmarks**

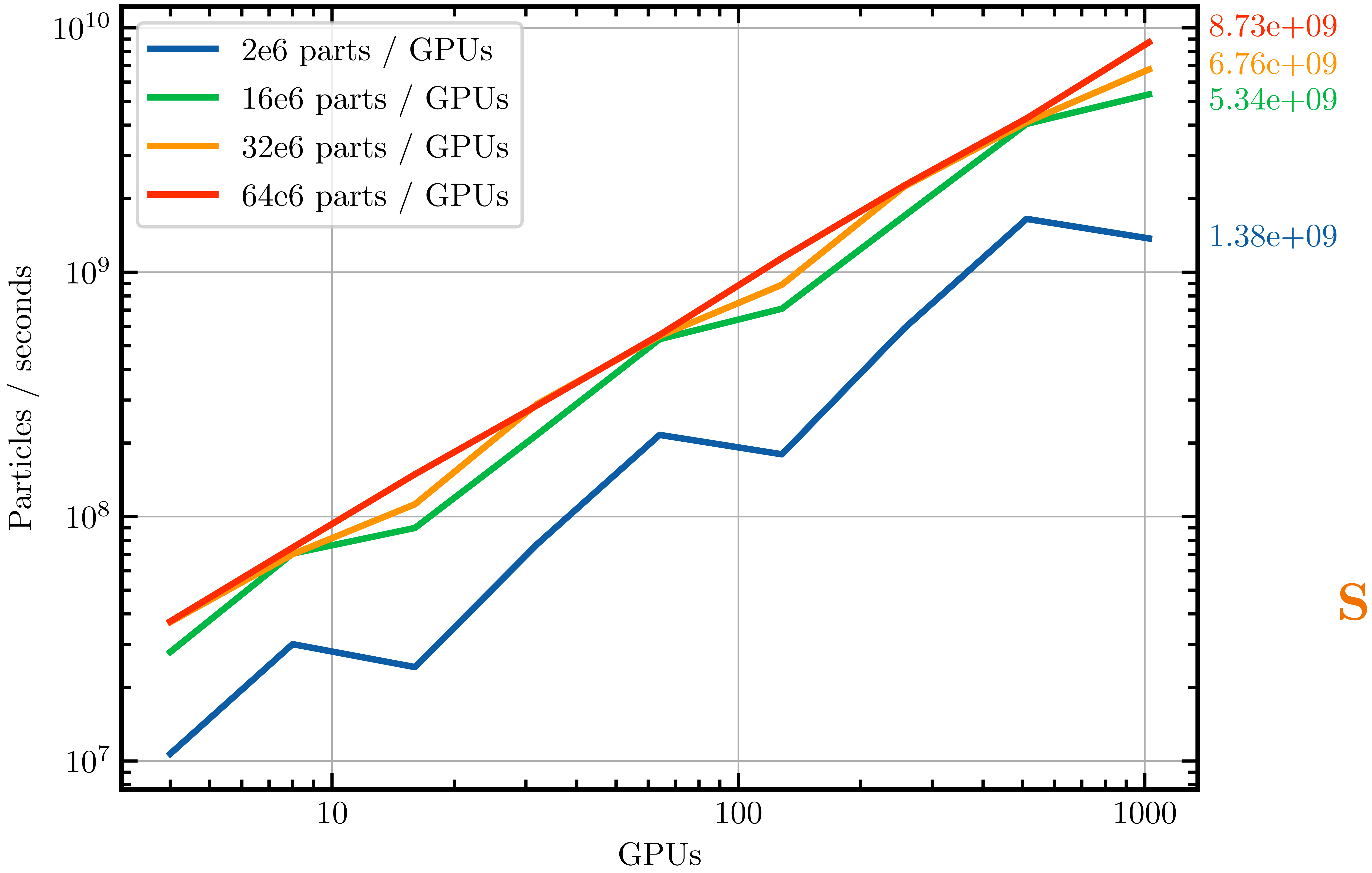
$\simeq 15\text{Mpart/sec}$

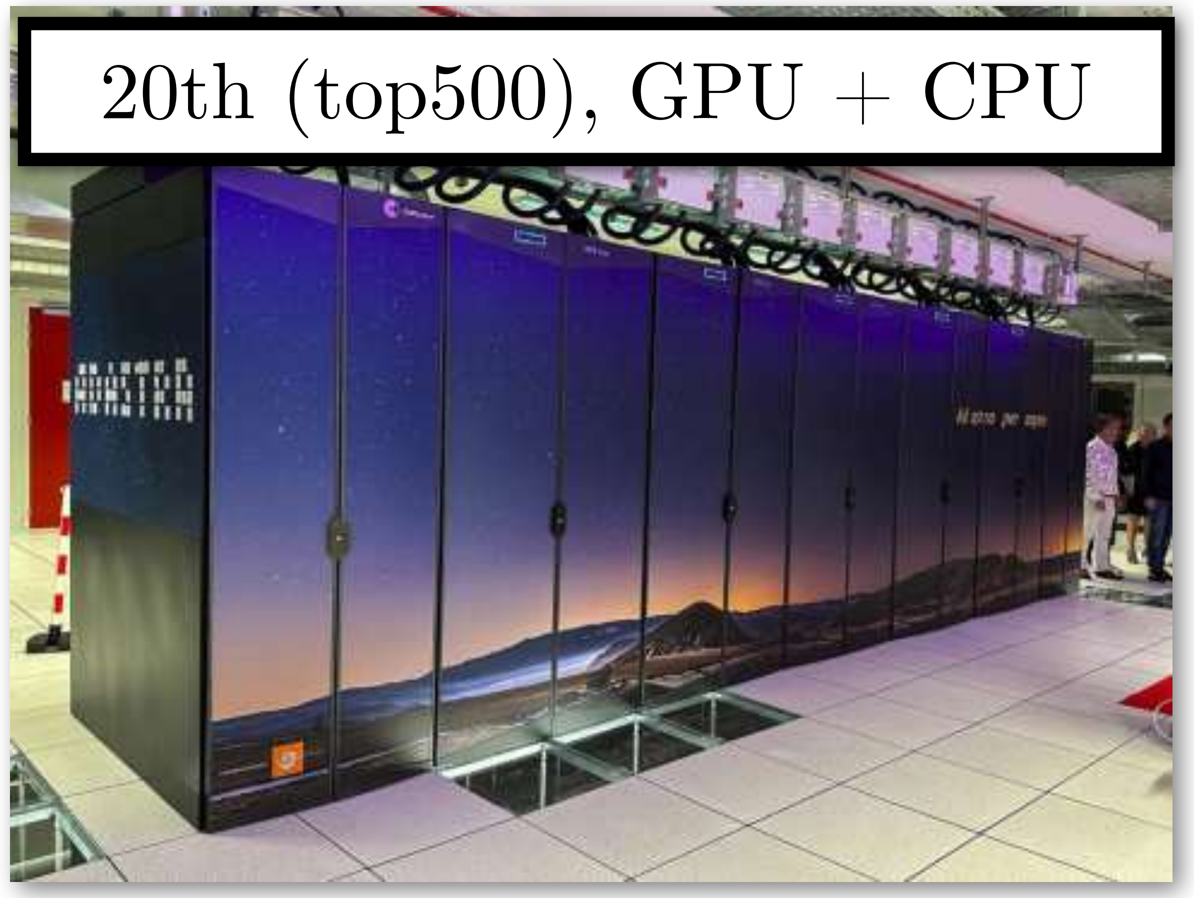
$\Longrightarrow$ **Acceleration by a factor 7/8 CPU⇒GPU**

$\Longrightarrow$ **Same power consumption !!!**

# Multi-GPU






**Shamrock :**

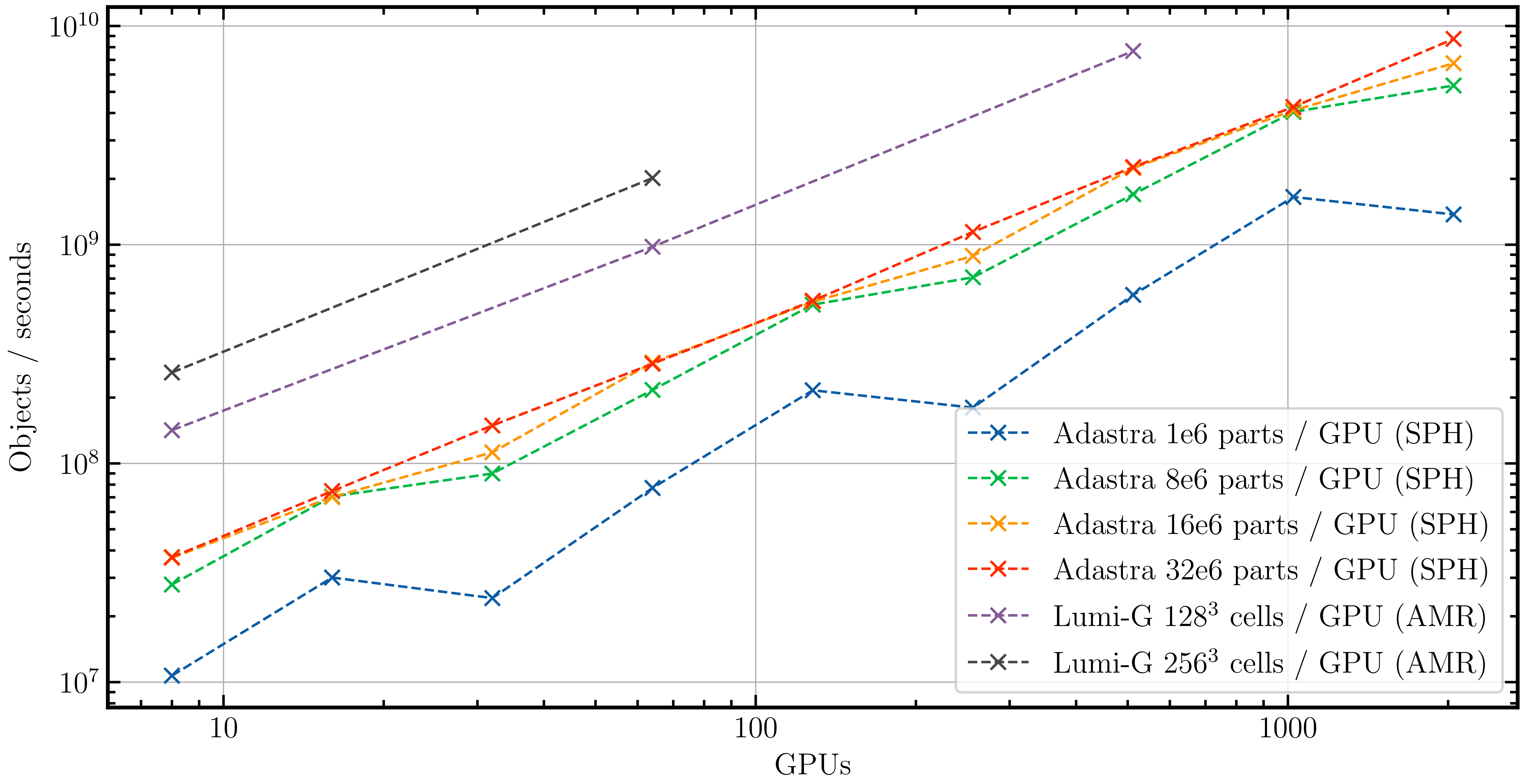
- **65G particles** $(4000^3)$
- **9G part / seconds**
- **7 sec / iterations**
- **1024 GPU (MI250x)**
- **92% parallel efficiency**

**Reference :**

Same test with Phantom SPH = 2e6 part / seconds

$\longrightarrow$ **x4500 speedup !**

# But also with Ramses




CAUTION WORK IN PROGRESS

**84%** parallel efficiency on LUMI **before optimisation**

**Next: Optimizing the solver to get at least a factor of 10**

# On the road to Exascale


Top 500 list :

| Rank | System | Cores | Rmax (PFlop/s) | Rpeak (PFlop/s) | Power (kW) |
|---|---|---|---|---|---|
| 1 | **El Capitan** - HPE Cray EX255a, AMD [illegible] Gen [illegible]C 24C 1.8GHz, AMD Instinct MI300A, Slingshot[illegible]OSS, HPE DOE/NNSA/LLNL United States | 11,039,616 | 1,742.00 | 2,746.38 | 29,581 |
| 2 | **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States | 9,066,176 | 1,353.00 | 2,055.72 | 24,607 |
| 3 | **Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States | 9,264,128 | 1,012.00 | 1,980.01 | 38,698 |
| 4 | **Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States | 2,073,600 | 561.20 | 846.84 | |
| 5 | **HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy | 3,143,520 | 477.90 | 606.97 | 8,461 |

Military

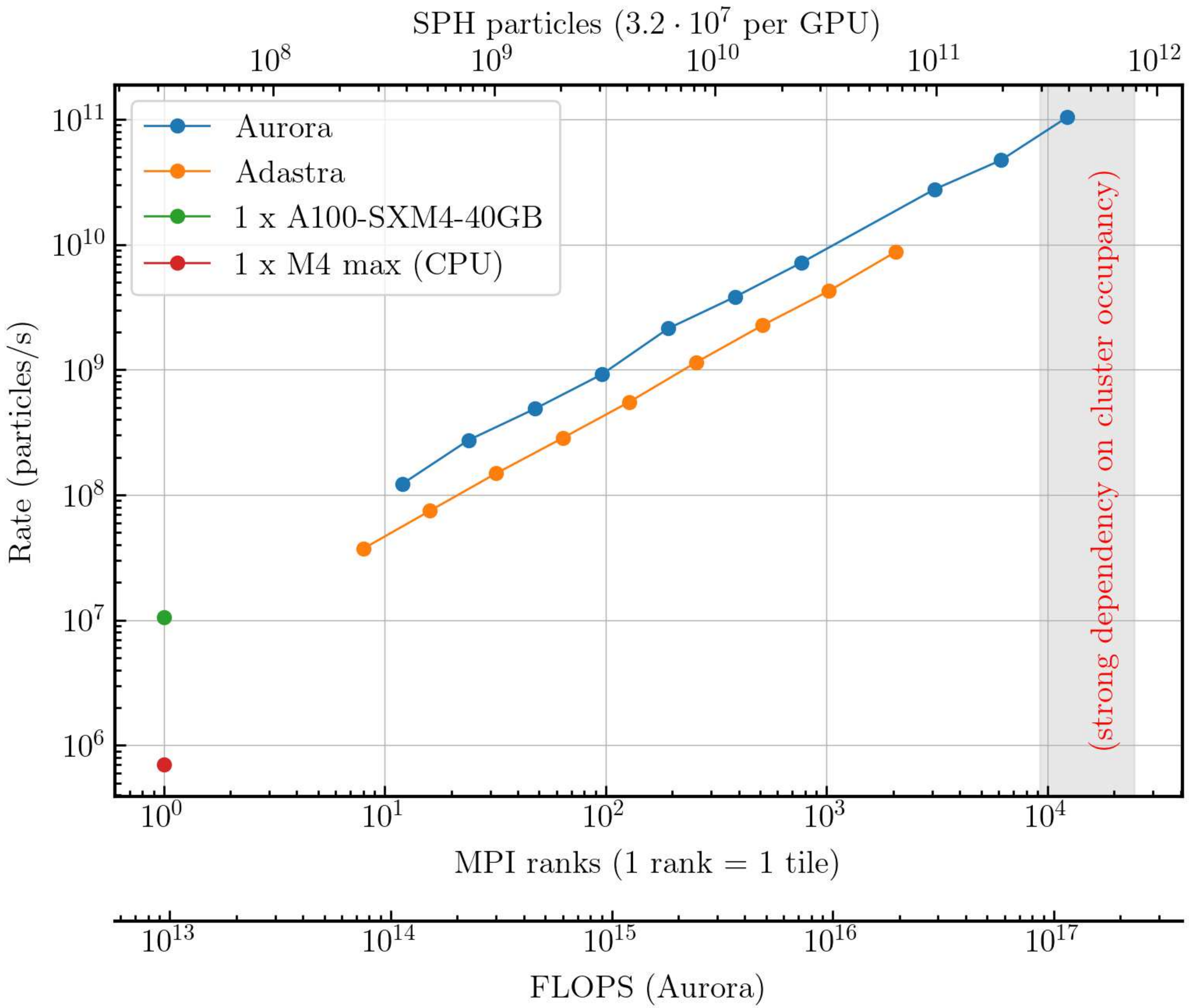
We have access to this one


120 000 GPUs

1 Exaflop

**Note:** getting GPU time is **really** easy

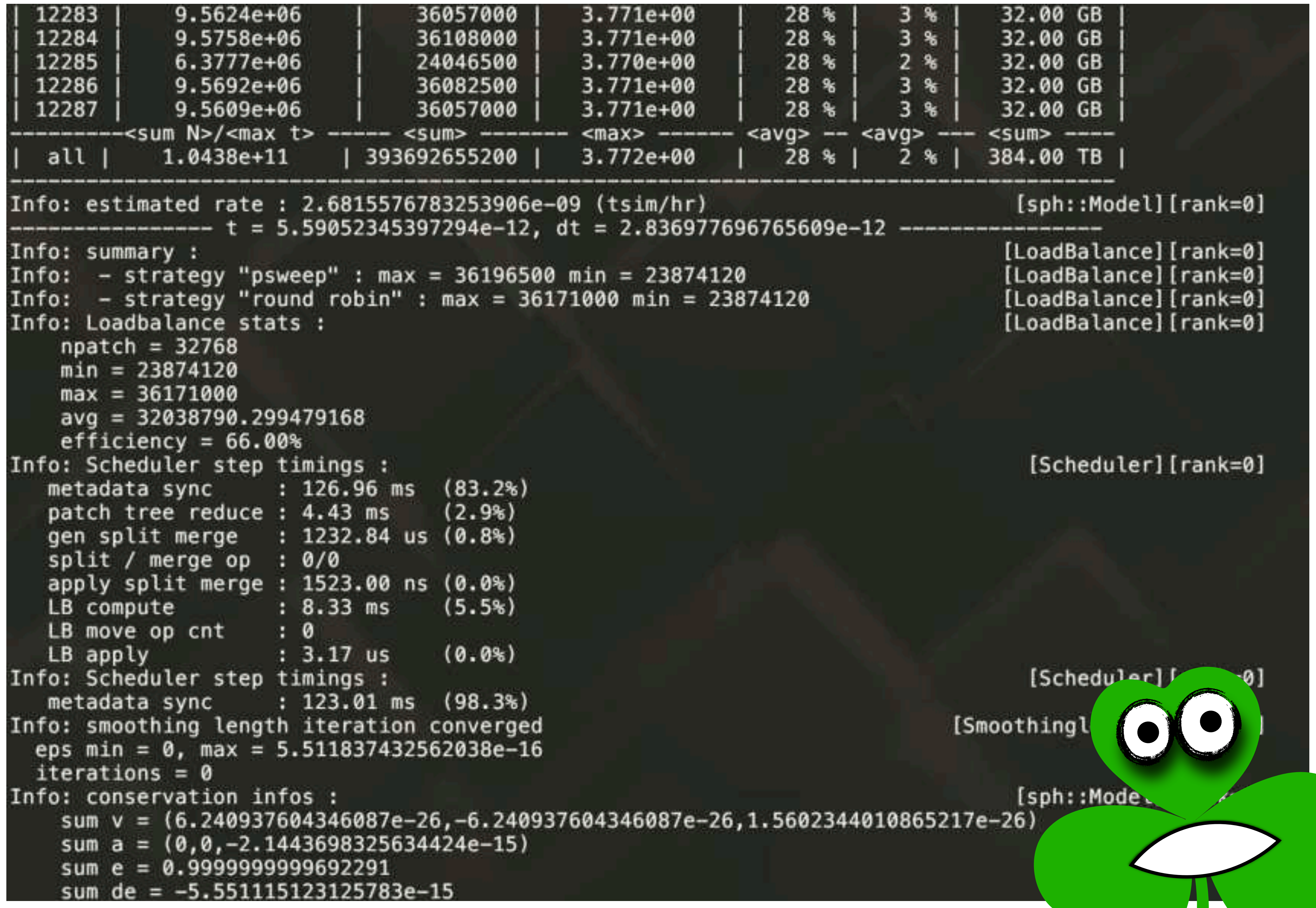
# On the road to Exascale


→ **x50000 speedup !**

**Shamrock :**

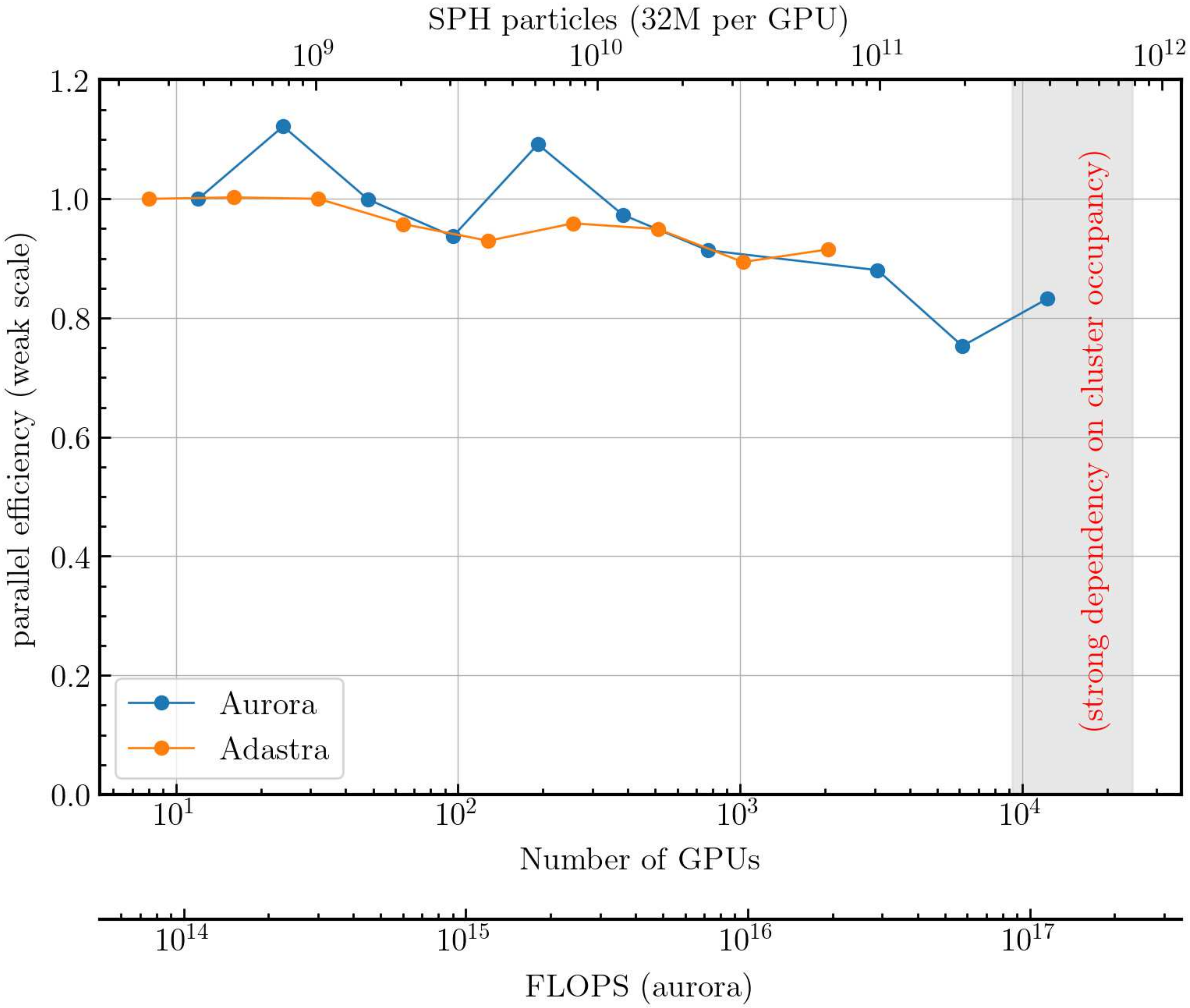
- **393.692.655.200 particles ($7324^3$)**
- **$10^{11}$ part / seconds**
- **4 sec / iterations**
- **12288 GPU (Intel PVC tiles)**



```
| 12283 |     9.5624e+06     |    36057000 |  3.771e+00     |   28 % |   3 % |   32.00 GB |
| 12284 |     9.5758e+06     |    36108000 |  3.771e+00     |   28 % |   3 % |   32.00 GB |
| 12285 |     6.3777e+06     |    24046500 |  3.770e+00     |   28 % |   2 % |   32.00 GB |
| 12286 |     9.5692e+06     |    36082500 |  3.771e+00     |   28 % |   3 % |   32.00 GB |
| 12287 |     9.5609e+06     |    36057000 |  3.771e+00     |   28 % |   3 % |   32.00 GB |
---------<sum N>/<max t> ----- <sum> -------- <max> ------- <avg> -- <avg> --- <sum> ----
|  all |    1.0438e+11    | 393692655200 |  3.772e+00     |   28 % |   2 % |  384.00 TB |
------------------------------------------------------------------------------------------
Info: estimated rate : 2.6815576783253906e-09 (tsim/hr)                     [sph::Model][rank=0]
---------------- t = 5.59052345397294e-12, dt = 2.836977696765609e-12 ----------------
Info: summary :                                                        [LoadBalance][rank=0]
Info:   - strategy "psweep" : max = 36196500 min = 23874120            [LoadBalance][rank=0]
Info:   - strategy "round robin" : max = 36171000 min = 23874120       [LoadBalance][rank=0]
Info: Loadbalance stats :                                              [LoadBalance][rank=0]
    npatch = 32768
    min = 23874120
    max = 36171000
    avg = 32038790.299479168
    efficiency = 66.00%
Info: Scheduler step timings :                                           [Scheduler][rank=0]
   metadata sync      : 126.96 ms  (83.2%)
   patch tree reduce : 4.43 ms     (2.9%)
   gen split merge    : 1232.84 us (0.8%)
   split / merge op   : 0/0
   apply split merge : 1523.00 ns (0.0%)
   LB compute         : 8.33 ms    (5.5%)
   LB move op cnt     : 0
   LB apply           : 3.17 us    (0.0%)
Info: Scheduler step timings :                                           [Scheduler][
   metadata sync      : 123.01 ms  (98.3%)
Info: smoothing length iteration converged                             [Smoothingl
  eps min = 0, max = 5.511837432562038e-16
  iterations = 0
Info: conservation infos :                                              [sph::Mode
    sum v = (6.240937604346087e-26,-6.240937604346087e-26,1.5602344010865217e-26)
    sum a = (0,0,-2.1443698325634424e-15)
    sum e = 0.9999999999692291
    sum de = -5.551115123125783e-15
```

# On the road to Exascale




**Next step is Exascale !**

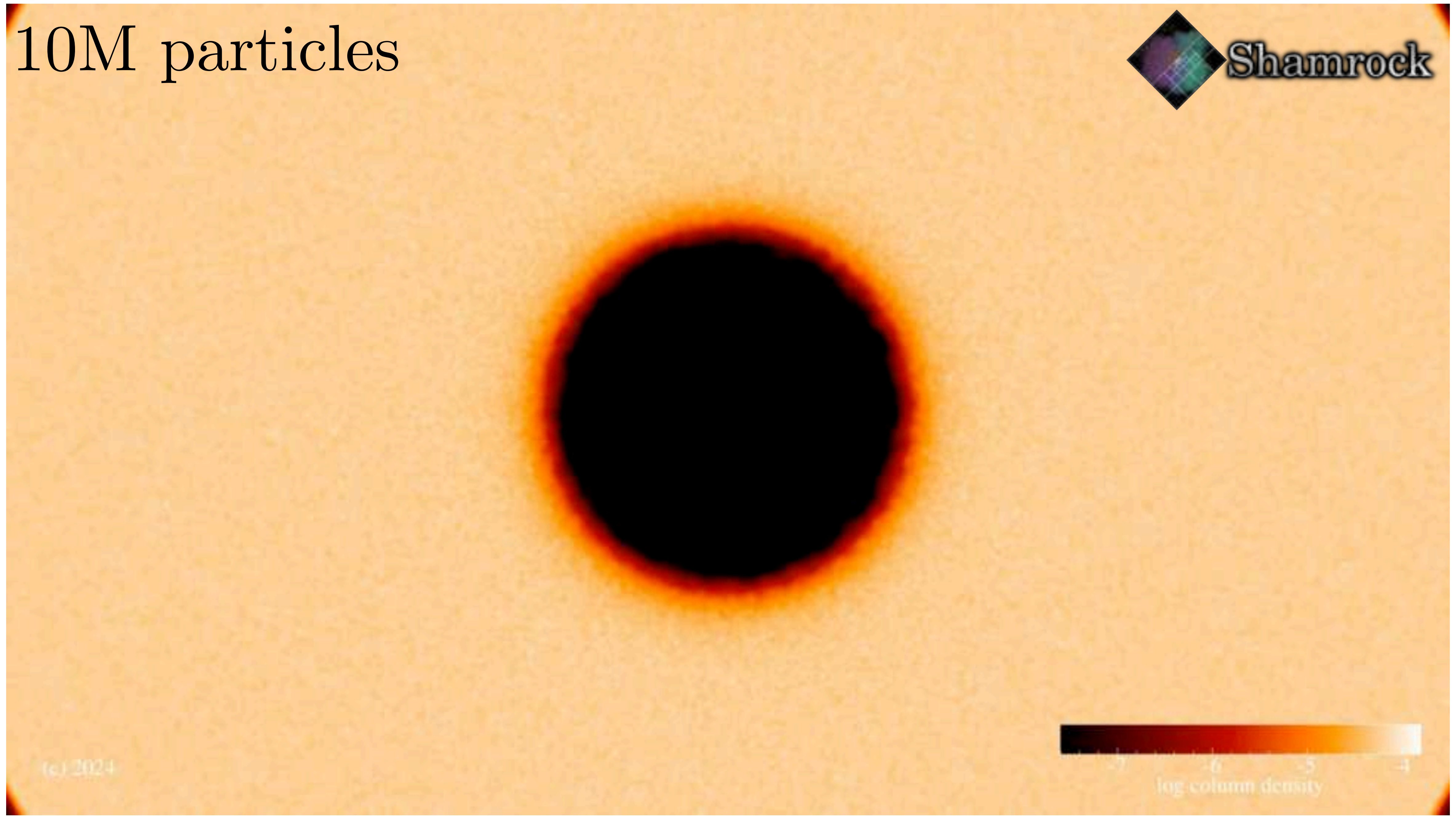
# Exemple of a simulation

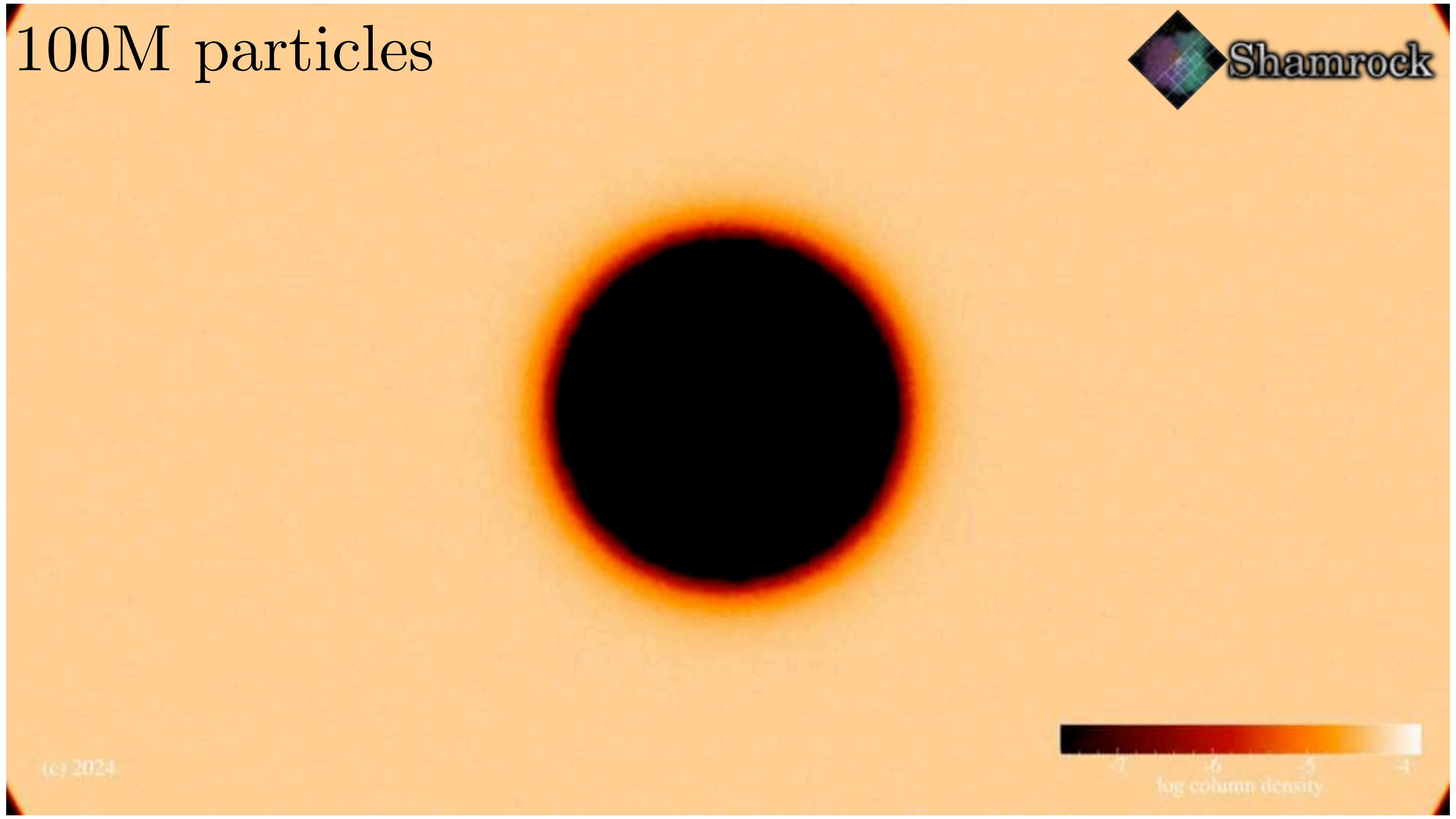
# Exemples : circum-binary discs



- Soft limit around 10-100M particles

But new physics at 1G particles



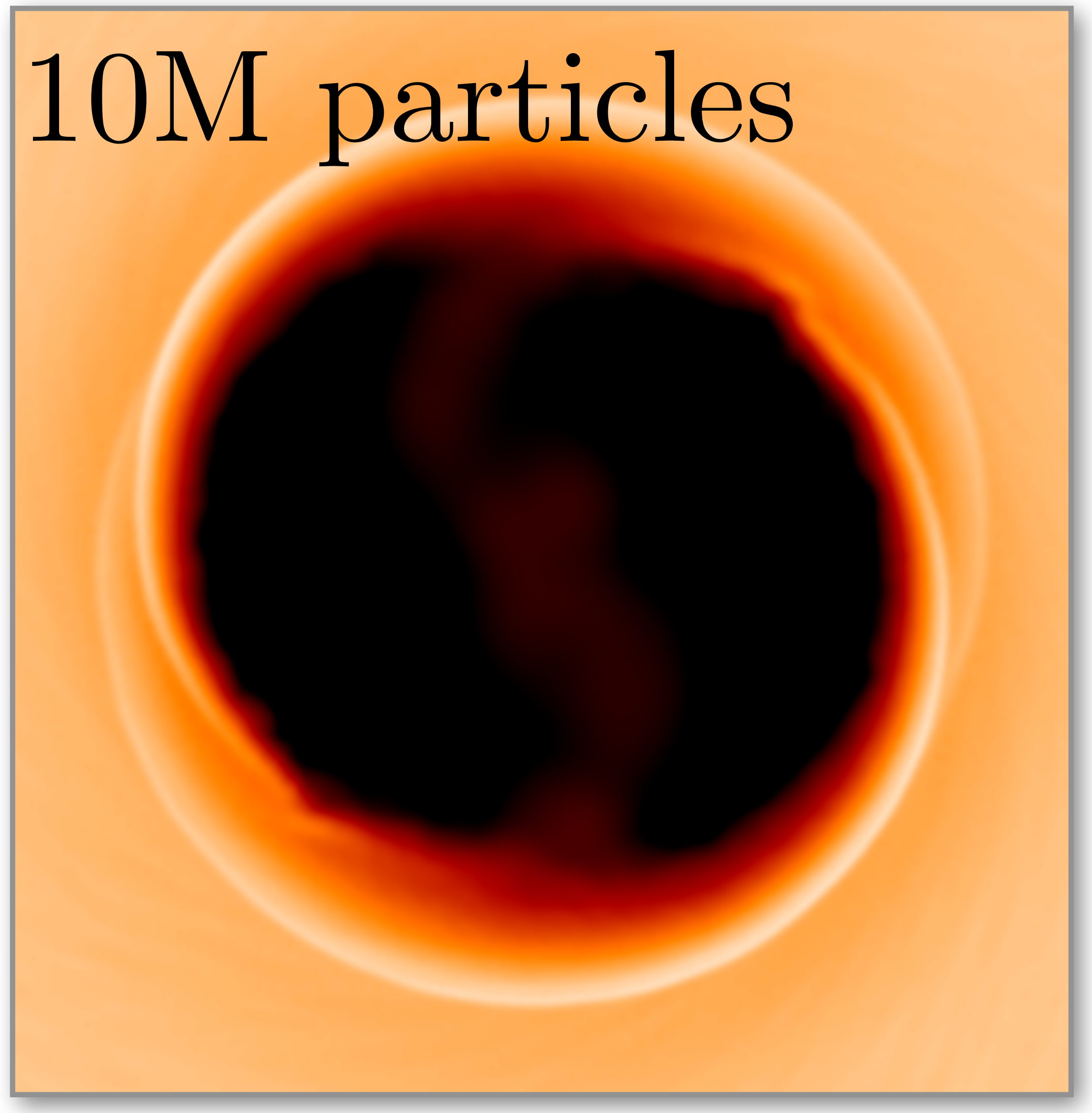



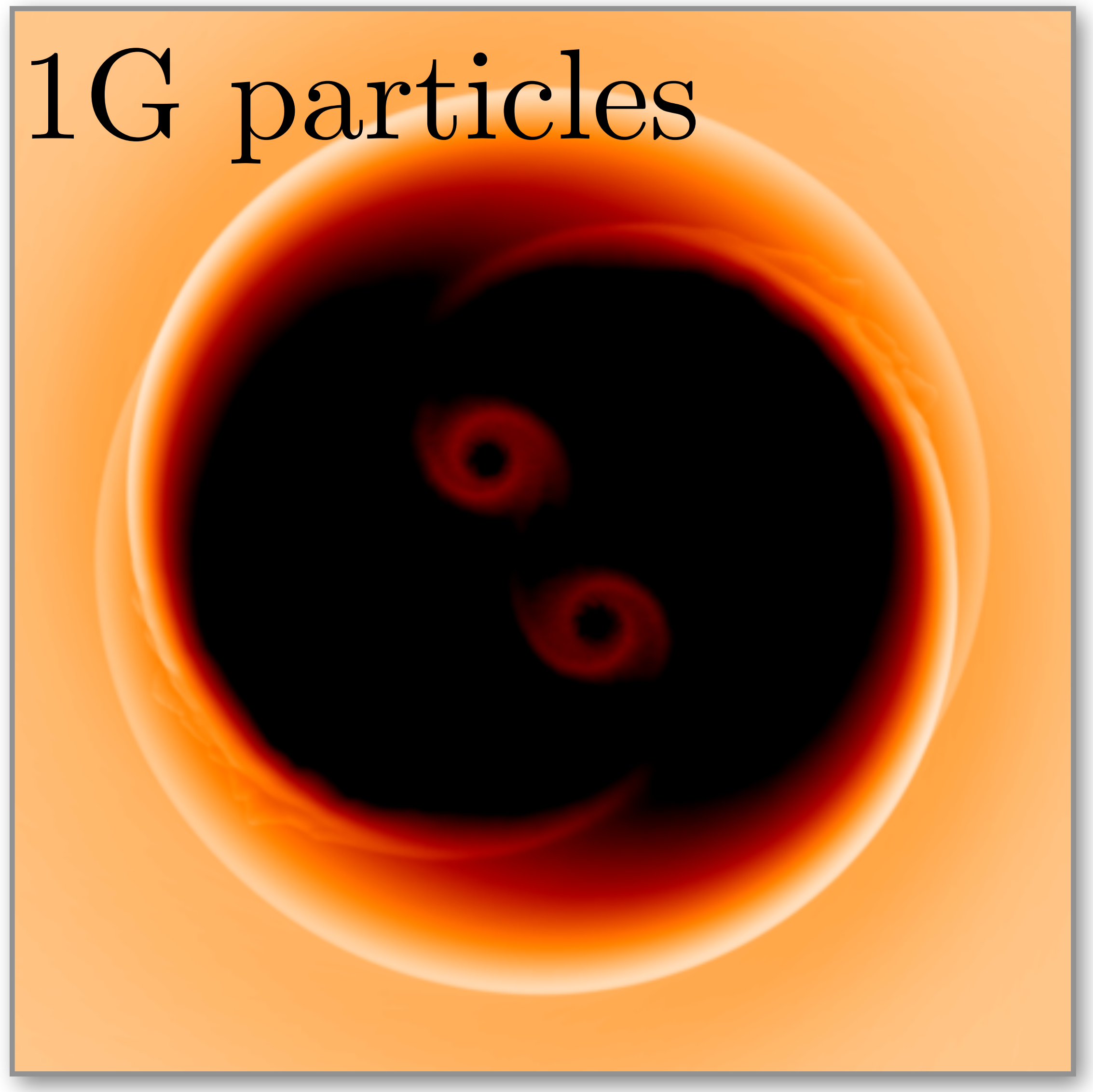

**Need more resolution !**

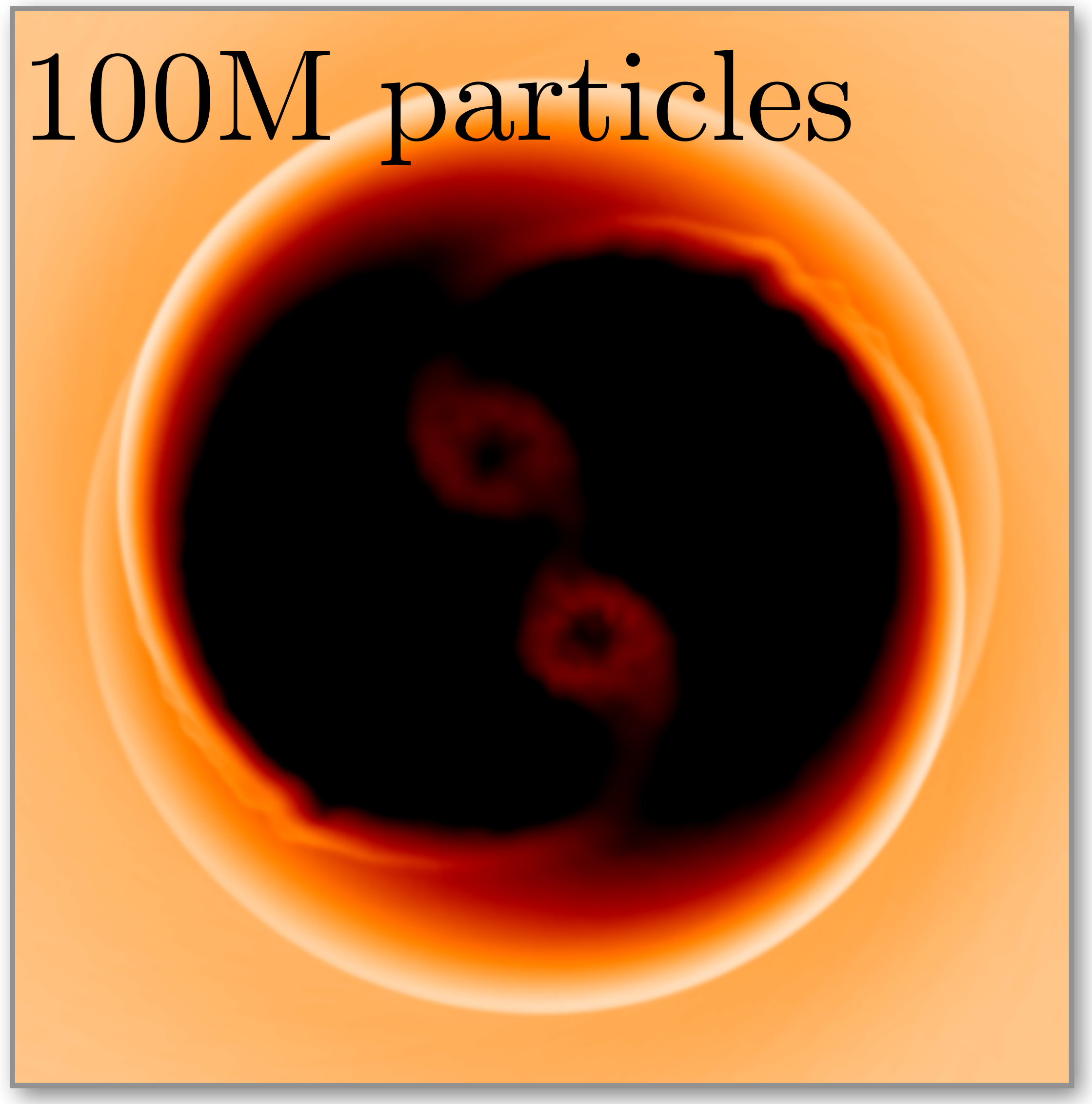

1G particles
Shamrock
-7
-6
-5
-4
log column density
(c) 2024

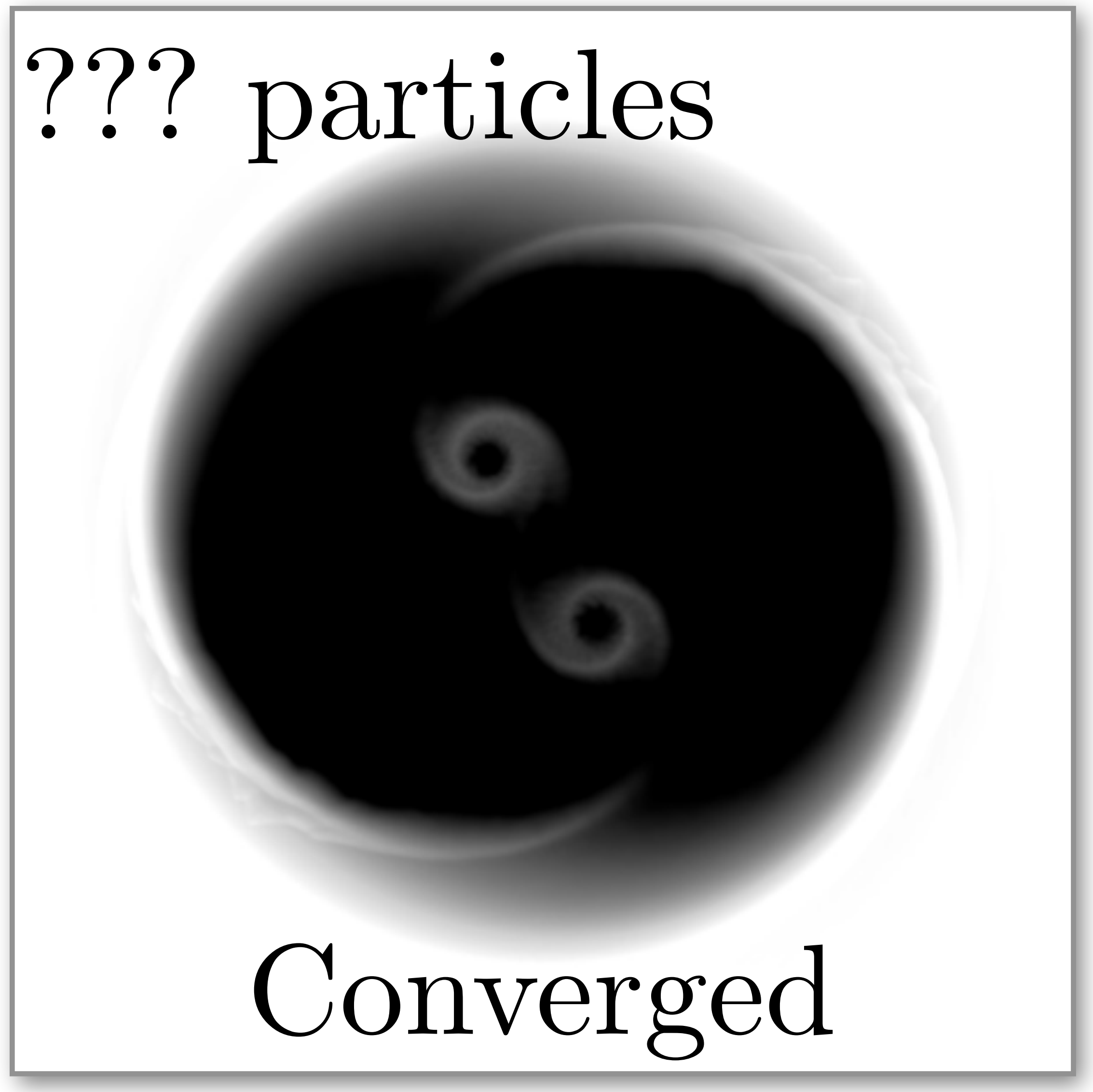
# Accretion rate



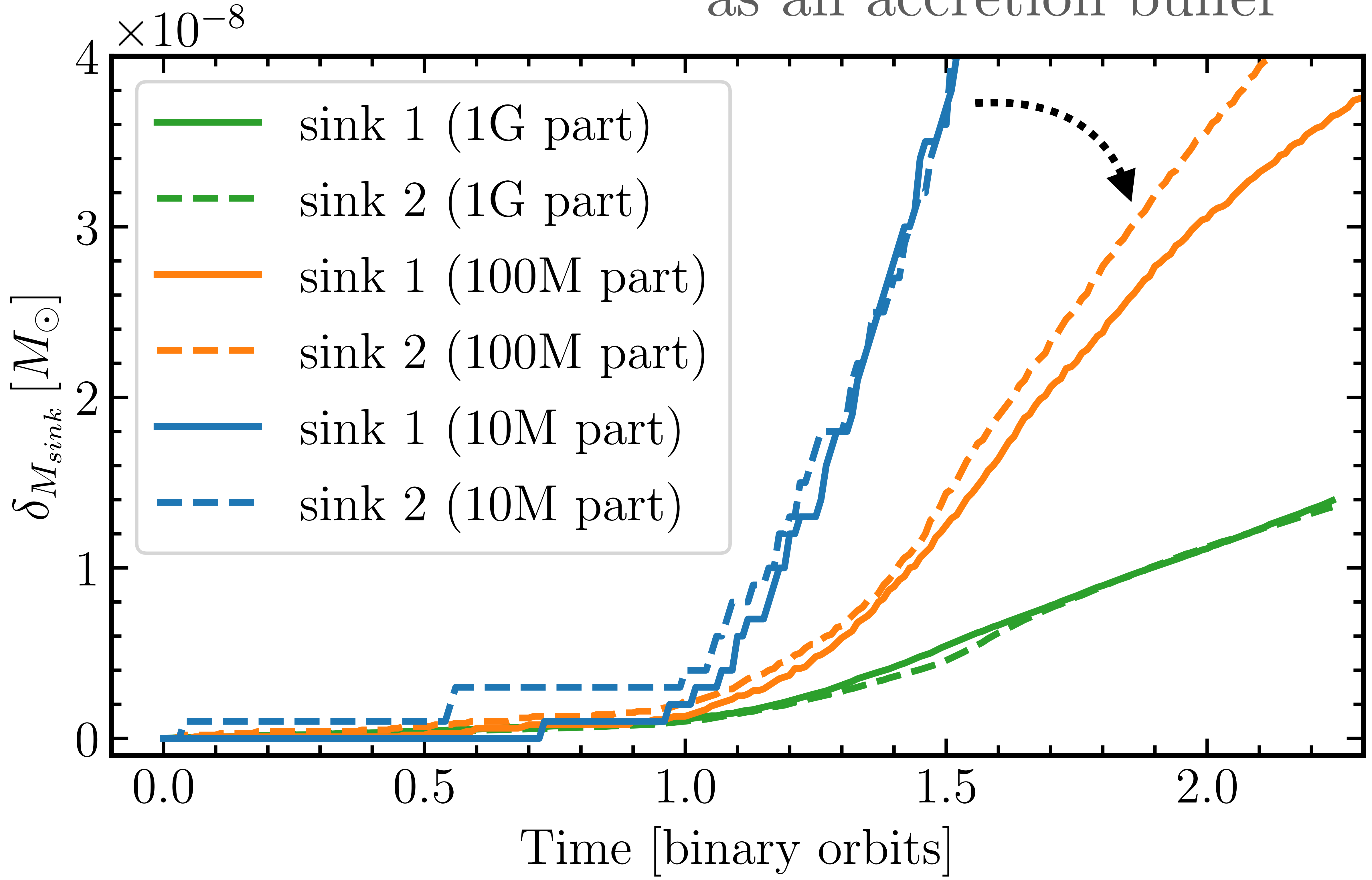


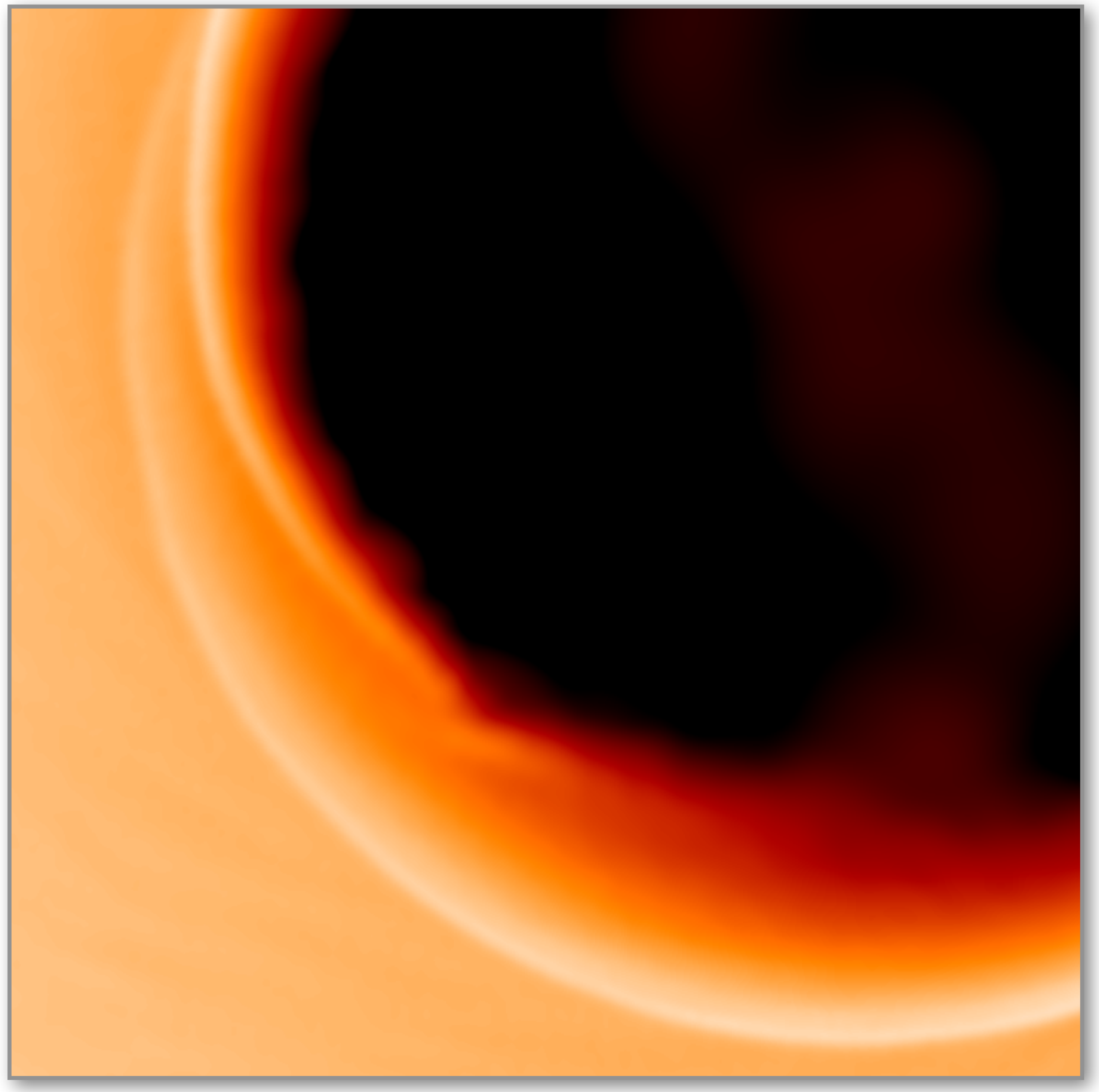

The circumstellaire discs act as an accretion buffer



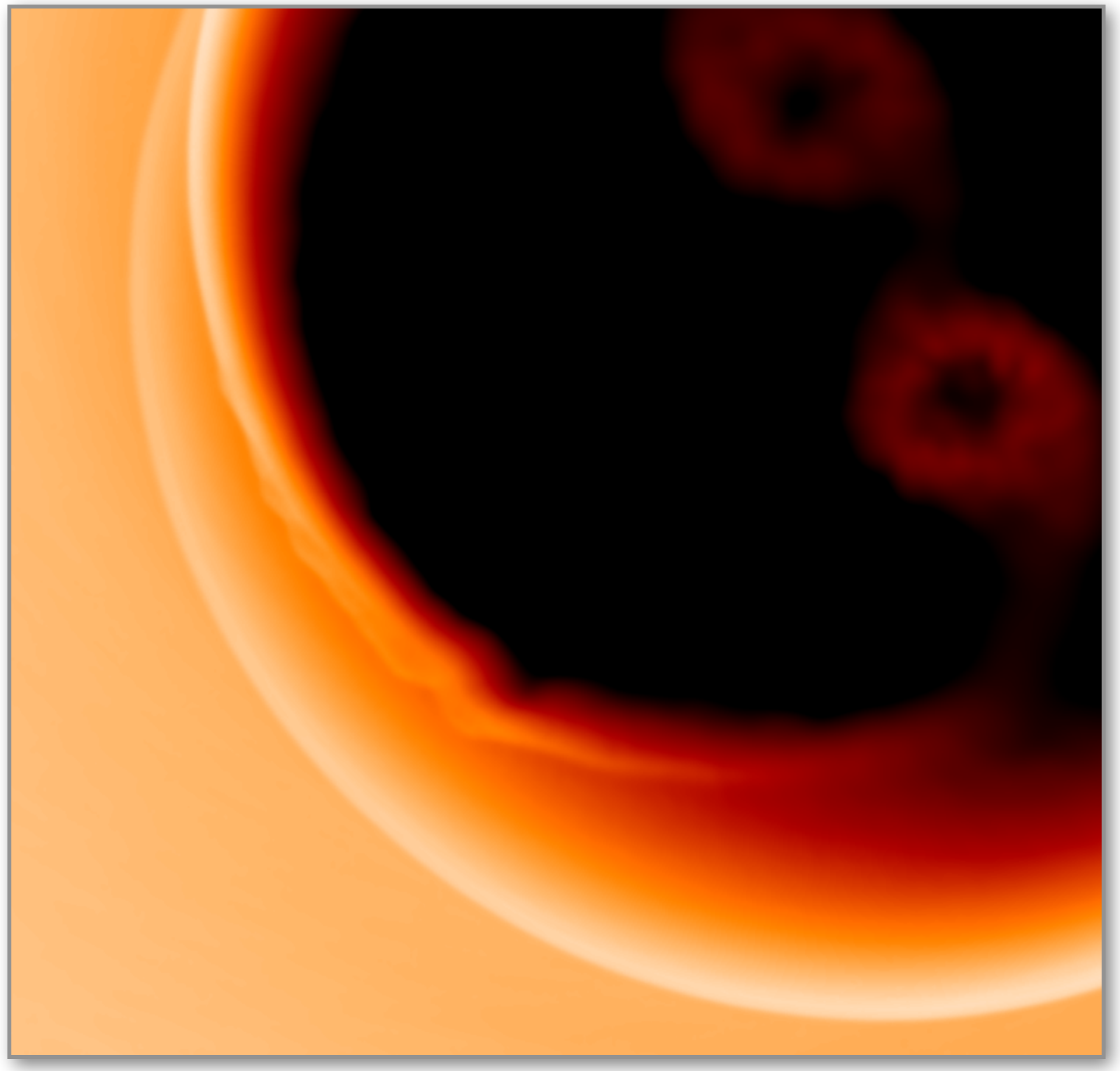
# Résolution verticale



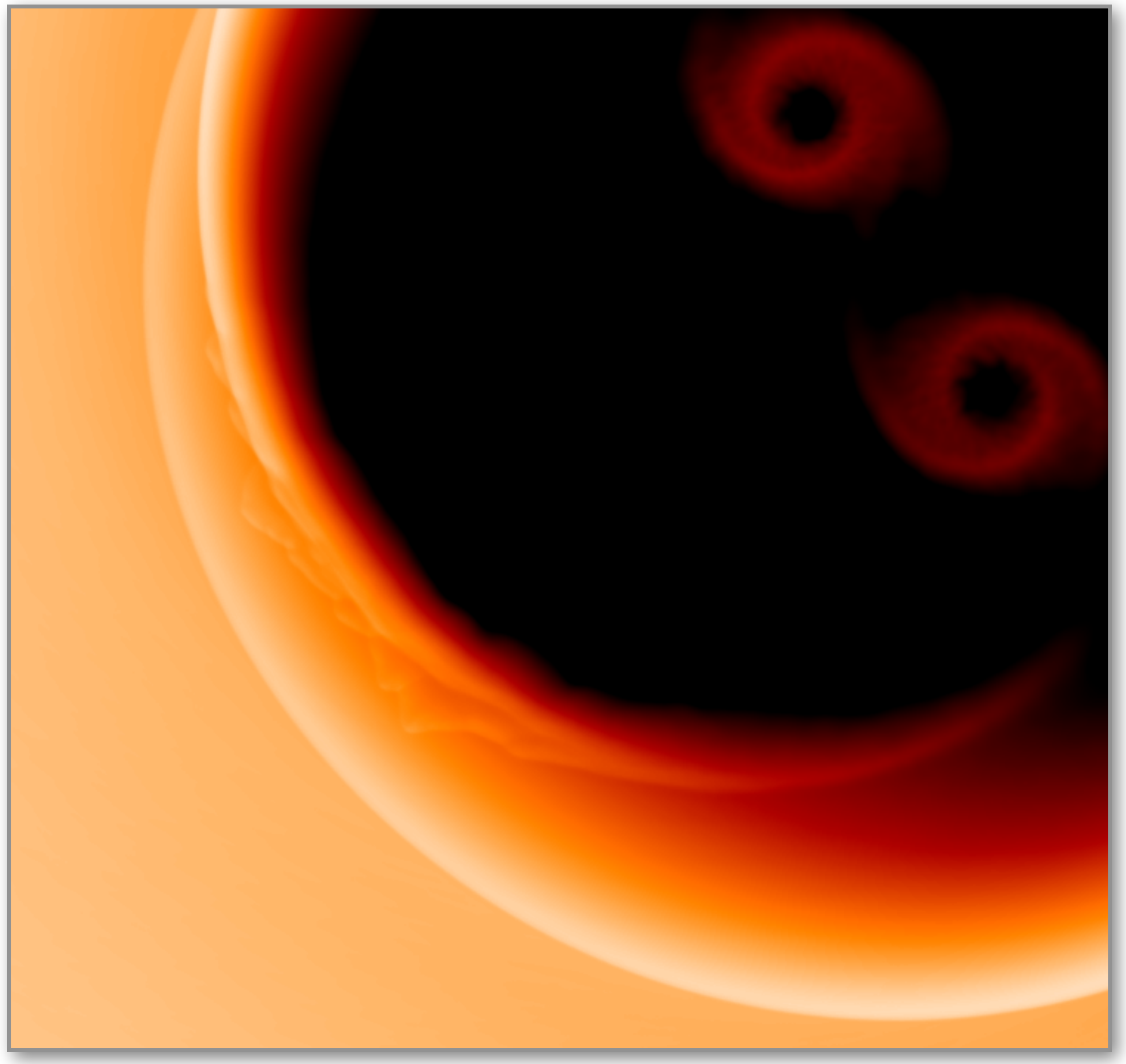
10M particles



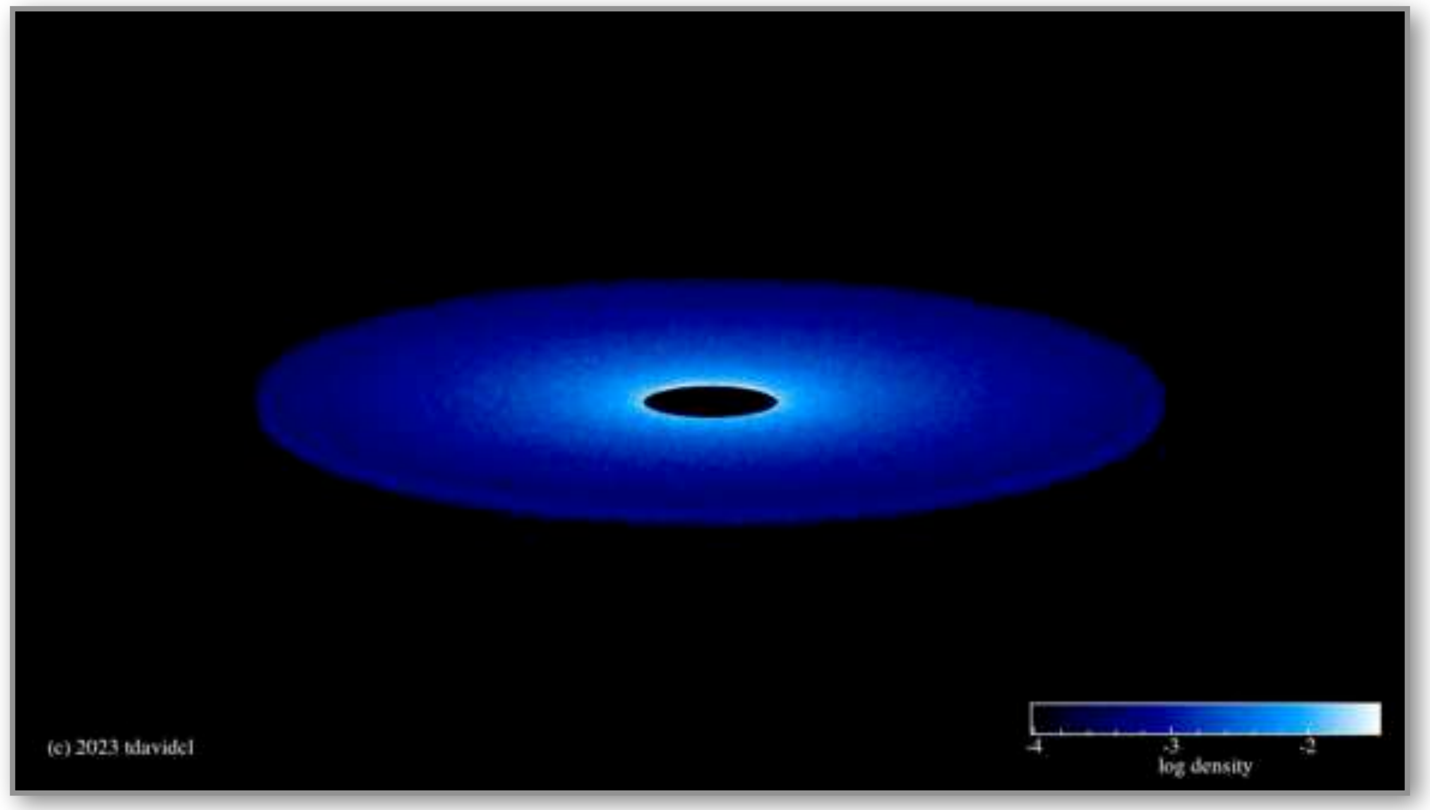
100M particles



1G particles



1G particules : <u>125</u> particules / H $\Rightarrow$ Enough for VSI, SI, parametric instability, ...




See Yona's talk
for discs


**Where do we go from there ????**

See guillaume talk

## SPH solver (production ready) :

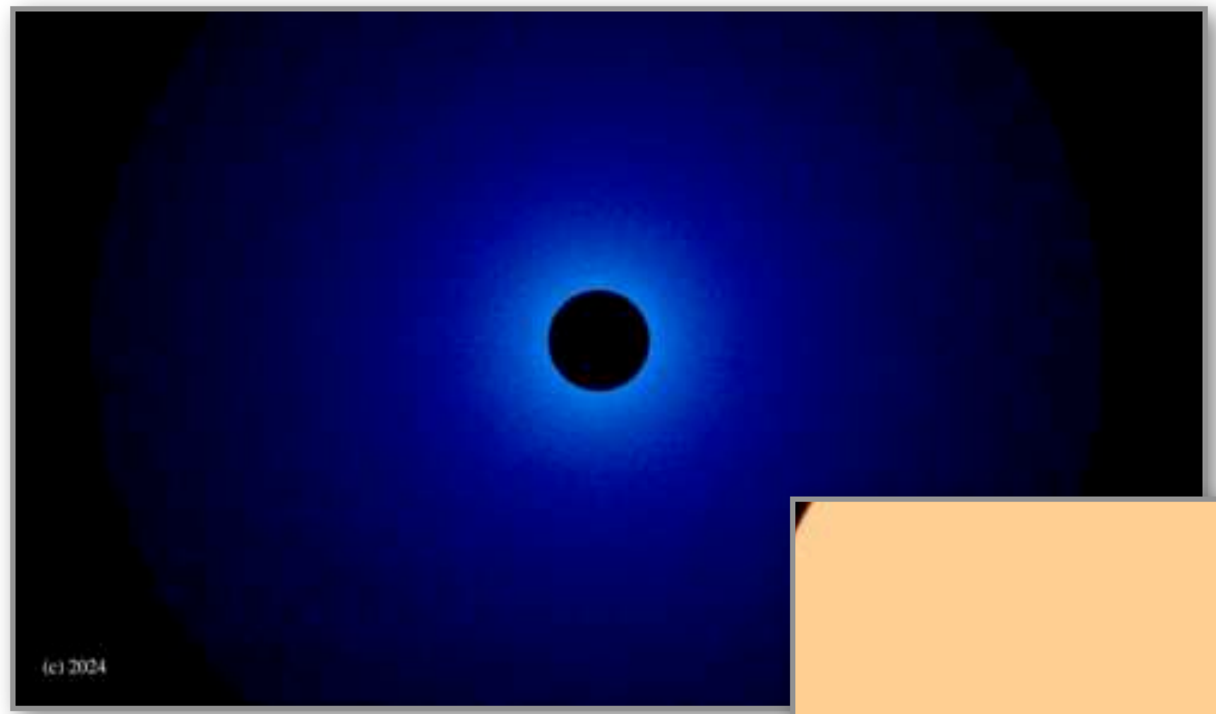
- Hydro ✅
- Discs ✅
- MHD ✅ (Y. Lapeyre)
- Black hole + warp ✅

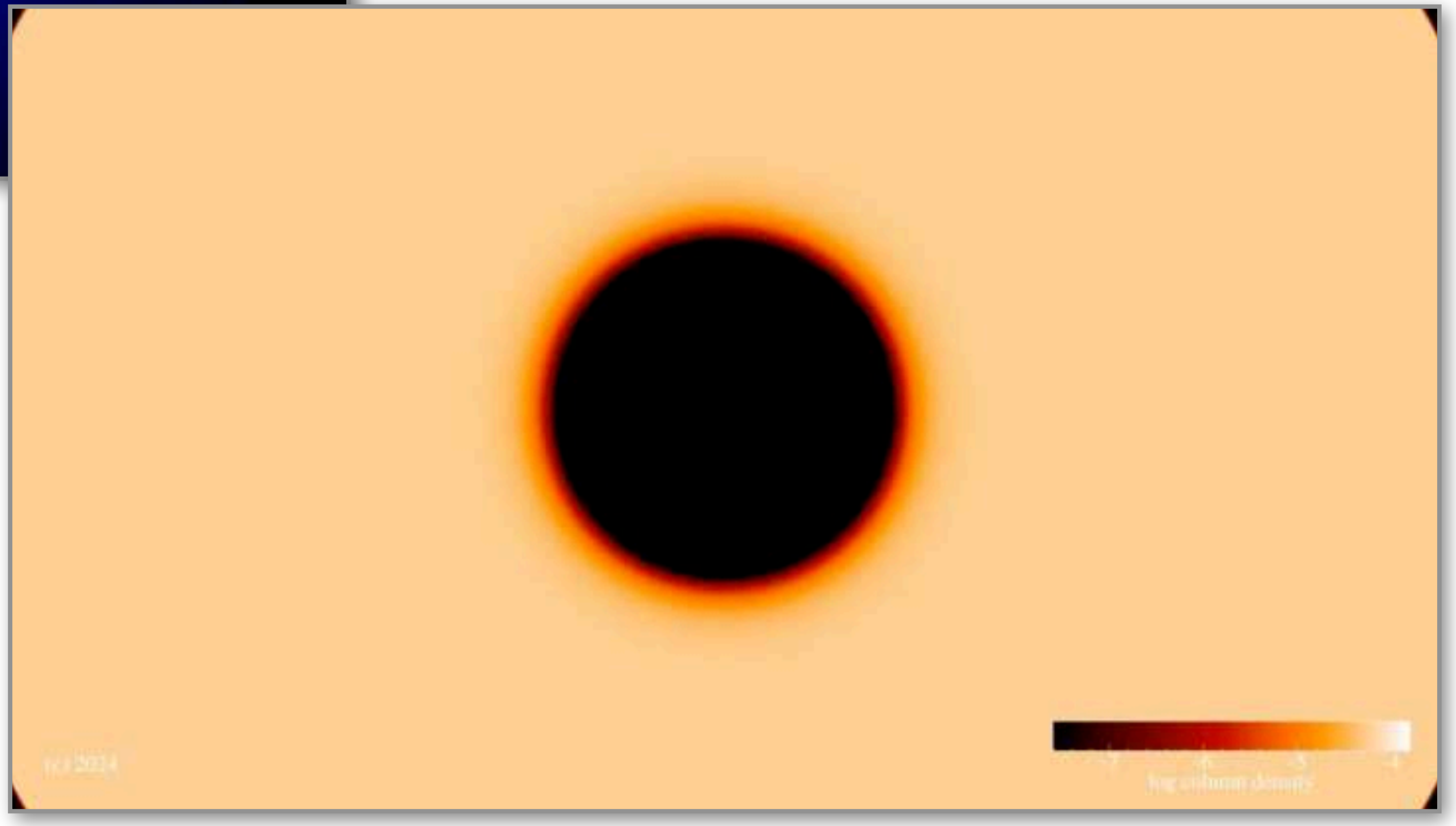
## AMR solver (WIP) :

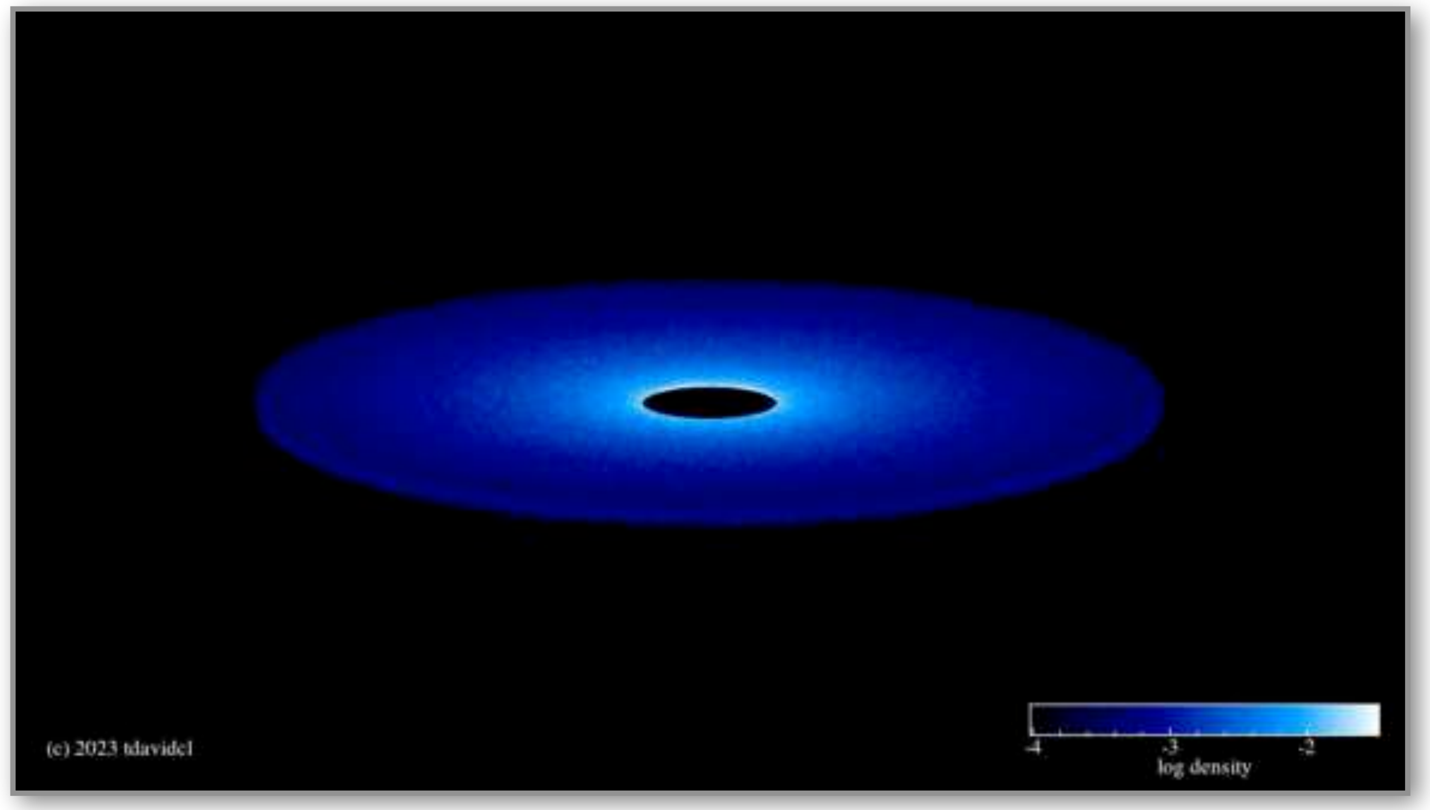
- Hydro ✅
- Dust ✅
- Self-gravity (WIP)

## Next :

- Exascale disc simulations
- Non local forces (self-gravity)
- Live MCMC radiative transfert







# Shamrock