3rd International Workshop on AI + Astronomy | Zhejiang Lab, Hangzhou | 19 Oct 2026

How Much Compute Does Good Science Need?

Co-Designing AI and Numerical Methods for Astronomy

Dr. Denisa Constantinescu
University of Basel, DMI HPC Group | EPFL ESL
Dr. Rubén Cabezón
University of Basel, sciCORE
Rubén Rodríguez Álvarez
EPFL, Embedded Systems Laboratory
Prof. Florina Ciorba
University of Basel, DMI HPC Group

denisaandreea.constantinescu@unibas.ch | denisa-c.com

Who we are: from chips to clusters to exascale astro codes









DMI HPC Group, University of Basel

Florina Ciorba | Denisa Constantinescu

- **SPH-EXA**: exascale fluid + gravity simulation code
- **Scheduling**, load balancing & energy of simulations, LCA, **scale**

sciCORE, University of Basel

Rubén Cabezón

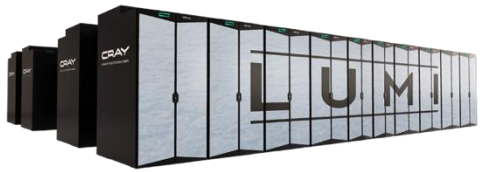
- The **astrophysics** inside **SPH-EXA**: supernovae, turbulence
- UniBasel research computing centre

Embedded Systems Lab, EPFL

Rubén Rodríguez Álvarez | D. Constantinescu

- **Reconfigurable chips** & custom **numeric formats** for SKA
- **Benchmarks** & HW **co-design**
- Control**, LCA (CO_2), TCO ($)**





HPCaTS

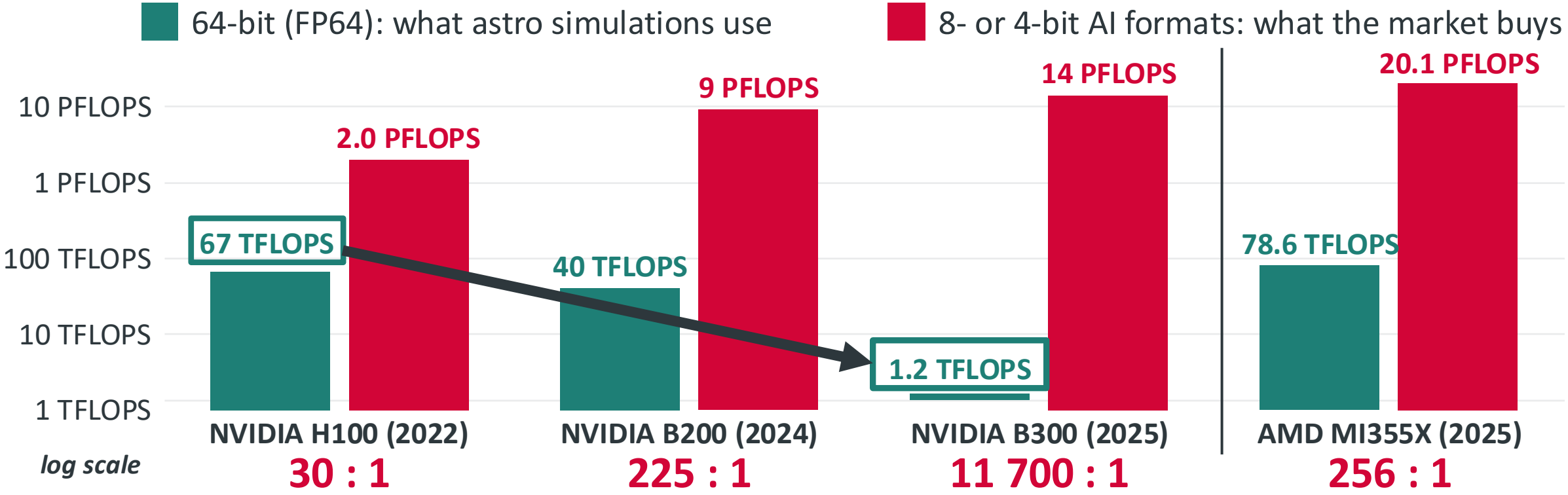
High Performance & Privacy through Confidential Computing, Trust and Security

The machines (instrumented allocations)

Alps (CSCS, GH200) | **LUMI-G** (EuroHPC, AMD MI250X)
Kuma (EPFL, NVIDIA H100) | **sciCORE** (UniBasel)
HPCaTS (DMI-UniBasel: GH200, MI300A)

GPUs are leaving 64-bit precision

Peak throughput per GPU, vendor datasheets: the 64-bit units we use versus the low-precision units AI buys



Systems "for science" follow: NVIDIA Vera Rubin, 7 EFLOPS of AI vs 5 PFLOPS of FP64 per system (1 400 : 1)

Green side: exact FP64 matrix multiply emulated on 8-bit integer units runs 2.3× faster than native FP64 on GB200

The AI industry will not solve our problems. Astronomy has to state what precision it needs

NVIDIA H100 / B200 / B300 and AMD MI355X datasheets (via G. Lockwood) · NVIDIA Vera Rubin press release, 2026 · Schwarz et al. (NVIDIA), HPCAsia 2026, cuBLAS 13.0 FP64 emulation.

FLOP/s says little about scientific return on investment (ROI, $)

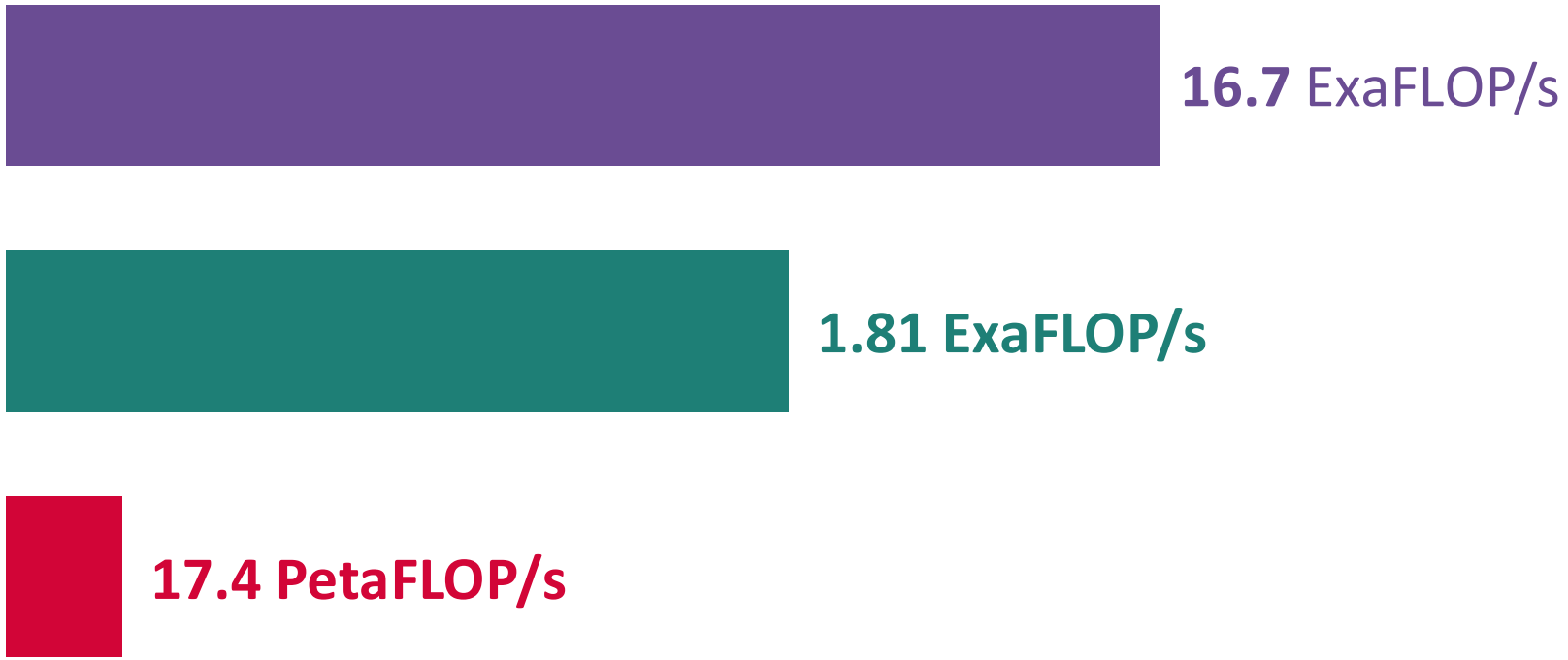
El Capitan, TOP500 #2 (June 2026): one machine, three benchmarks

HPL-MxP
mixed-precision
High-**P**erformance **L**inpack
AI bench – near 100% utilization

HPL - High-**P**erformance **L**inpack
dense FP64 linear algebra
TOP500 bench

HPCG - HP Conjugate **G**radient
sparse, memory-bound solver
what our codes look like

bar length is logarithmic



960 : 104 : 1

For memory-bound astronomy codes the useful number is 1 % of the TOP500 headline

TOP500, HPCG and HPL-MxP lists, June 2026 (top500.org). The new #1, LineShine (42 MW, no GPUs): HPCG is 1.0 % of its HPL as well.

Scientific GOODPUT: where allocation of an SKA imaging node goes

astroCAMP, PASC 2026: 100 instrumented runs of WSClean + IDG on AMD EPYC + NVIDIA H100, OSKAR simulated SKA-Low data

Compute utilization

hardware counters (uProf, Nsight), roofline

4 to 14 % imaging arithmetic

86 to 96 %
stalled: single-threaded I/O, data movement, orchestration

→ **2.8×**
max CPU speed-up stalls at 16 threads

Energy per run

wall-plug PDU, RAPL, NVML, PMT

15 to 20 % dynamic: computing

80 to 85 %
static: node powered on, waiting

→ **90-95 %**
of carbon is operational
2.1× more in SA than in WA

Science products

*to measure: RMS, SSIM dynamic range vs **reference***

?
valid at the required resolution & dynamic range

?
not valid, or never checked

= 100 %

the split is unknown: quality is not yet scored against the SKA requirements

astroCAMP

Performance	Energy	Carbon	System	Economics	Quality*

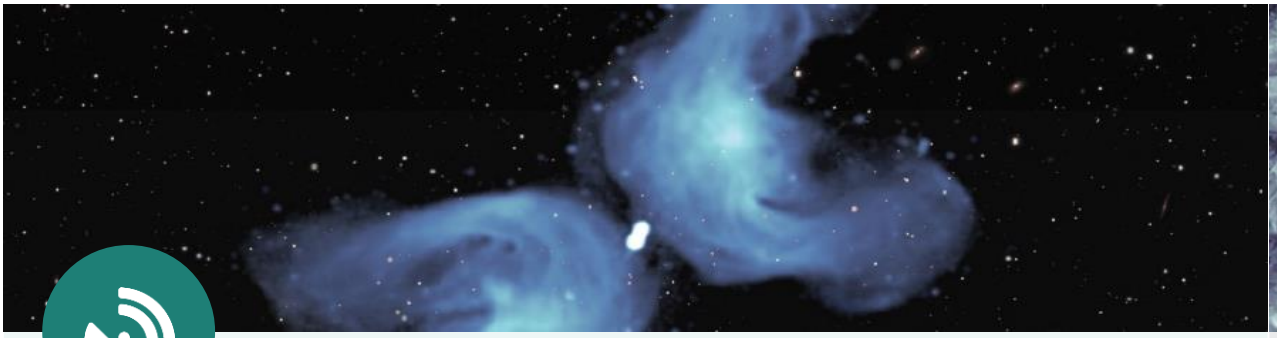
***Quality** (residual noise, image similarity, dynamic range, astrometry, photometry) is defined but not quantified

GOODPUT = valid science products ÷ node-hours, joules, $

Constantinescu, Rodríguez Álvarez et al., astroCAMP, PASC 2026 (arXiv:2512.13591): Sections 6.2, 6.3, 7.1 · data doi:10.5281/zenodo.20093790 · SKA Level-0 requirements SKA-TEL-SKO-0000007.

Two kinds of astro science, two kinds of "good enough"

"Start from the science" means: write down what a result must keep



Observations and surveys

Ground truth partly known → tolerances & specs

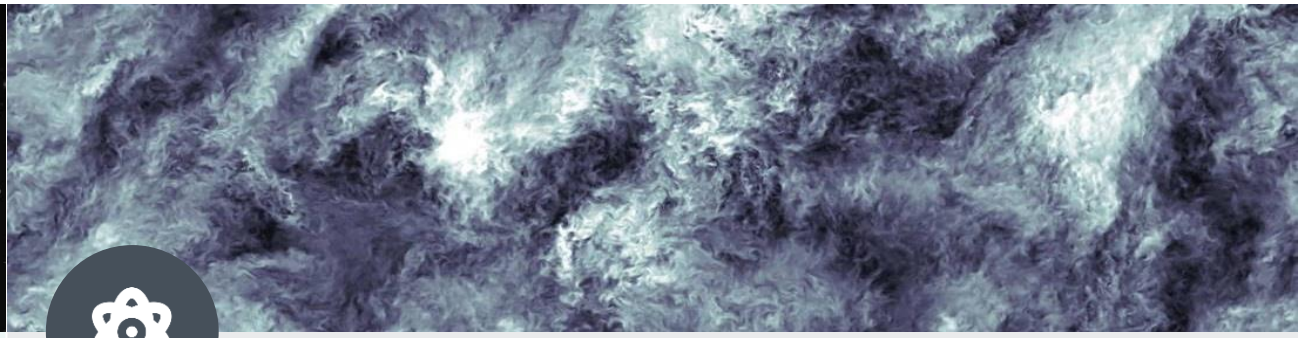


SKA-Mid: 60 dB image dynamic range, 1 % flux scale



Euclid: weak-lensing shear bias $|\Delta m| < 2 \cdot 10^{-3}$

Rubin | Gaia:
- 5 mmag photometry
- 20-30 µas parallax



Frontier simulations

Ground truth unknown → invariants instead

Conservation: mass, momentum, energy, entropy

Convergence: with resolution: a tidal-disruption stream only at 10^{10} particles

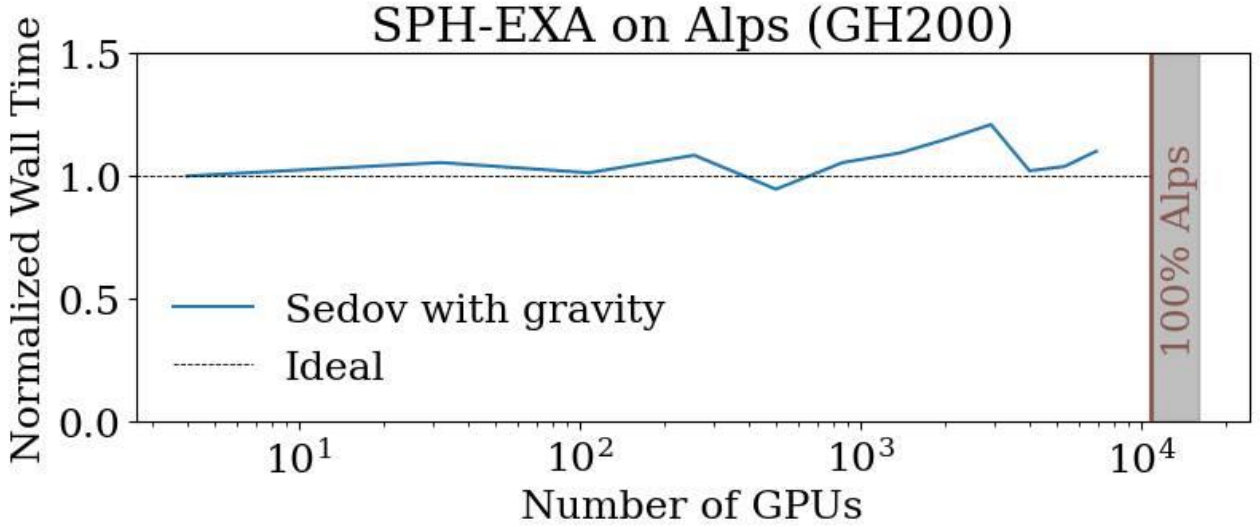
Cross-code agreement: 4 supernova codes within ~10 %; power spectra

Fidelity map = error tolerances on the left + invariants on the right

SKA-TEL-SKO-0000007 Rev 02 · Euclid Collaboration 2024 (arXiv:2405.13491) · LSST LPM-17 · Gaia EDR3 (Lindegren+ 2021) · Cabezón+ 2018, A&A 619 · Kubli+ 2025 (SPH-EXA TDE).

Case 1 | SPH-EXA: performance first, and the science kept

UniBasel + CSCS exascale smoothed-particle hydrodynamics (SPH) & gravity code



***Weak scaling** on **Alps** (NVIDIA GH200), **blast wave with self-gravity**, 432 M particles /GPU*
S. Keller, CSCS, 2024



2.98 trillion

particles: the largest SPH simulation we know of

90 %

weak-scaling efficiency on 6 912 GPUs (64 % of Alps)

13.9× fewer node-hours than planned for the **10 000³ turbulence run** on **LUMI-G**: 5.4 M planned, under 400 k used. **Individual time-stepping** did it, not new hardware.

Performance portability - C++20 source:
- Piz Daint (NVIDIA P100)
- LUMI-G (AMD MI250X)
- Alps (NVIDIA GH200) → accepted with this code

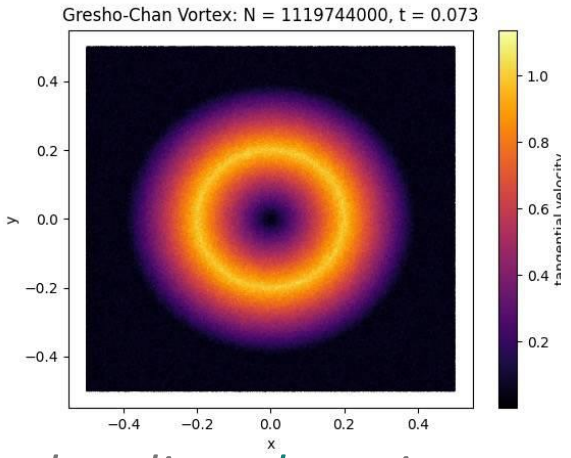
Biggest saving ever was an algorithm
Goodput is heavily decided in the code before the hardware is chosen

Platform for Advanced Scientific Computing

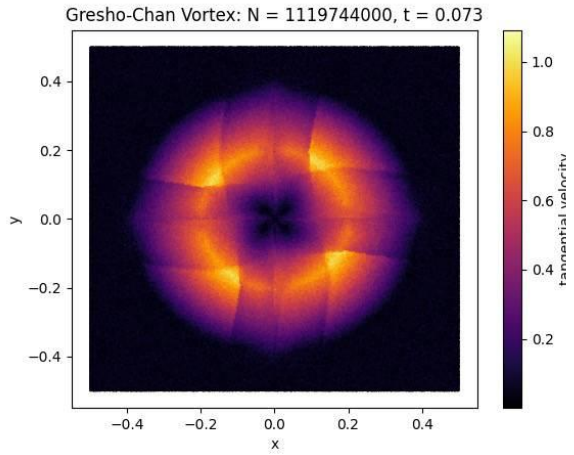
Keller et al., PASC 2023 (Cornerstone) · Ciorba, CARLA 2026 keynote · EuroHPC EHPC-EXT-2023E01-031 (22 M GPU-h on LUMI-G) · github.com/unibas-dmi-hpc/SPH-EXA (MIT).

Case 1 | 23 % faster, but physically wrong

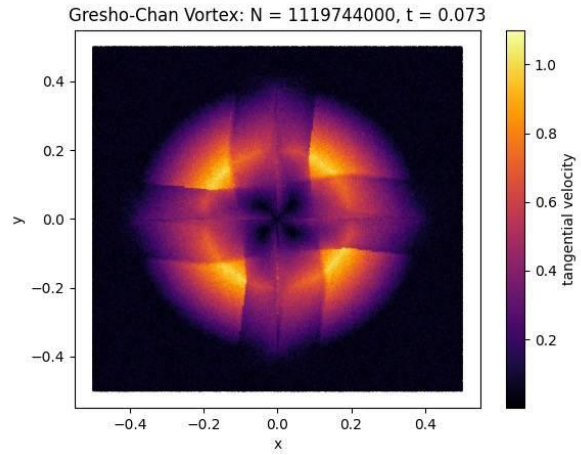
Re-partitioning the domain every 20 steps instead of every step (Gresho-Chan vortex test, 1.12 billion particles)



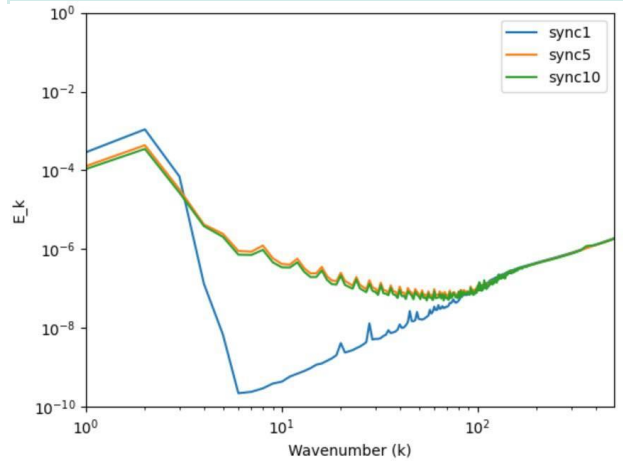
baseline: clean ring



every 5 steps: spokes



every 10 steps: wrong



turbulence energy spectrum: every 10 steps departs

–23 %

time-to-solution
11 445 s → 8 778 s

stale neighbours

copies of neighbouring particles go out of date: spokes, more steps, wrong spectrum

1 % trigger

rebalance only when a rank drifts > 1 % and update the tree: baseline physics, fires only when needed

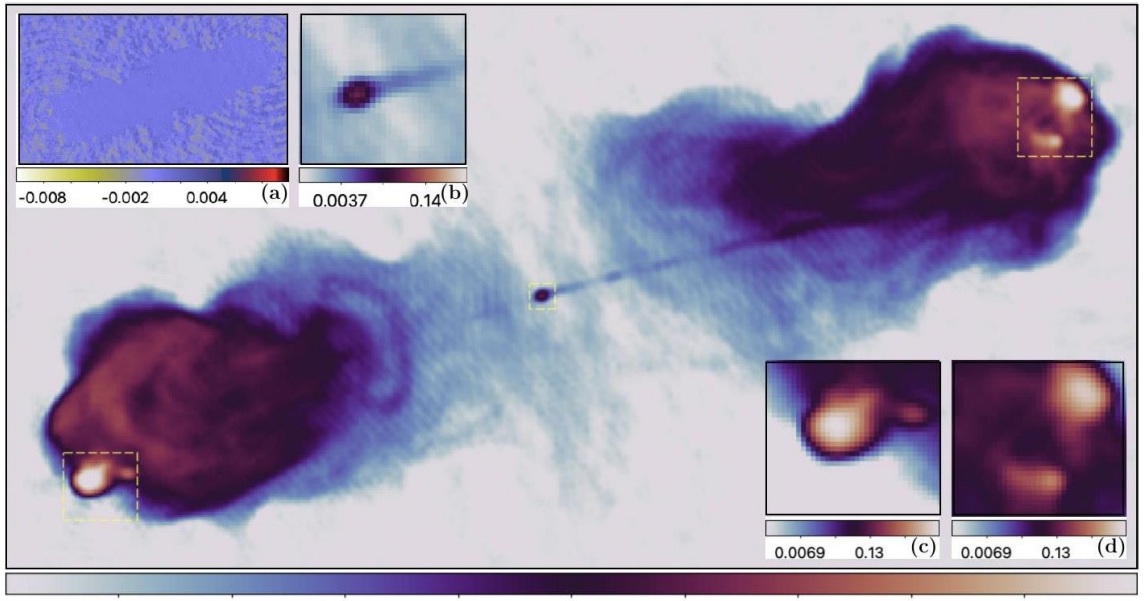
Every performance decision needs an automatic science check

Platform for Advanced Scientific Computing

O. S. Simsek and F. M. Ciorba, Cost-Benefit Analysis of Domain Re/Decomposition in SPH-EXA, Leogang HPC Workshop, Feb 2026 · Ciorba, CARLA 2026 keynote · Daint.Alps GH200.

Case 2 | Two ways to compute a SKA image

Same object, same VLA data, 512^2 pixels: the numerical pipeline we benchmarked vs a learned alternative



Cygnus A reconstructed with WSClean
Yellow boxes: the hotspots compared on the right

SKA asks for 60 dB dynamic range (10^6 : 1) at 0.5" and 1 % flux

Shown so far: 52 dB ($1.7 \cdot 10^5$) on a 512^2 image, both methods.
Neither is scored against the SKA numbers at SKA resolution

East hotspot | West hotspot | residual

CLEAN

WSClean + IDG

numerical



33.7 s on 1 CPU core | residual noise $8.6 \cdot 10^{-4}$ | error controlled by the numerics: 32-bit gridding, tunable accuracy target

R2D2

learned CLEAN

AI



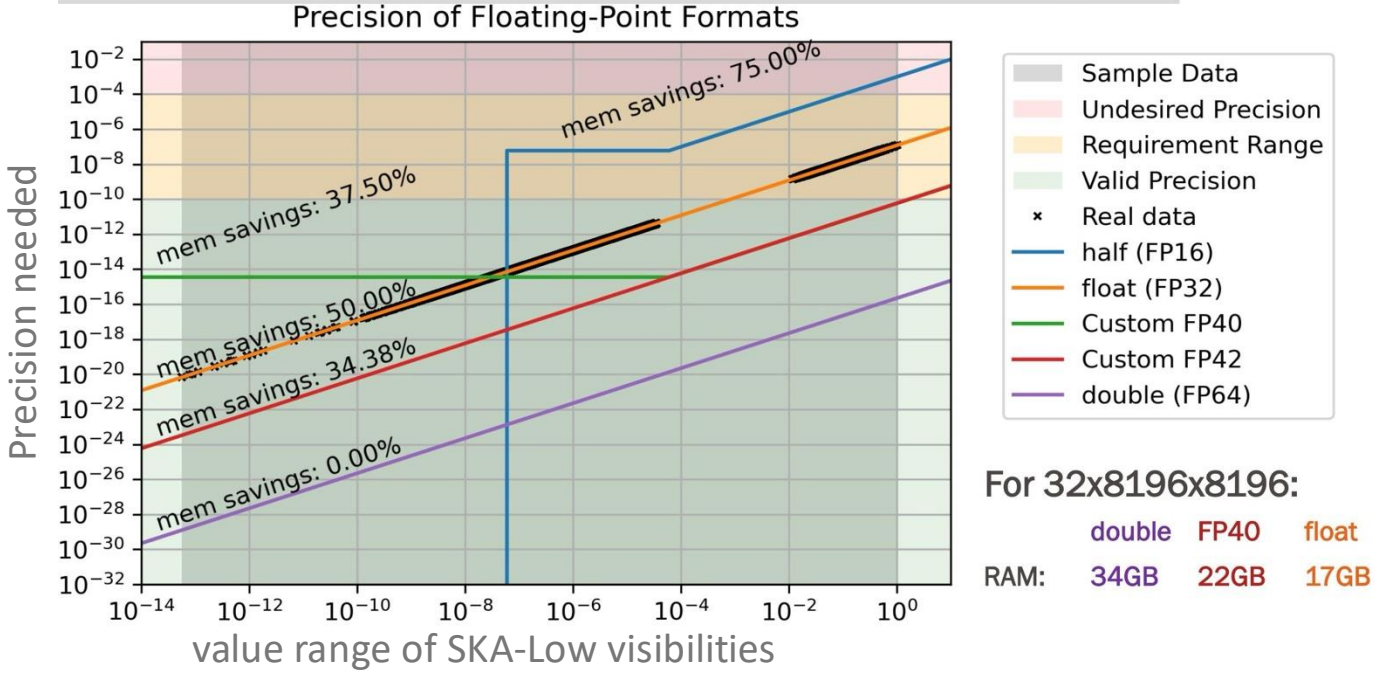
3.3 s on 1 GPU | residual noise $11.7 \cdot 10^{-4}$ | error comes from 20 000 simulated training pairs, no numerical bound
"one cannot ascertain whether this sharpness is physical"

Which one is right? Learned or numerical?
neither has been scored against the SKA tolerance at the required resolution

Images and numbers: Dabbech, Aghabiglou et al., CLEANing Cygnus A deep and fast with R2D2, ApJL 2024 (Figs. 1, 2, Table 2) · Veenboer & Romein 2020 (IDG) · astroCAMP, PASC 2026 · SKA-TEL-SKO-0000007.

Computer science for science: hardware arch. meets tolerance

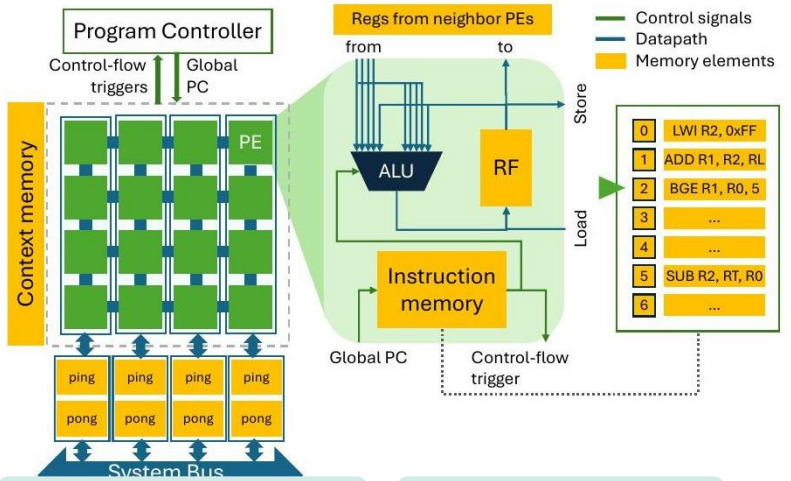
SEAMS: build a **busy** & **frugal CGRA chip** with numeric formats the astro science needs



For 32x8196x8196:

	double	FP40	float
RAM:	34GB	22GB	17GB

TAC-Co **CGRA**
4×4 reconfigurable array, TSMC 16 nm (ASP-DAC 2027)



1.79 GHz
highest clock of any FP32 reconfigurable array

52 % busy
cells utilised on average: FFT 61 %, conv 52 %, matmul 44 %

FACETs
C compiler: automatic flow → practical barrier of CGRAs removed

format must stay in the green band. R. Rodríguez Álvarez, SKA Days 2024

34 → 22 GB
with the custom 40-bit format

16-bit fails
value range too small for visibilities; 32-bit only above $3 \cdot 10^{-5}$

Goodput / silicon area
busy cells, not dark silicon: 3.56 GFLOP/s of FFT

Energy / result →$, CO_2
pipelining: 2.2× clock at ~no leakage;
fast cells: 20× leakage (0.67 → 13.7 mW).
Chosen per kernel

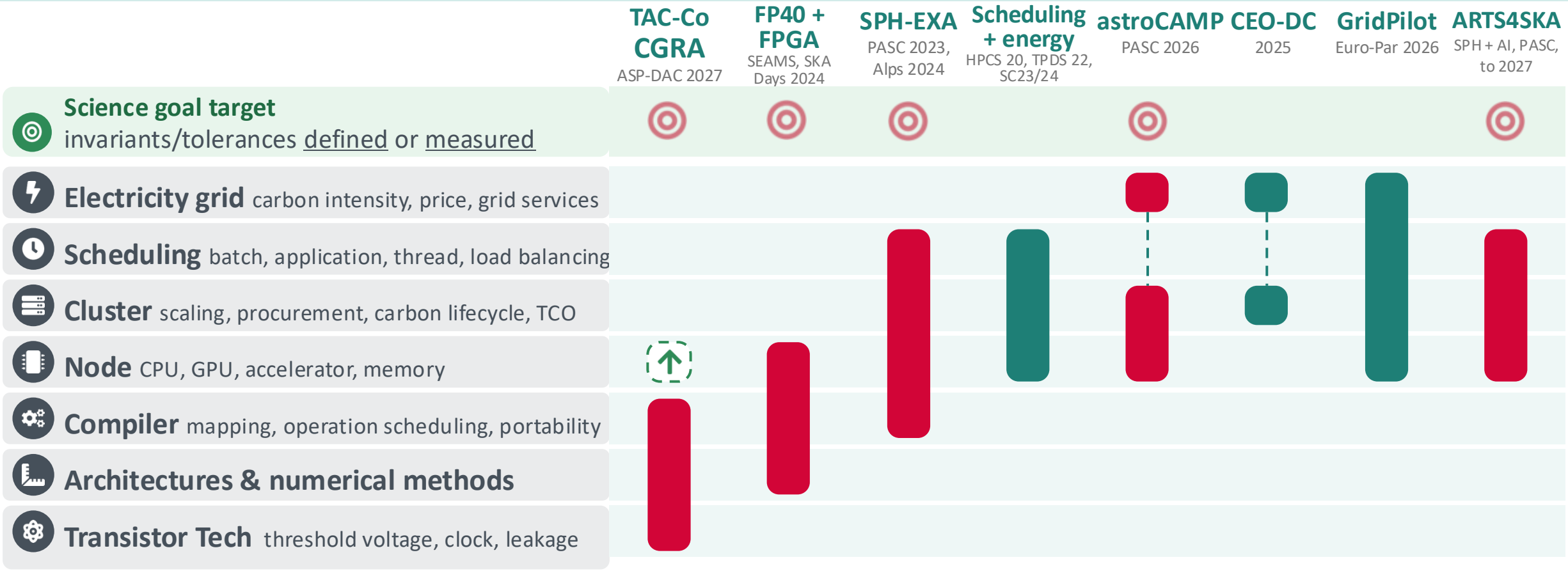
Next: - scale to SKA imaging needs (gridding, FFT @ SDP scale)
- custom-precision array for LLMs and FFTs

With the tolerance as input
the numeric format, the utilisation per chip area & energy per result become HW co-design variables

Rodríguez Álvarez et al., SKA Days 2024 · Arras et al. 2021 (wgridder) · Corda et al. 2021 · Rodríguez Álvarez, Wang, Constantinescu, Ansaloni, Peón-Quirós, Atienza, TAC-Co, ASP-DAC 2027 (Tables 4, 5; R-Blocks comparison).

Co-design across the whole stack, for GOODPUT and SUSTAINABILITY

From the transistor to the electricity grid: each bar is one of our works, spanning the layers it co-designs



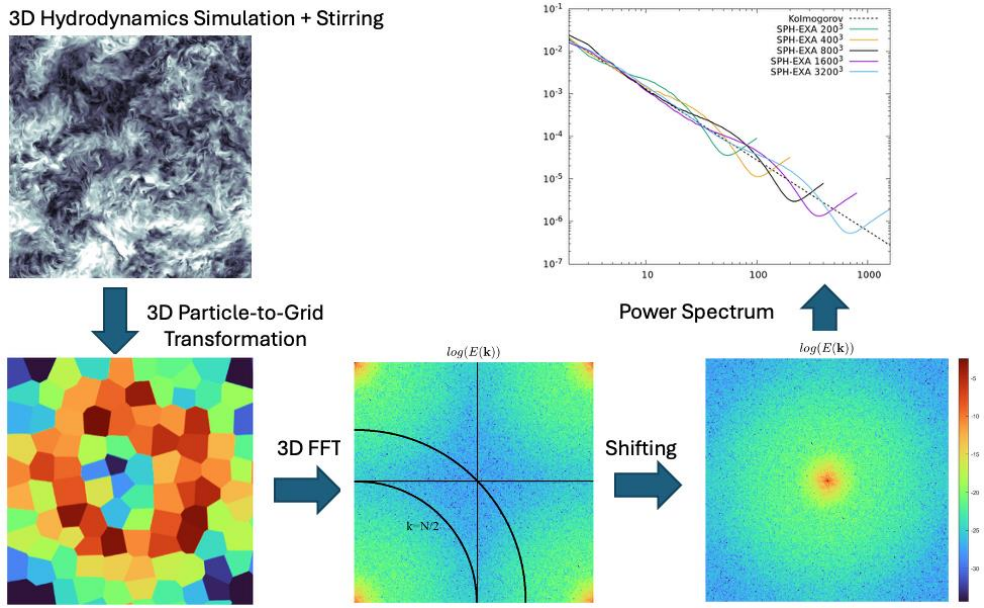
■ published co-design ▢ next step: scale the TAC-Co CGRA to node level ◎ hits the target: science quality defined or measured

Who computes what, where and when is a runtime scheduling & co-design decision

TAC-Co: Rodríguez Álvarez et al., ASP-DAC 2027 · SEAMS FPGA / FP40: SKA Days 2024 · SPH-EXA: Keller et al., PASC 2023; Alps acceptance 2024 · Scheduling + energy: Eleliemy, Mohammed, Müller Korndörfer, Simsek, Ciorba · astroCAMP: PASC 2026 · CEO-DC 2025 · GridPilot: Euro-Par 2026 · ARTS4SKA: PASC (ETH, ZHAW, Basel, EPFL).

ARTS4SKA: SPH + AI in one simulation

PASC project of ETHZ, ZHAW, UniBasel and EPFL: solver & AI judged by the same invariants



P2G: particles → grids → observables (power spectrum, 21-cm, radiative transfer). O. Simsek, SKA Days 2026

Accelerated Radiative Transfer Simulations for SKA: SPH-EXA physics, AI operators, 21-cm signal for SKA-Low

SPH-EXA	- **fine-grained** - conservative to machine precision - **double precision** throughout
AI surrogate	- **coarse-grained** ranges - **100× cheaper**, no guarantee
Particle-to-grid bridge (P2G)	both become observables: power spectra, 21-cm maps, radiative transfer
Invariant check	- Conservation energy spectrum - Convergence - cross-code agreement

The question we will answer: how much AI error keeps every observable inside its tolerance, and at what resolution does the solver have to take over?

A simulation 100× cheaper that passes its invariants is GOODPUT
One that does not is NOISE and wasted compute

Platform for Advanced Scientific Computing

ARTS4SKA (PASC; ETH Zurich, ZHAW, Basel, EPFL), cosmology.ethz.ch · SPH-EXA2 (PASC), github.com/unibas-dmi-hpc/SPH-EXA · Simsek et al., GPU particle-to-grid library, Swiss SKA Days 2026 · Cabezón et al. 2025, arXiv:2503.10273.

What computer scientists need to maximize GOODPUT?

astroCAMP

A fidelity map for astronomy science cases and benchmarks



Radio imaging

60 dB dynamic range
1 % flux scale

SKA Level-0 requirements



Weak-lensing shear

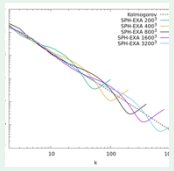
bias $|\Delta m| < 2 \cdot 10^{-3}$

Euclid

Photometry, astrometry

5 mmag (Rubin)
20 to 30 µas (Gaia)

LSST, Gaia



Matter power spectrum

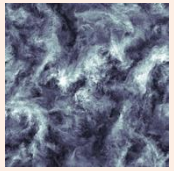
1 % to k = 1
3 % to k = 10 h/Mpc

N-body code comparison

Posterior coverage

coverage test passes
what else must hold when the model is wrong?

? simulation-based inference



Simulation observables

energy spectrum, stream width, halo mass function: which, to what %?

? you

AI surrogate error budget

per observable, per regime, with a fallback rule

? together

Your science product

name it, give its tolerance and a test case

? this week

SKA-TEL-SKO-0000007 Rev 02 · Euclid Collaboration 2024 · LSST LPM-17 · Gaia EDR3 · Schneider et al. 2016 (JCAP) · Lemos et al. 2023 (TARP) · Modi et al. 2023.

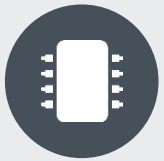
What we have, know how to do, and a roadmap for collaboration

What we need from you: a science product, its tolerance, a reference pipeline, one test case



Benchmarking pipelines

time, energy, carbon, $
& quality per result
(GPUs, CPUs)



Operators HW-SW co-design

reconfigurable-chip operators (FPGA, CGRA) @ precision your tolerance allows: gridding, FFT, calibration, inference



Simulate with invariants

SPH-EXA + particle-to-grid + AI surrogates under one scheduler, automatic invariant check

This week

Thursday **OneAstronomy** discussion:
bring one tolerance per science case

2027

fidelity-map workshop:
joint **astroCAMP** + **OneAstroBench**
with quality metrics

2027 to 2028

joint proposal: AI + numerical co-design
for SKA and survey simulations
shared benchmarks, hardware prototypes

Astro + computer science co-design of numerical systems to serve

more valid science per joule, per $, per telescope-year

Thank you!

Open Science & Artifacts

SPH-EXA

exascale SPH + gravity, Cornerstone octree, P2G (MIT)

github.com/unibas-dmi-hpc/SPH-EXA

astroCAMP

SKA imaging benchmark: code, 18 metrics, traces; data on Zenodo

github.com/SEAMS-Project/astroCAMP

DaphneSched

scheduler for integrated HPC + AI + data pipelines (DAPHNE)

github.com/daphne-eu/daphne

CEO-DC

AI cluster procurement planning
Carbon + Economy analysis framework

github.com/denisa-c/gridpilot

Also open (or in the process)

LB4OMP, **DLS4LB** (Basel) | **cgra_x_heep** | **SPHYNX** (Cabezón) | **astroCAMP data** v1.0: zenodo 20093790 | papers: **astroCAMP** PASC 26, **CEO-DC** 25, **GridPilot** Euro-Par 26, **TAC-Co** ASP-DAC 27, **Cornerstone** PASC 23

Machines we can run with you

HPCaTS
High Performance & Privacy through Confidential Computing, Trust and Security



HPCaTS (GH200, MI300A) | **Alps** (CSCS, GH200)
LUMI-G (EuroHPC, MI250X) | **Kuma** (EPFL, H100)
sciCORE (UniBasel): allocations & energy instrumentation

With **thanks to** S. Keller (CSCS), O. Simsek, L. Mayer, A. Refregier, D. Atienza, M. Peón-Quirós, E. Orliac, X. Ouvrard, the SEAMS, SKACH, SPH-EXA2 and ARTS4SKA teams, and EuroHPC JU (LUMI-G)

Funding, projects

3rd International Workshop on AI + Astronomy | Zhejiang Lab, Hangzhou | 19 Oct 2026

How Much Compute Does Good Science Need?

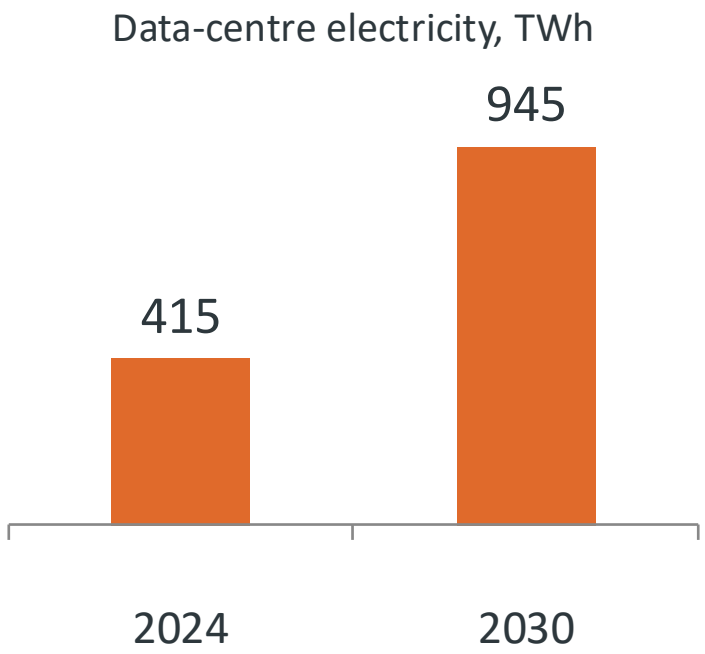
Co-Designing AI and Numerical Methods for Astronomy

Dr. Denisa Constantinescu
University of Basel, DMI HPC Group | EPFL ESL
Dr. Rubén Cabezón
University of Basel, sciCORE
Rubén Rodríguez Álvarez
EPFL, Embedded Systems Laboratory
Prof. Florina Ciorba
University of Basel, DMI HPC Group

denisaandreea.constantinescu@unibas.ch | denisa-c.com

Backup | The budgets are finite: joules, francs, megawatts at the telescope

Six numbers from six different instruments and studies, one message



IEA, Energy and AI (2025): 2030 is more than Japan uses today



22 t CO_2e / yr

per astronomer, from supercomputing alone

Stevens et al. 2020, Australian astronomers



1 MW per SDP

SKA cap per data processor: 2.8 to 29.5 kt CO_2 / yr

SKA Observatory



835 000 node-h

Euclid Flagship 2: 4 trillion particles

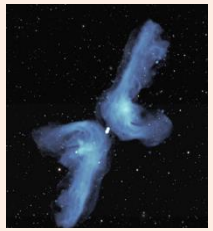
PKDGRAV3 on Piz Daint, CSCS



22 M GPU-h

SPH-EXA campaign, largest EuroHPC allocation

LUMI-G, Finland



4 to 14 %

of peak reached by radio-imaging pipelines

astroCAMP, PASC 2026

50 years

SKA lifetime: inefficiency paid for half a century

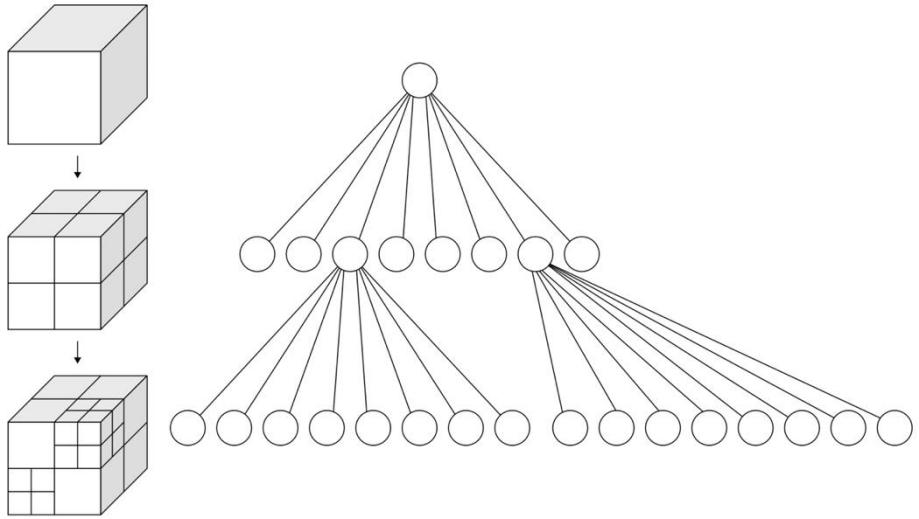
SKA Observatory

More science per joule is not an ethics slide. It is how the science we plan fits the budgets we have

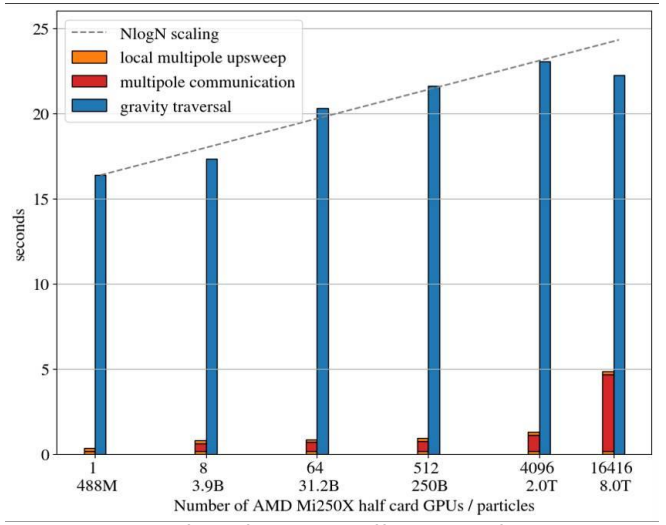
IEA, Energy and AI (2025) · Stevens et al., Nat. Astron. 4, 843 (2020) · SKAO · Stadel, ISPDC 2022 (Euclid Flagship 2) · EuroHPC EHPC-EXT-2023E01-031 · Constantinescu et al., astroCAMP, PASC 2026.

Backup | SPH-EXA under the hood

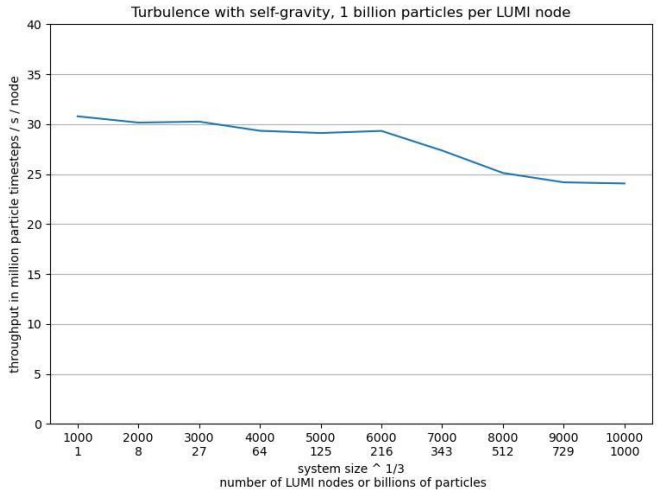
Cornerstone octree on a space-filling curve, everything resident on the GPU



Cornerstone: octree cells = key ranges of the Hilbert curve; no per-particle pointers



Gravity weak scaling to 8 trillion particles, 16 416 GPUs, LUMI-G



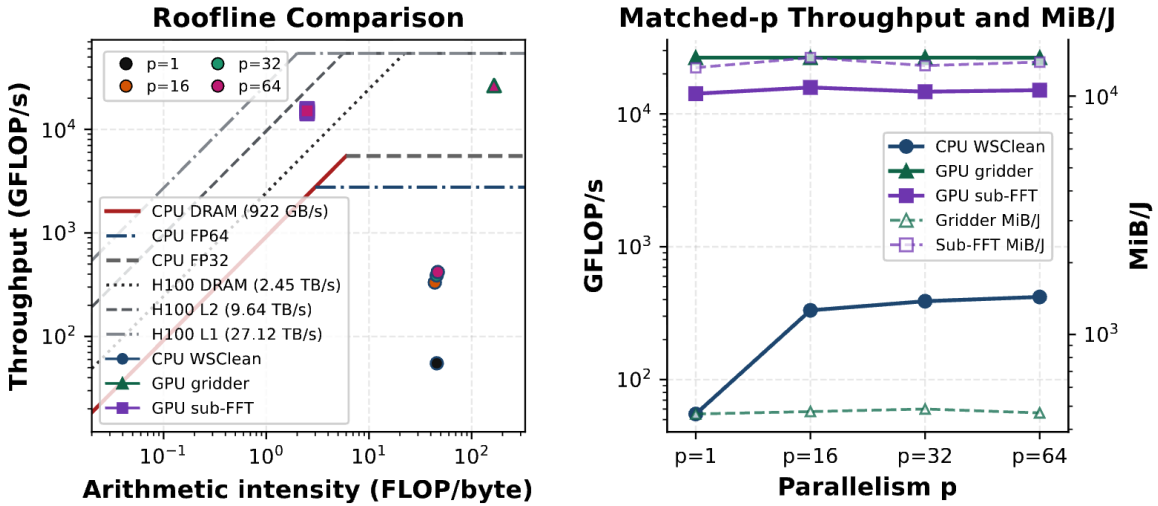
Hydro + turbulence + gravity: 1 billion particles per node, to 1 trillion

- Every rank knows the global tree and builds a locally essential tree for Barnes-Hut gravity (Ryoanji, 70 to 80 % of A100 peak)
- Domain decomposition every step (10 to 20 % of a step): assignment, particle exchange, halo discovery; halos within ~2h, 100 to 150 neighbours per particle
- MPI + OpenMP + CUDA / HIP, C++20, MIT licence; double precision in all production runs; mixed precision is future work

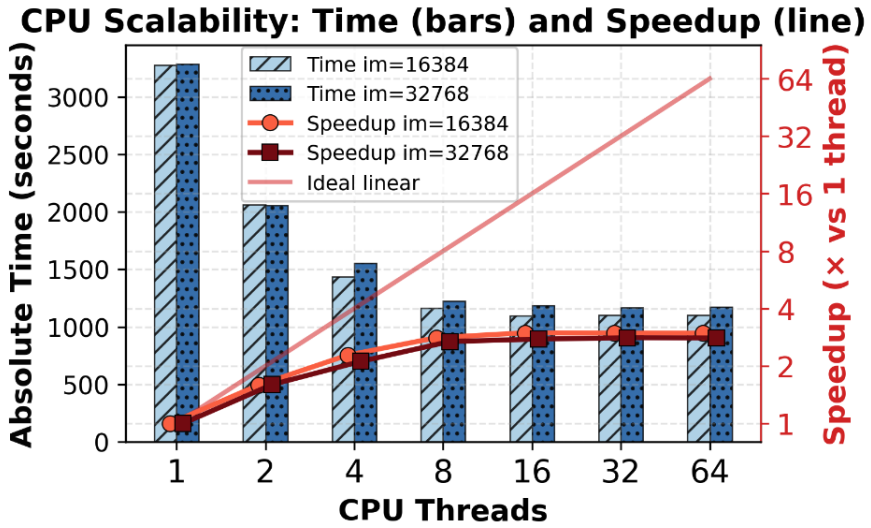
Keller, Cavelan, Cabezón, Mayer, Ciorba, PASC 2023, arXiv:2307.06345 · Ciorba, WSSP35 2023 · github.com/sekelle/cornerstone-octree.

Backup | astroCAMP: setup, metric suite, roofline

What was measured and how



CPU and GPU roofline: memory- and orchestration-bound, not compute-bound



CPU strong scaling: 2.82× maximum, efficiency 4.4 % at 64 threads

- Node: Lenovo SR675 V3 (Kuma, EPFL), 2× EPYC 9334, 4× H100 94 GB; runs use 16 cores + 1 H100; server 29 875 USD, 513 $kgCO_2e$ embodied
- 16 OSKAR SKA-Low datasets (512 stations, GLEAM sky), time × channel ∈ {1, 8, 64, 128, 256}; images 4096^2 to 32768^2; 100 monitored runs
- Instrumentation: PDU, RAPL, NVML, PMT, AMD uProf, Nsight, I/O counters; 18 metrics in 6 families; all figures regenerable from the repository

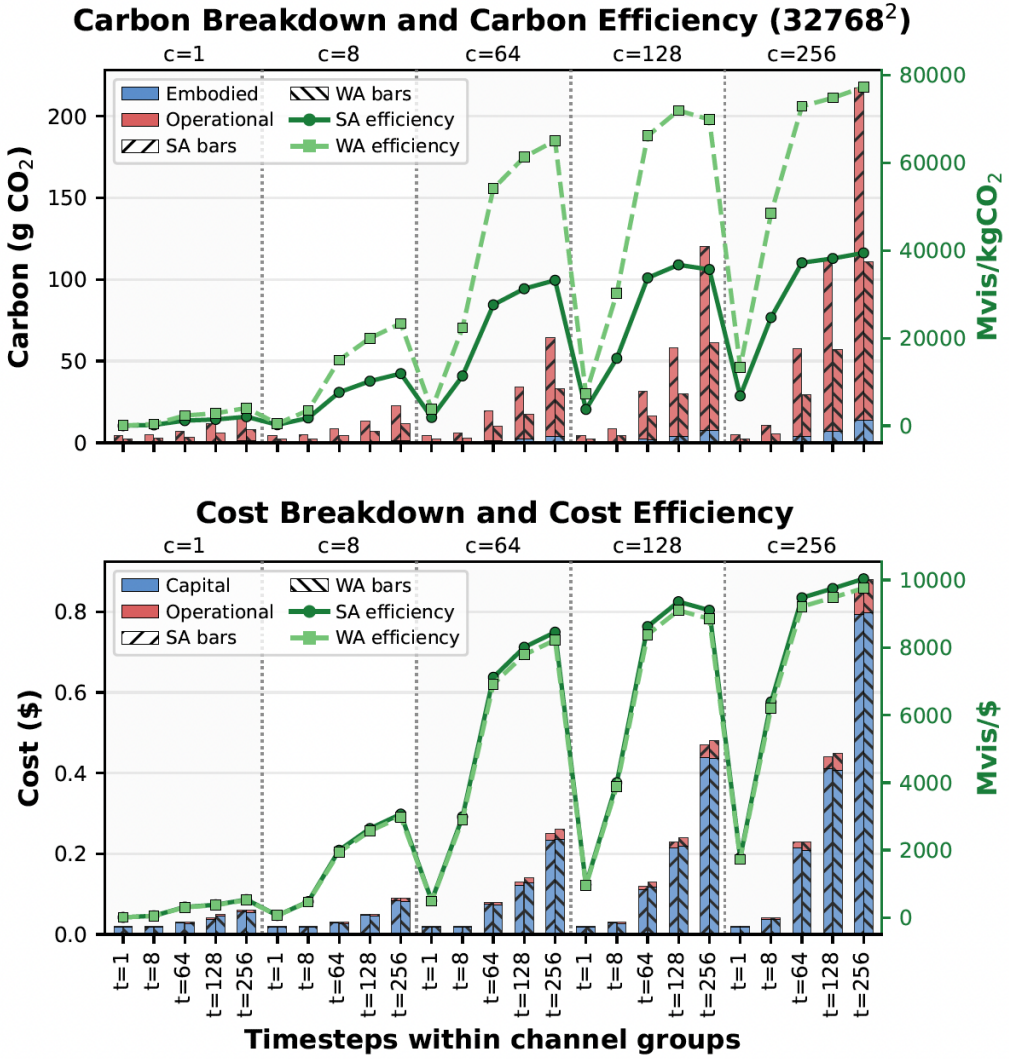
Constantinescu et al., astroCAMP, PASC 2026, arXiv:2512.13591, Sections 5 and 6, Figures 10 to 13.

Backup | CEO-DC: carbon and cost are different decisions

Rodríguez Álvarez, Constantinescu et al., 2025

SKA imaging on EPYC + H100: carbon (top) and cost (bottom)

Rodríguez Álvarez*, Constantinescu* et al., CEO-DC, arXiv:2507.08923 (2025).



- Operational emissions 90 to 95 % of carbon; hardware CAPEX 89 to 92 % of cost
- 72 % of AI platforms improve slower than demand grows (6.6× per 4 years)
- Replacing > 4-year-old platforms in high-carbon grids cuts emissions ≥ 75 %
- Modular servers save 50 to 75 % embodied carbon
- GPU prices +24 % per year (V100 to B200)
- Energy efficiency improves 18 % slower than performance
- Carbon-optimal and cost-optimal choices diff country