

Accelerating radiative transfer with GPUs

Oliver Zier

Center for Astrophysics | Harvard & Smithsonian

With Aaron Smith, Xuejian (Jacob) Shen, Rongrong Liu, Mark Vogelsberger, Volker Springel, Ruediger Pakmor, Rahul Kannan, …

Simulations GEARS, Kavli IPMU · 5 November 2026

Why is radiation transport expensive?

$$\frac{1}{\tilde{c}}\frac{\partial I_\nu}{\partial t} + \boldsymbol{n}\cdot\boldsymbol{\nabla} I_\nu = j_\nu - \alpha_\nu I_\nu$$

Many dimensions

$$I_\nu(\boldsymbol{x}, \boldsymbol{n}, t) = \frac{\mathrm{d}E_{\mathrm{rad}}}{\mathrm{d}t\,\mathrm{d}A_\perp\,\mathrm{d}\nu\,\mathrm{d}\Omega}$$

3 position + 2 direction + 1 frequency

Fast propagation

$$\Delta t_{\mathrm{RT}} \lesssim C_{\mathrm{RT}}\frac{\Delta x}{\tilde{c}}$$

Radiation crosses cells much faster than the gas

M1: evolve moments of the radiation field

$$E = \frac{1}{\tilde{c}} \int\int I_\nu \, \mathrm{d}\Omega \, \mathrm{d}\nu \qquad \boldsymbol{F} = \int\int \boldsymbol{n} I_\nu \, \mathrm{d}\Omega \, \mathrm{d}\nu \qquad \mathbb{P} = \frac{1}{\tilde{c}} \int\int (\boldsymbol{n} \otimes \boldsymbol{n}) I_\nu \, \mathrm{d}\Omega \, \mathrm{d}\nu$$

$$\frac{\partial E}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{F} = S_E$$

$$\frac{\partial \boldsymbol{F}}{\partial t} + \tilde{c}^2 \boldsymbol{\nabla} \cdot \mathbb{P} = \boldsymbol{S}_F$$

$\tilde{c}$: reduced speed of light

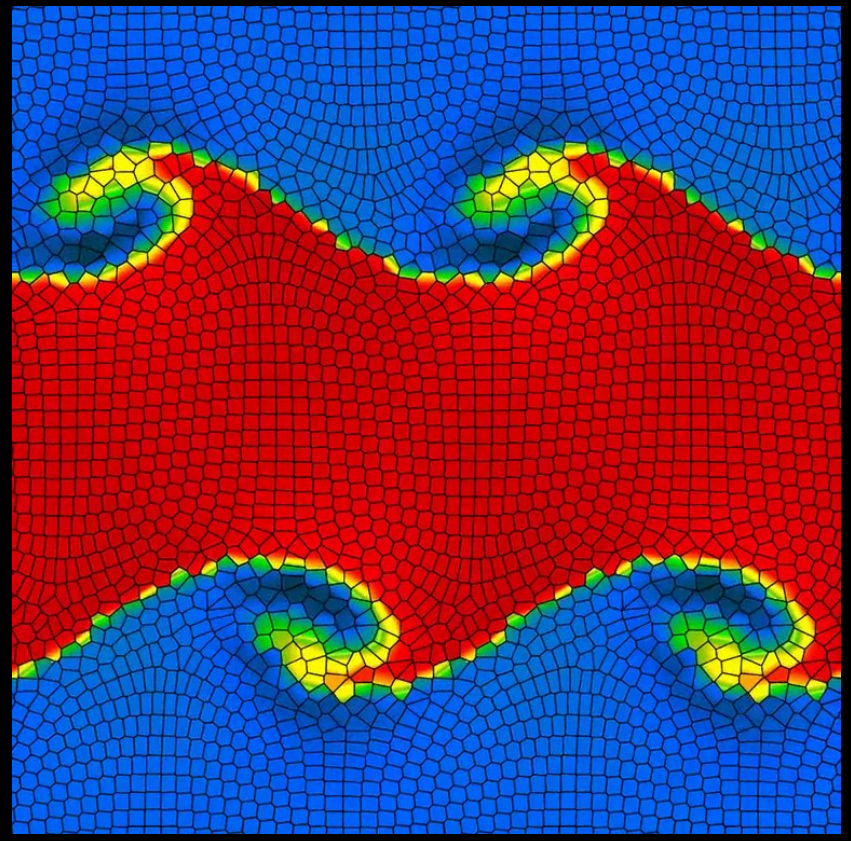
4 variables per frequency bin

M1 closure

$$\mathbb{P} = E \, \mathbb{D}_{\mathrm{M1}}(E, \boldsymbol{F})$$

Kannan et al. (2019)

Radiation transport in AREPO



Kelvin–Helmholtz instability

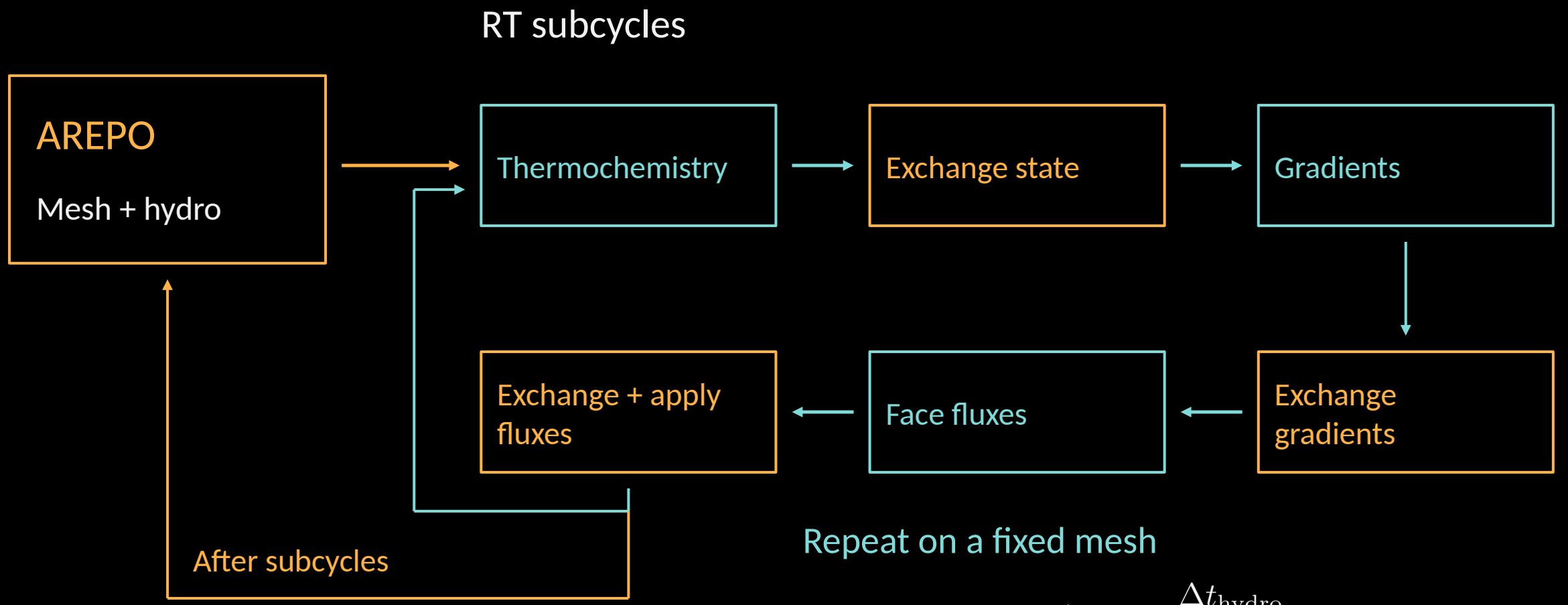
Springel (2010); Kannan et al. (2019)

Moving Voronoi mesh

Finite-volume hydrodynamics

RT on the same mesh

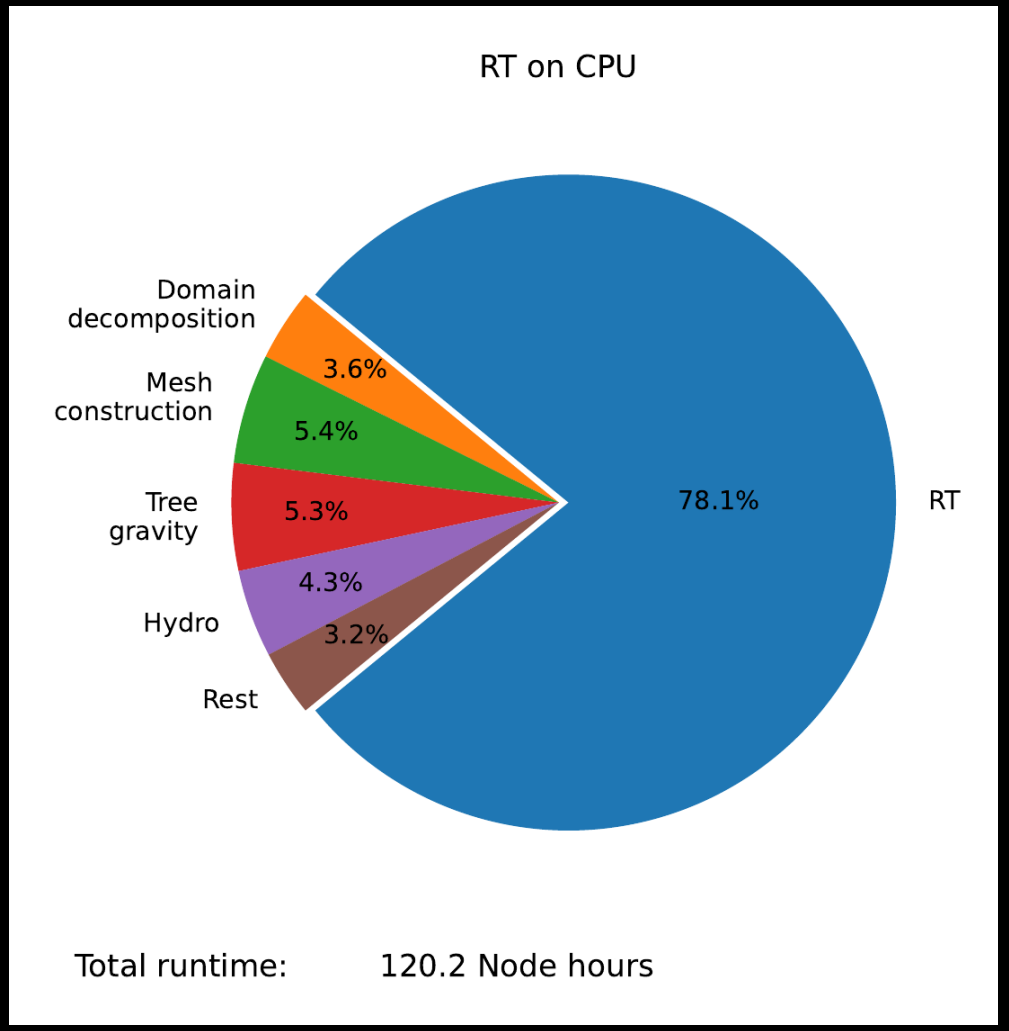
Structure of the RT solver



$$\Delta t_{\mathrm{RT}} = \frac{\Delta t_{\mathrm{hydro}}}{N_{\mathrm{sub}}}$$

Zier et al. (2024)

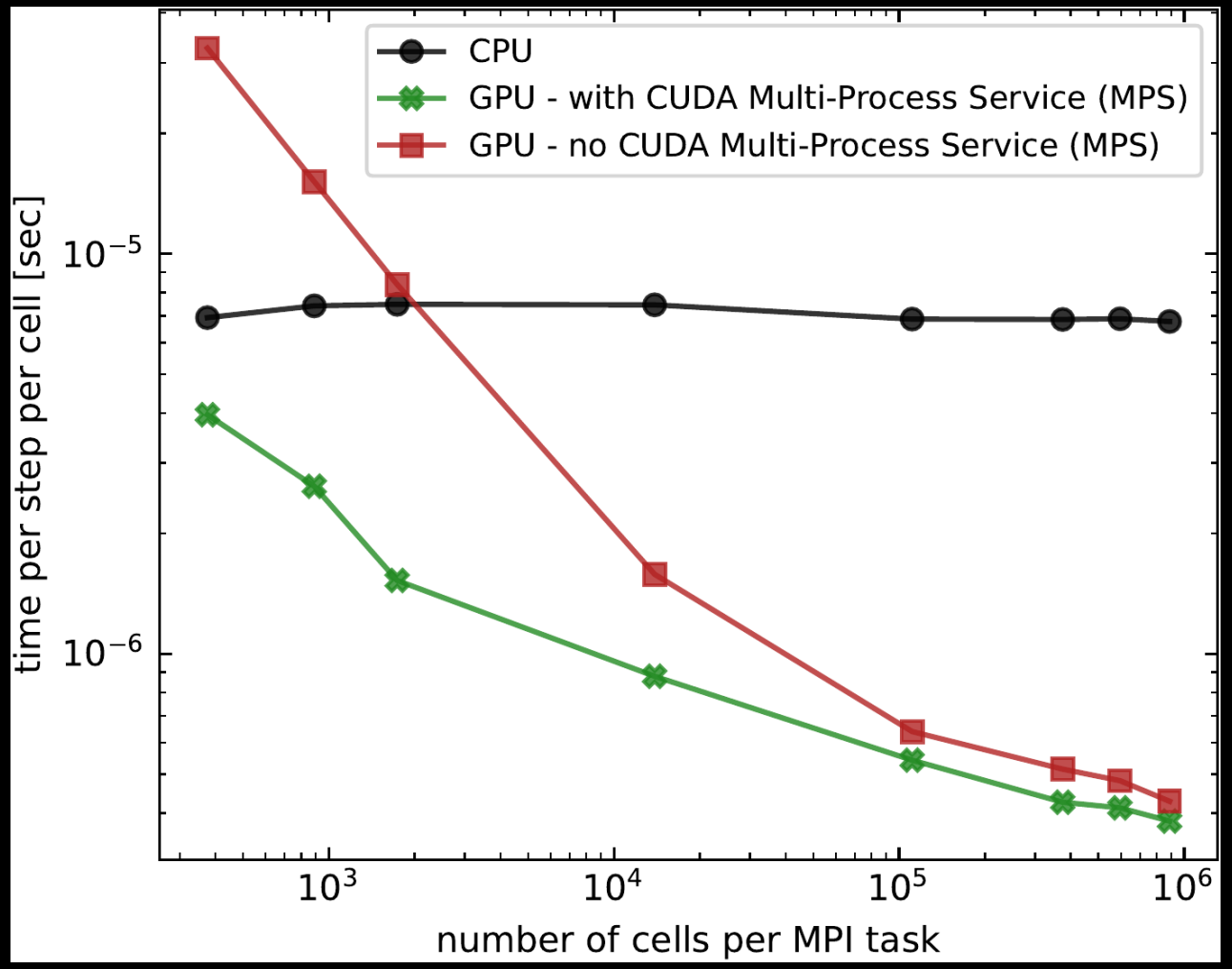
TNG + AREPO-RT



64 subcycles

Zier et al. (2024)

Radiation-wave performance



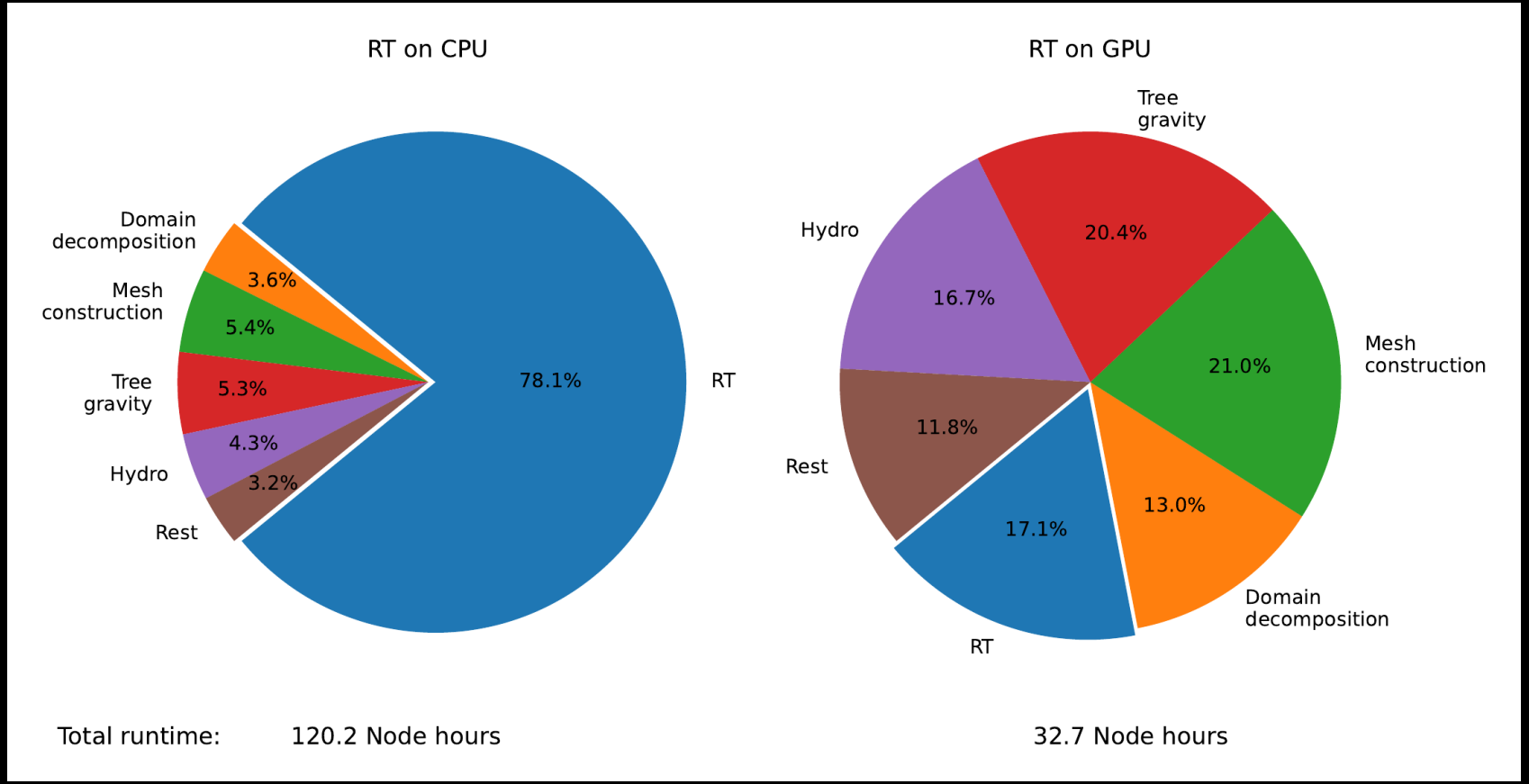
More work per GPU

Lower cost per cell

Zier et al. (2024)

GPU acceleration in a cosmological box

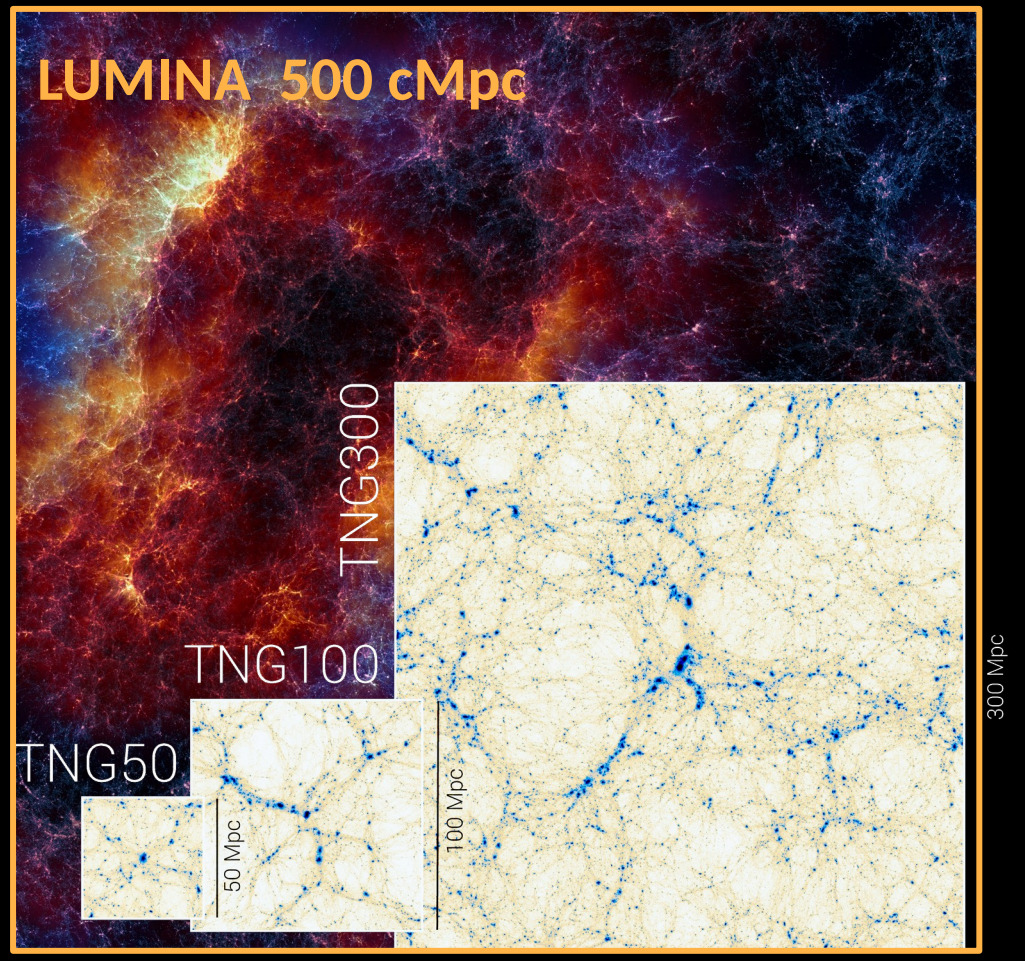
72 CPU cores vs 72 CPU cores + 4 A100 GPUs



17× faster RT, 3.7× faster full simulation

Zier et al. (2024)

The LUMINA simulation



500 cMpc box

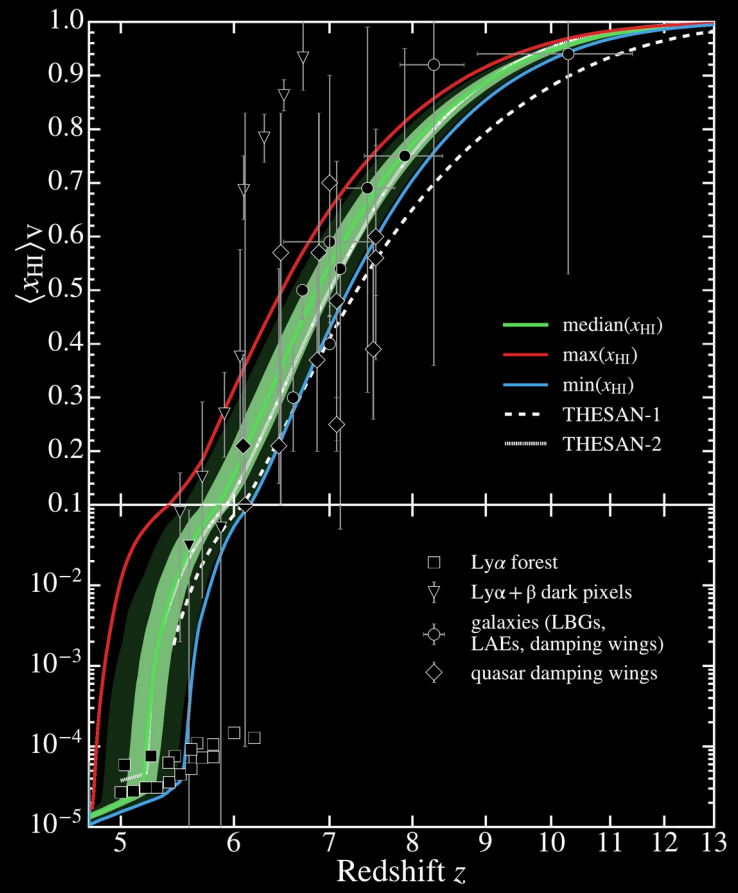
6000^3 gas cells
+ 6000^3 dark-matter particles

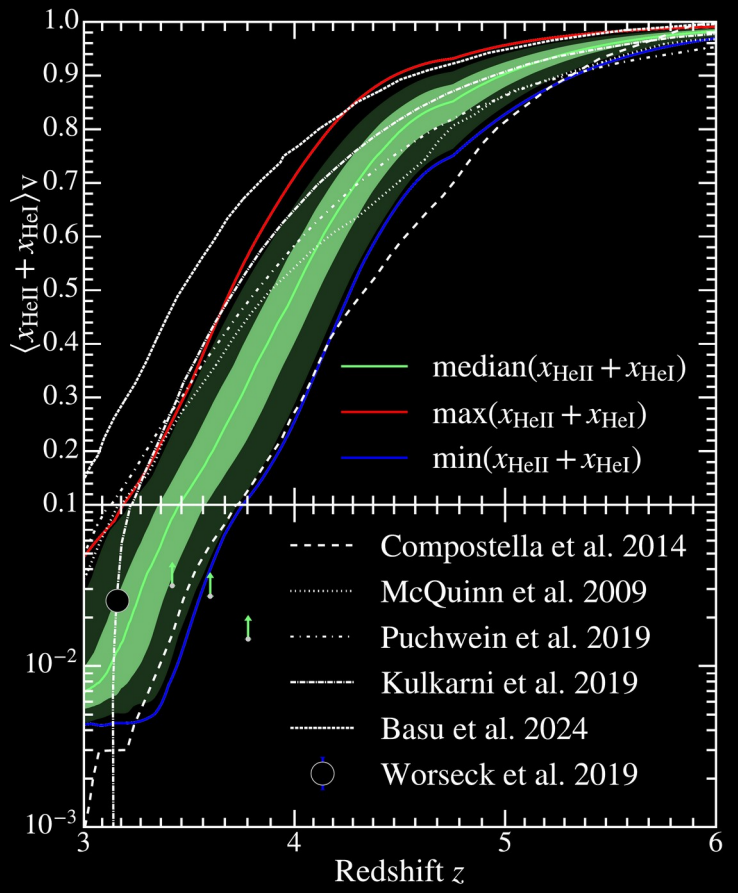
Hydrogen + helium reionisation
to z = 3

IllustrisTNG + on-the-fly RT

Zier et al. (2026)

LUMINA: hydrogen and helium reionisation





Hydrogen: stars

Helium: AGN

Zier et al. (2026)



Application: LUMINA

105,504
CPU cores

7,536
AMD MI250X GPUs

109 M
CPU hours

600 TB
peak memory

250 TB
per checkpoint, written at 800 GB/s

140 PB
restart files written in total

Frontier, OLCF/ORNL (CC BY 2.0)

Why RT still costs time

RT work per time step (64 subcycles)

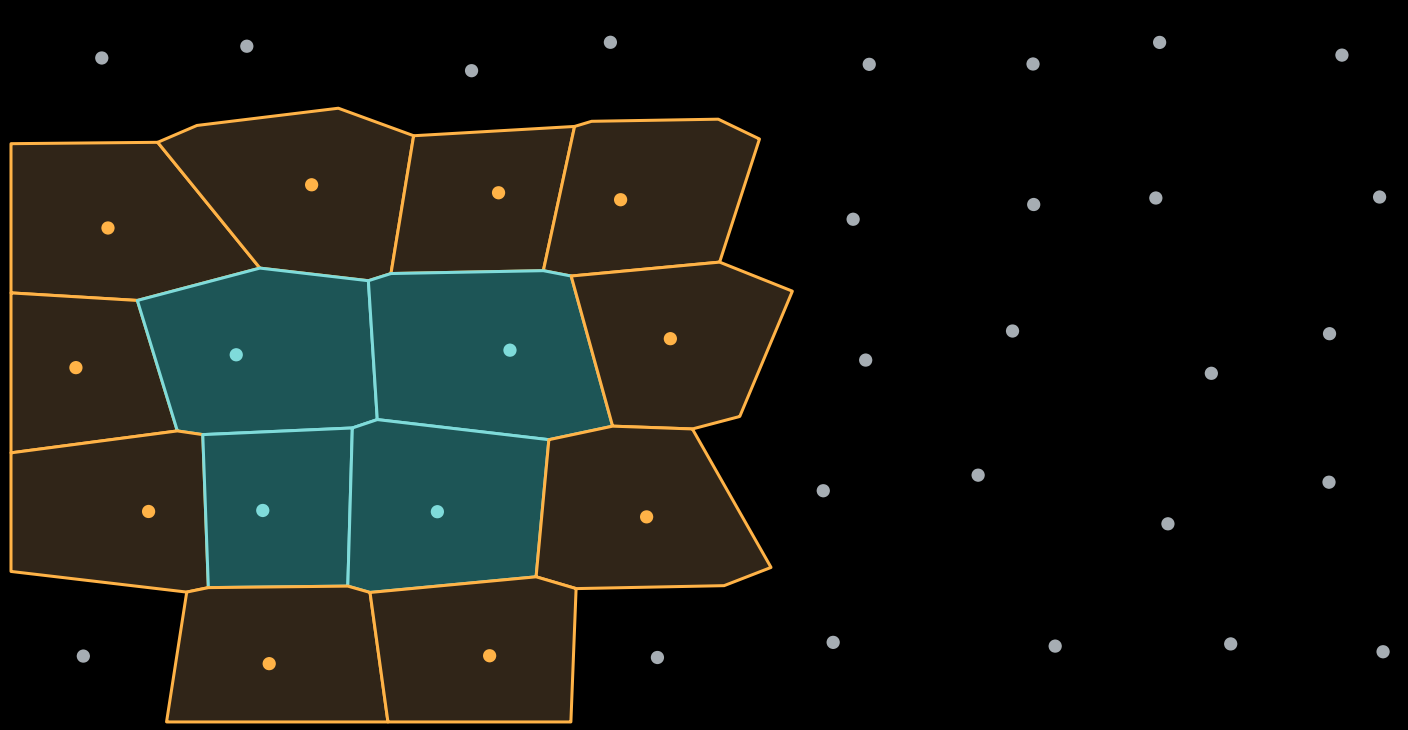
- 64× thermochemistry
- 64× fluxes
- 64× gradients

Cell timestep

$$\Delta t_{\mathrm{cell}} = \min(\Delta t_{\mathrm{hydro}},\ \Delta t_{\mathrm{grav}},\ N_{\mathrm{sub}} \Delta t_{\mathrm{RT}})$$

Gas on the effective EOS is transparent to radiation

Individual timesteps and RT subcycling



Active cells

Passive neighbours

Points only:
mesh not constructed

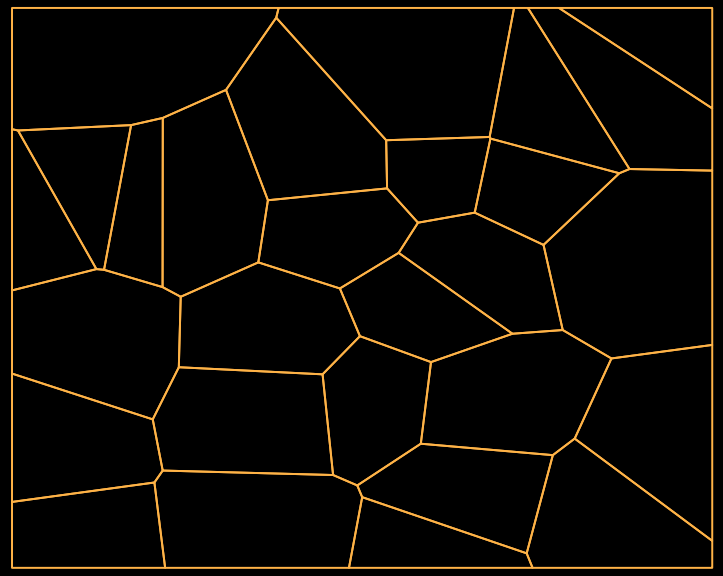
Fluxes only across faces touching an active cell

Current AREPO-RT: radiation cannot pass the passive layer

RT on a separate AMR grid

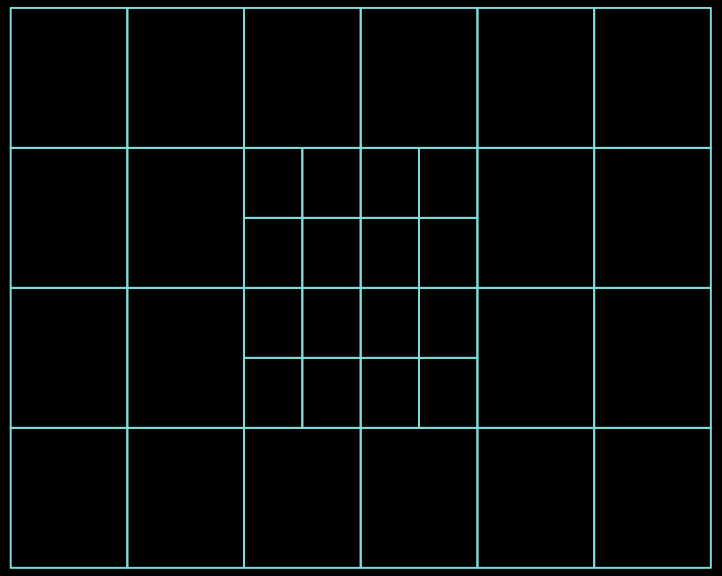
CPU: AREPO

GPU: radiation transport



Bin density, temperature, and ions onto AMR

Map ionisation rate back to Voronoi



Moving Voronoi mesh

Adaptive radiation grid

Zier, Pakmor, et al., in prep.

Future work

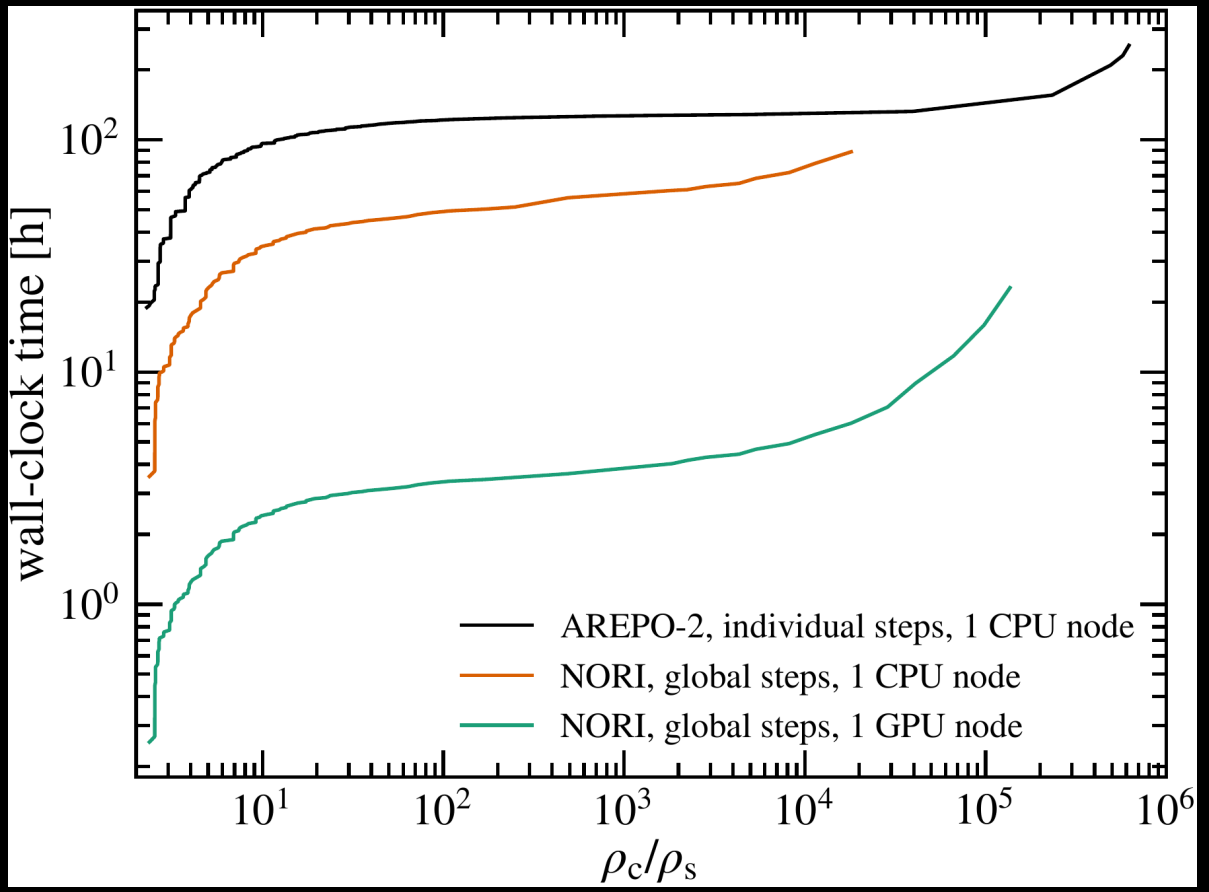
IGM refinement

- Constant spatial resolution in the IGM
- Increase resolution in the CGM
- Decrease resolution in the ISM

Beyond M1

- Higher-order moments: M2?
- Ray tracing? Discrete ordinates?

NORI: self-interacting dark matter



≈25× faster

NORI: 5.2 hours

4 A100 GPUs

AREPO-2: 5.4 days

112 CPU cores

$N_{200} = 10^7$; compared at ρc / ρs = 10^4

Zier et al. (2026)

Summary

RT is expensive
GPUs can help

LUMINA is an application
of GPU acceleration

New RT solver:
AMR grid + Voronoi mesh

LUMINA at z = 3



Video: Rongrong Liu