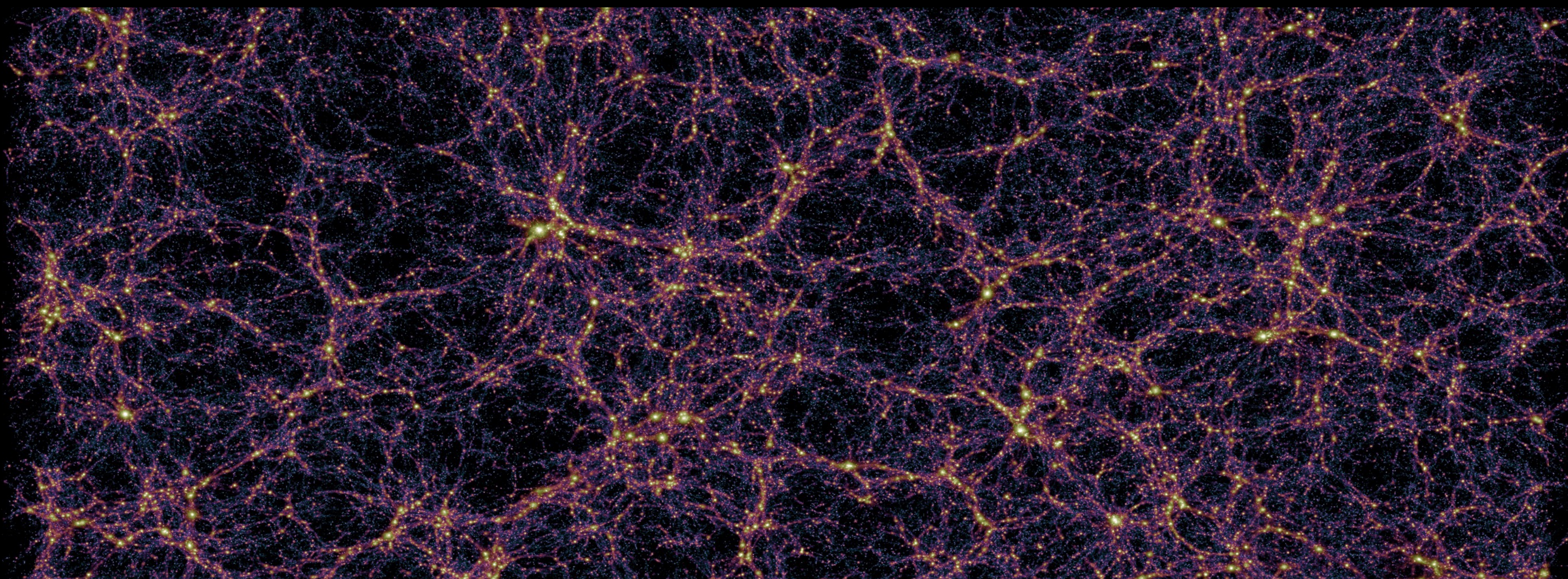
Cosmological Hydrodynamics at Exascale

A Trillion-particle Leap in Capability



Nicholas Frontiere
CPS Division, Argonne National Laboratory

ALCF Webinar
Jan 28th 2026

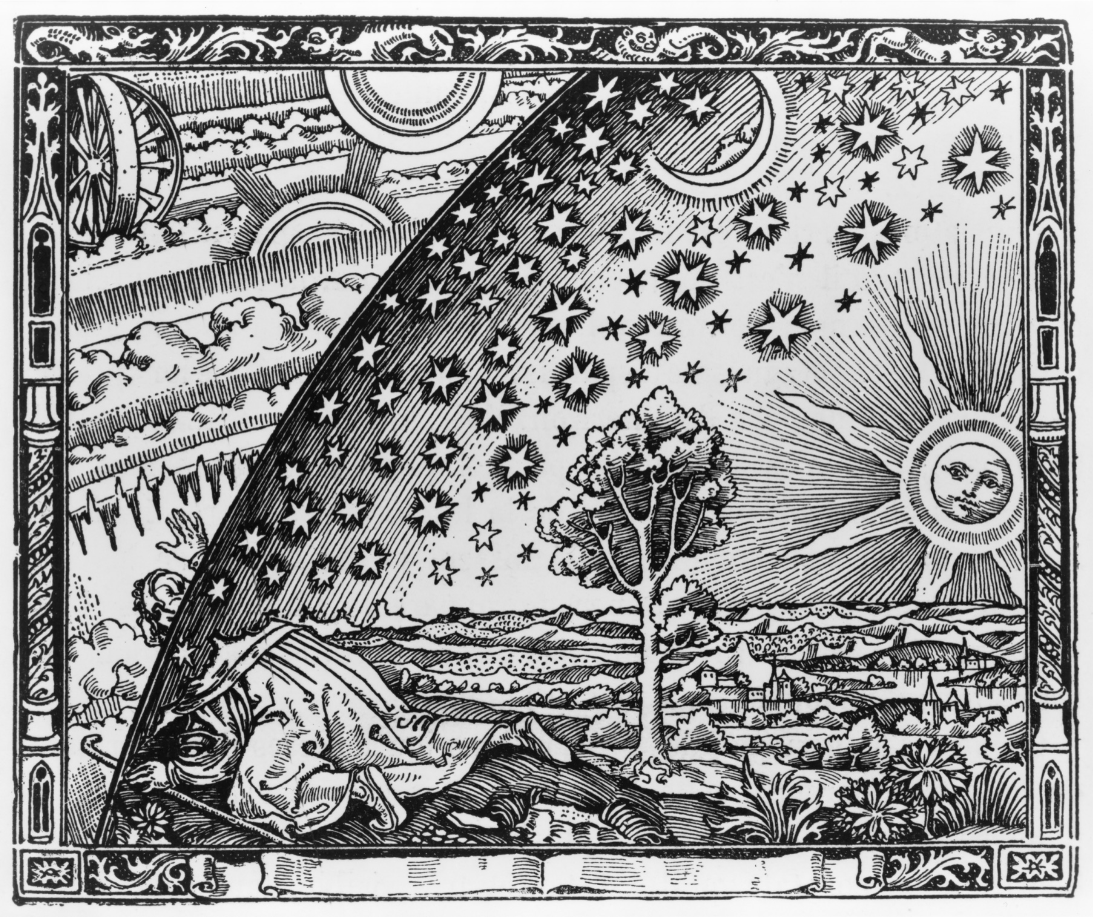
Cosmological Simulations

Modern Cosmology

- The study of the nature of the Universe
- Characterize the properties of dark matter and dark energy
- Use cosmology as a probe of fundamental physics

Observations and Simulations

- A new generation of extremely sophisticated telescopes
- Simulations provide theoretical forward model predictions
  - Mock realizations of the full sky at the same precision as surveys
  - **CRK-HACC** cosmology framework designed for exascale machines to run the largest ever cosmological simulations



Flammarion Engraving

Simulation Characteristics

Domain

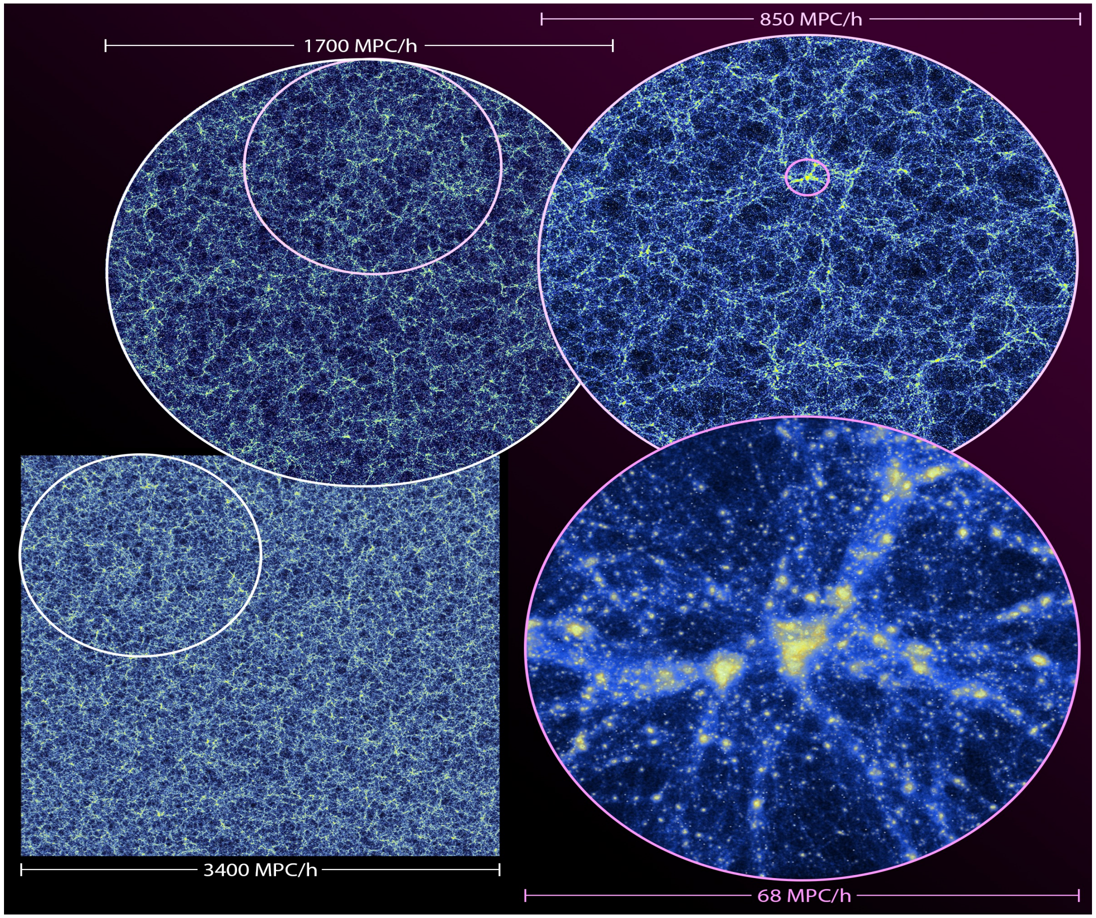
- Extent of visible Universe – billions of light years
- Particle mass equal to many millions of stars (~10^{13} Earths)
- Dynamic range of a million to one in both mass and time (Gyr-kyr)

Gravity – N-body

- Vlasov-Poisson equation
- Dominates on intergalactic scales
- Sampled with trillions of particles

Gas Physics – CRKSPH

- Euler fluid equations
- Radiative cooling and heating
- Supernovae and star formation
- Supermassive black hole feedback



Last Journey Simulation

Simulation Characteristics

Domain

- Extent of visible Universe – billions of light years
- Particle mass equal to many millions of stars (~10^{13} Earths)
- Dynamic range of a million to one in both mass and time (Gyr-kyr)

Gravity – N-body

- Vlasov-Poisson equation
- Dominates on intergalactic scales
- Sampled with trillions of particles

Gas Physics – CRKSPH

- Euler fluid equations
- Radiative cooling and heating
- Supernovae and star formation
- Supermassive black hole feedback

Vlasov-Poisson Equation

$$\frac{\partial f_i}{\partial t} + \dot{\mathbf{x}}\frac{\partial f_i}{\partial \mathbf{x}} - \nabla\phi\frac{\partial f_i}{\partial \mathbf{p}} = 0, \qquad \mathbf{p} = a^2\dot{\mathbf{x}},$$

$$\nabla^2\phi = 4\pi G a^2(\rho(\mathbf{x},t) - \langle\rho_{\rm dm}(t)\rangle) = 4\pi G a^2 \Omega_{\rm dm}\delta_{\rm dm}\rho_{\rm cr},$$

$$\delta_{\rm dm}(\mathbf{x},t) = (\rho_{\rm dm} - \langle\rho_{\rm dm}\rangle)/\langle\rho_{\rm dm}\rangle,$$

$$\rho_{\rm dm}(\mathbf{x},t) = a^{-3}\sum_i m_i \int d^3\mathbf{p} f_i(\mathbf{x},\dot{\mathbf{x}},t).$$

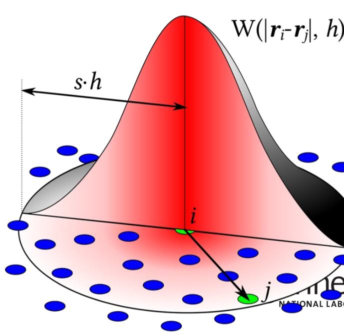
Use N-body methods to sample distribution

Euler Fluid Equations

$$\frac{du}{dt} = -\frac{P}{\rho}\nabla\cdot\boldsymbol{v} \qquad \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\rho\nabla\cdot\boldsymbol{v} \qquad \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\frac{1}{\rho}\nabla P$$

Fields are smoothed over an SPH kernel

Astrophysics processes are stochastically integrated sources of mass, momentum and energy



Simulation Characteristics

Domain

- Extent of visible Universe – billions of light years
- Particle mass equal to many millions of stars (~10^{13} Earths)
- Dynamic range of a million to one in both mass and time (Gyr-kyr)

Gravity – N-body

- Vlasov-Poisson equation
- Dominates on intergalactic scales
- Sampled with trillions of particles

Gas Physics – CRKSPH

- Euler fluid equations
- Radiative cooling and heating
- Supernovae and star formation
- Supermassive black hole feedback

Vlasov-Poisson Equation

$$\frac{\partial f_i}{\partial t} + \dot{\mathrm{x}}\frac{\partial f_i}{\partial \mathrm{x}} - \nabla\phi\frac{\partial f_i}{\partial \mathrm{p}} = 0, \qquad \mathrm{p} = a^2\dot{\mathrm{x}},$$

$$\nabla^2\phi = 4\pi G a^2(\rho(\mathrm{x},t) - \langle\rho_{\mathrm{dm}}(t)\rangle) = 4\pi G a^2\Omega_{\mathrm{dm}}\delta_{\mathrm{dm}}\rho_{\mathrm{cr}},$$

$$\delta_{\mathrm{dm}}(\mathrm{x},t) = (\rho_{\mathrm{dm}} - \langle\rho_{\mathrm{dm}}\rangle)/\langle\rho_{\mathrm{dm}}\rangle,$$

$$\rho_{\mathrm{dm}}(\mathrm{x},t) = a^{-3}\sum_i m_i \int d^3\mathrm{p}\, f_i(\mathrm{x},\mathrm{x},t).$$

Use N-body methods to sample distribution

Extreme dynamic range *and* ***complexity*** *of the underlying physics models make cosmology one of the* ***hardest*** *classes of computational problems in HPC.*

Euler Fluid Equations

$$\frac{du}{dt} = -\frac{P}{\rho}\nabla\cdot v \qquad \frac{d\rho}{dt} = -\rho\nabla\cdot v \qquad \frac{dv}{dt} = -\frac{1}{\rho}\nabla P$$

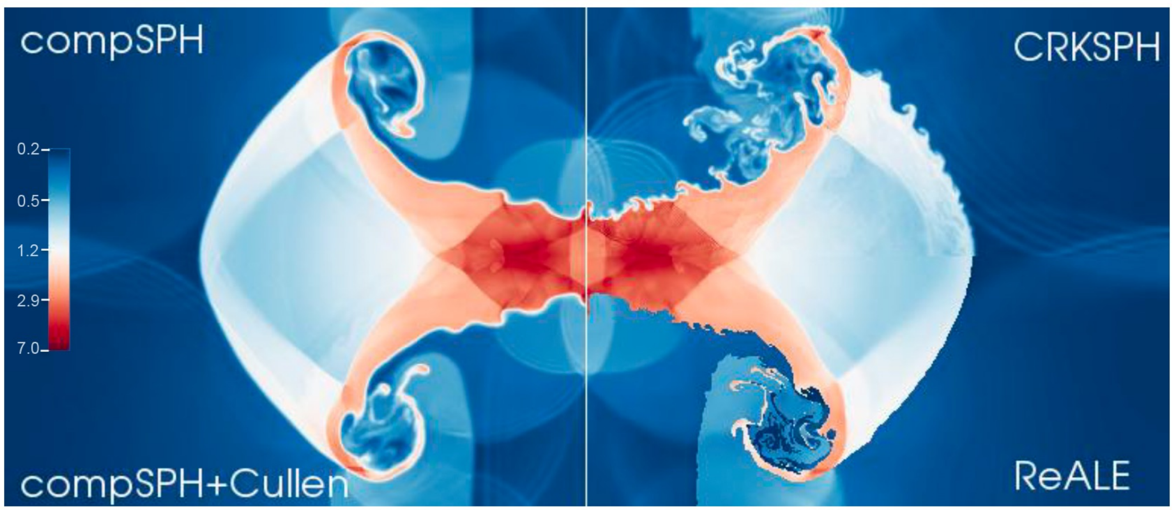
Fields are smoothed over an SPH kernel

Astrophysics processes are stochastically integrated sources of mass, momentum and energy

CRKSPH Algorithm

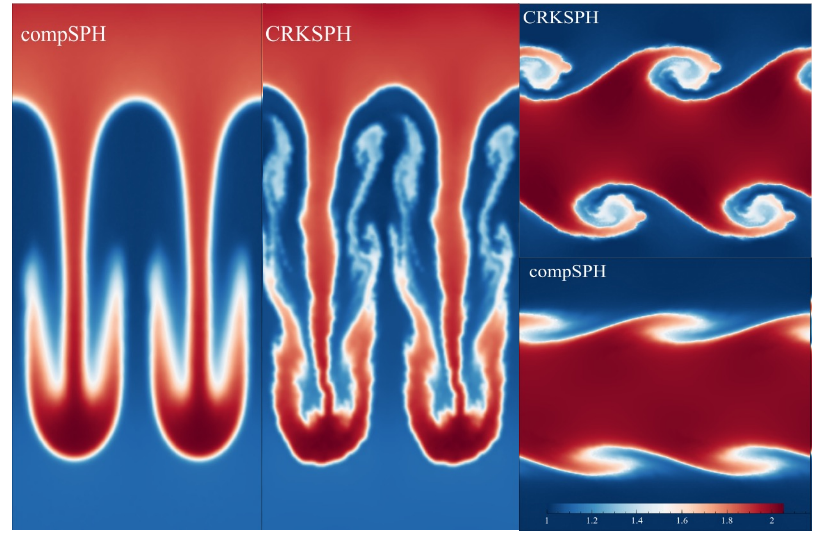
Conservative Reproducing Kernel SPH

- Smoothed particle hydrodynamics variant
- Adaptive fluid dynamics solver without a fixed grid
- Higher-order kernels exactly reproduce linear order fields
- Maintains exact conservation of energy and momentum
- Significantly reduces excessive SPH artificial diffusion
- No artificial conductivity required

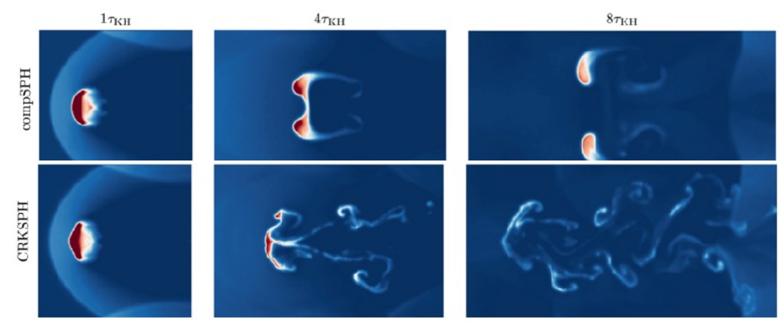


Triple-point shock

Rayleigh-Taylor **Kelvin-Helmholtz**



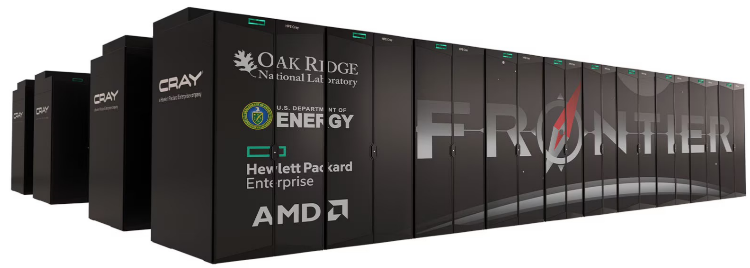
Frontiere et al., J. Comp. Phys. 332, 160 (2017)



'Blob' test with CRK-SPH showing blob disruption due to Kelvin-Helmholtz and Rayleigh-Taylor instabilities

Age of Exascale

- **Next generation of supercomputers**
  - Frontier ORNL 1.353 Eflop/s
  - Aurora ANL 1.012 Eflop/s
  - El Capitan LLNL 1.742 Eflop/s
  - GPU accelerated
  - Mixed precision



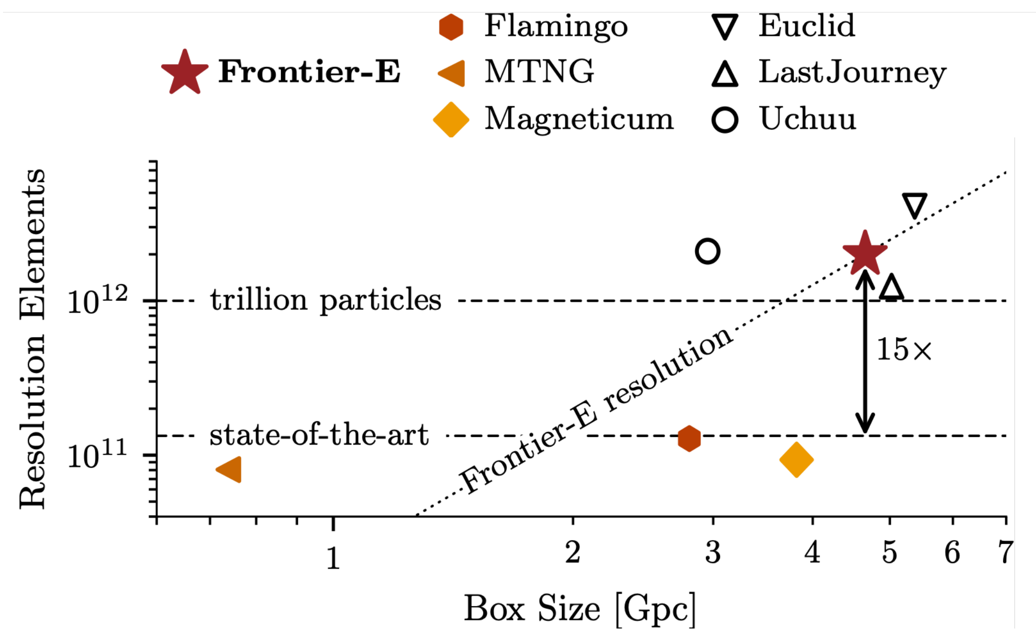
- **Exascale class capabilities**
  - 10-50x improvement in performance
  - Hydro simulations are typically 10-30x slower
  - Previous state-of-the-art GO simulations can now include gas physics (trillions of particles)
  - Next-generation GO simulations in the 10+ trillion particle class

Frontier Exascale Simulation

Frontier-E Specifications

- Evolves **4 trillion** particles
- Resolves **1 billion** galaxies
- **15x** jump in capability
- **1.8 million** node hrs on Frontier
- **100 PB** data generated

- Complete **end-to-end** simulation
- **12 PB** of stored science data
- The first leadership scale **GPU-accelerated** hydrodynamic cosmology simulation



Exascale unlocks the ability to run previous **state of the art** gravity only simulations now including **hydrodynamics** in a **week** of machine time (~ 1 yr CPU)

Gravity only simulations break into the **10+ trillion** particle class

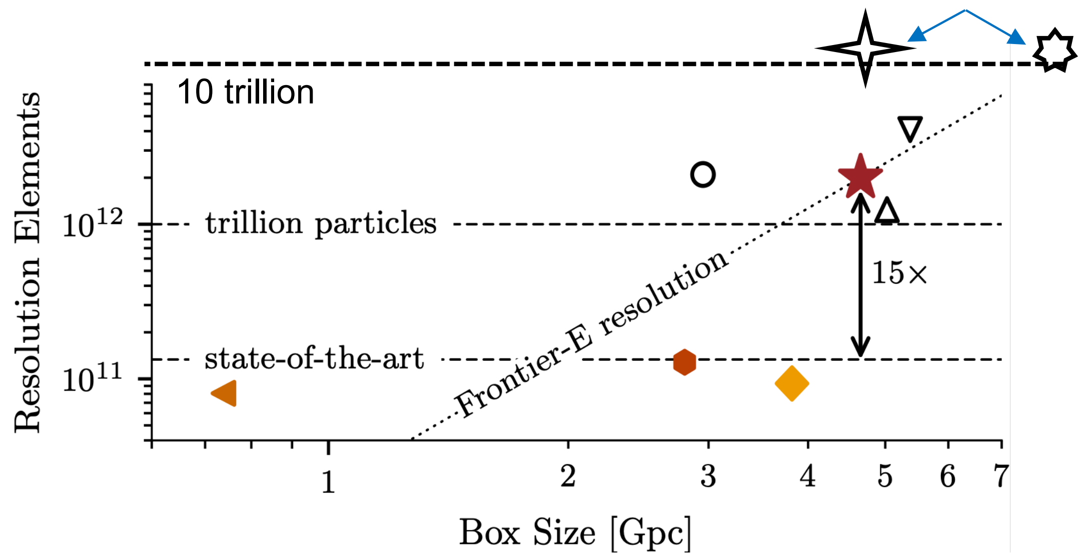
Frontier Exascale Simulation

Frontier-E Specifications

- Evolves **4 trillion** particles
- Resolves **1 billion** galaxies
- **15x** jump in capability
- **1.8 million** node hrs on Frontier
- **100 PB** data generated

- Complete **end-to-end** simulation
- **12 PB** of stored science data
- The first leadership scale **GPU-accelerated** hydrodynamic cosmology simulation

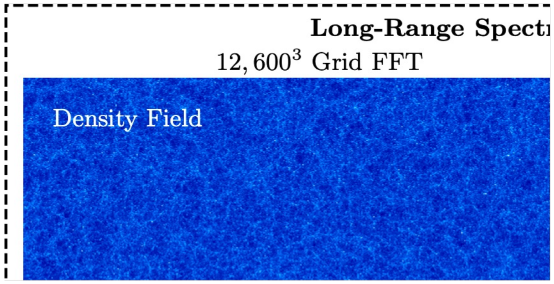
Aurora Exascale Sims



Exascale unlocks the ability to run previous **state of the art** gravity only simulations now including **hydrodynamics** in a **week** of machine time (~ 1 yr CPU)

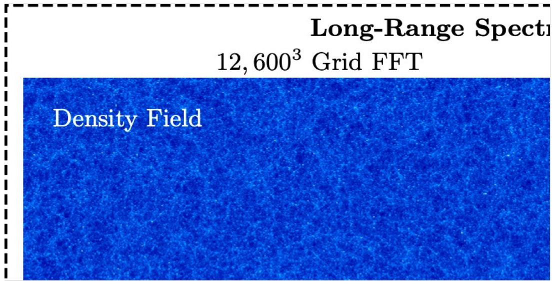
Gravity only simulations break into the **10+ trillion** particle class

CRK-HACC Runtime Overview



Four key innovations to running at exascale

CRK-HACC Runtime Overview



Separation of scales approach

- Domain divided into cuboid volumes
- Overloaded regions eliminate frequent MPI communication
- CM cells spatially decompose short-range interaction volumes
- Long-range global gravitational potential evaluated on a uniform grid of **$12{,}600^3$ cells**
- Custom high-performance FFT distribution code SWFFT
- Facilitates **ideal** weak scaling

CRK-HACC Runtime Overview

Optimized Short-range Solver

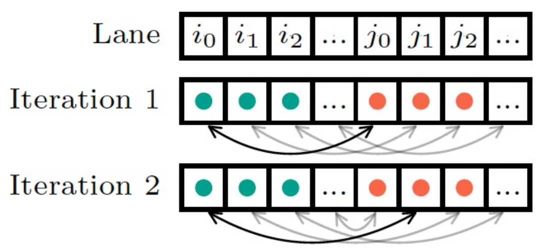
- Data GPU-resident for short-range integration – minimal transfer penalties
- CM cells subdivided into k-d trees
- Leaves constructed once and dynamically evolve – no re-indexing penalties
- Flat tree-leaf interactions for contiguous walk
- 50 GPU optimized science kernels, including short-range gravity, hydrodynamics, and stochastically integrated astrophysics models (supernovae, galactic winds, …)
- **Warp-splitting** approach for efficient particle interactions

CRK-HACC Runtime Overview

Algorithm 1 Warp Splitting Example

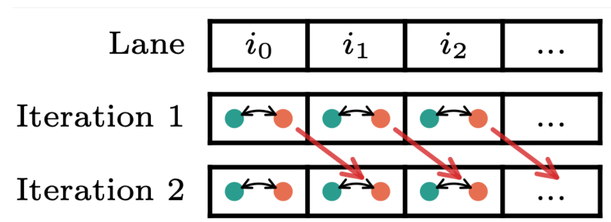
Warp loads particle data from global memory: $\mathbf{r}, \alpha, \ldots$ (half threads load from leaf i, other half from leaf j)
for each partner j in half-warp **do**
    Exchange data via warp shuffle: $\mathbf{r}_j$ = shuffle($\mathbf{r}_i$)
    Eval $f_i(\alpha_i, \ldots)$, $g_i(\alpha_i, \ldots)$, $h_{ij}(|\mathbf{r}_i - \mathbf{r}_j|, \ldots) \cdots$
    Exchange partials: g_j = shuffle(g_i)
    Eval $\phi_{ij} = f_i * g_j * h_{ij} \cdots$
    Accumulate $\phi_i \mathrel{+}= \phi_{ij}$
end for
Perform atomic update of ϕ_i to global memory

Illustration of the first two iterations of warp-shuffle operations on a single split warp



2x16[2] Interactions

Common interaction kernels use 2x the memory



32[2] Interactions

$$\phi_{ij} = f_i(\alpha_i, \ldots) * g_j(\alpha_j, \ldots) * h_{ij}(|\mathbf{r}_i - \mathbf{r}_j|, \alpha_i, \ldots) \cdots ,$$

Generalized Interaction

CRK-HACC Runtime Overview

Accelerated Analysis

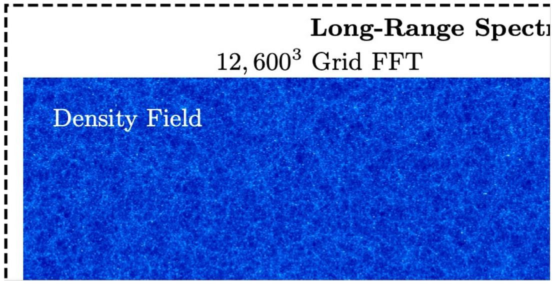
- Too expensive to post-process PBs of scientific data
  - Must be performed in situ
- Galaxy and halo finding are the most computationally demanding operations
- DBSCAN and spherical range queries
- Co-developed the publicly available ArborX library: GPU-native spatial indexing and traversal routines
- All analyses are performed on the GPU as the simulation progresses

CRK-HACC Runtime Overview

Asynchronous I/O

- Low mean time to interrupt (MTTI) for full machine jobs is problematic (both in the AI and HPC spaces). MTTI ~ hr(s)
- Mitigation – checkpoint simulation as often as possible
- Challenge – each checkpoint is hundreds of TB
- Solution: Staged I/O
  - Data is synchronously copied into fast memory (NVMe or RAM disk)
  - Files are asynchronously bled to PFS while simulation continues
  - I/O pattern optimal (no file contention), maximizes PFS bandwidth

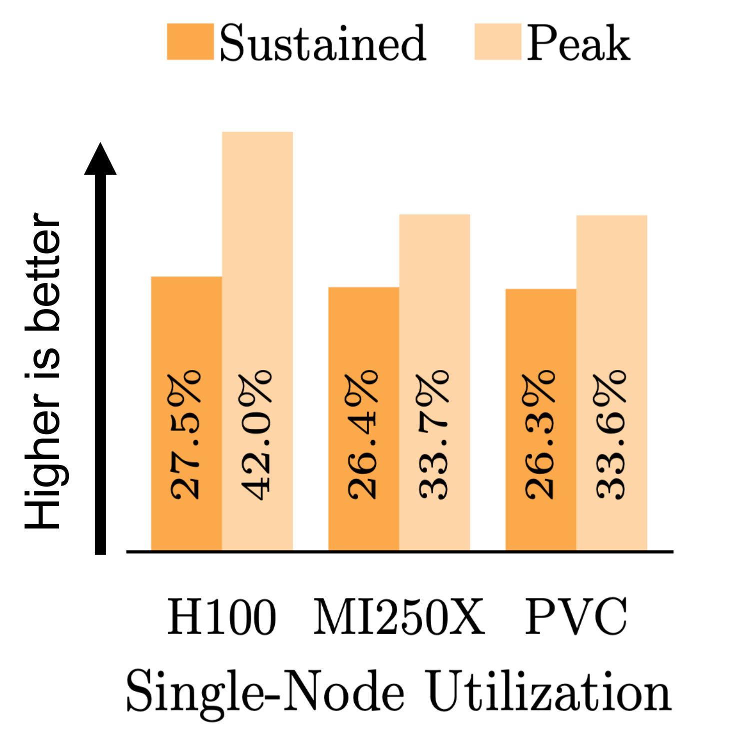
CRK-HACC Runtime Overview



All major components are portable, performant, and generalizable

CRK-HACC Portability

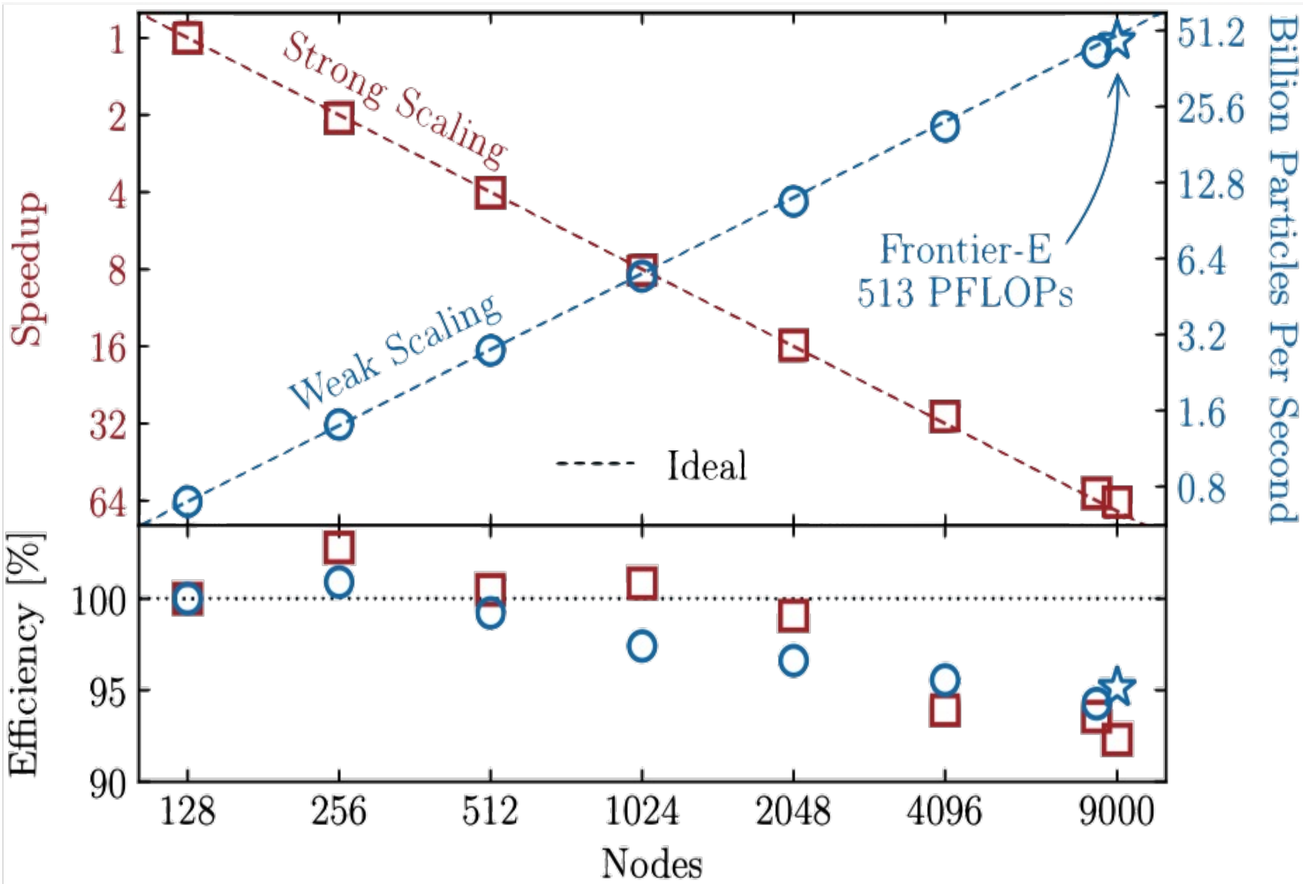
- All solver components written in vendor specific languages (CUDA, HIP, SYCL, …) optimized for each GPU supercomputing hardware
- Consistent performance across Nvidia, AMD, and Intel hardware of similar generation GPUs
- High-level software abstraction layers maintains one source code
- Dedicated study on performance portability*



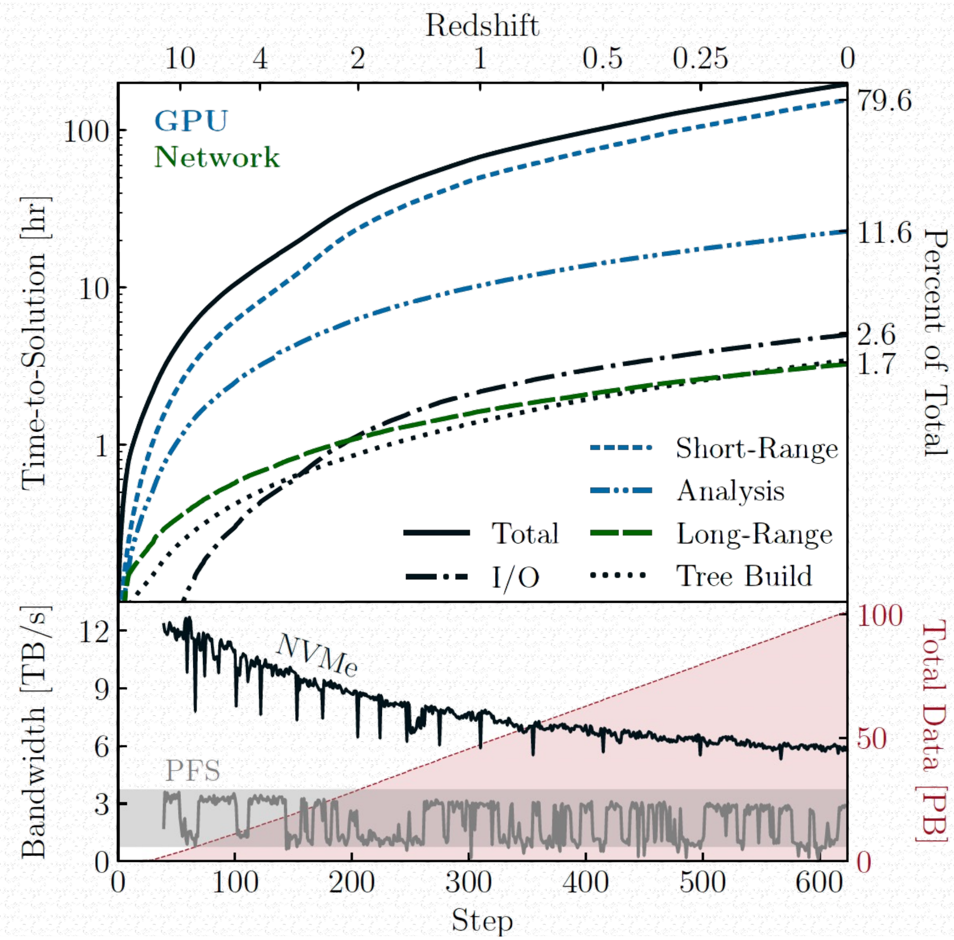
*Rangel et al. (SC-W ’23), A Performance-Portable SYCL Implementation of CRK-HACC for Exascale

Frontier-E Scaling

- **Near-ideal** weak (**95%**) and strong (**92%**) scaling up to **9,000** nodes
- Achieved **513 PFLOPs** peak performance
  420 PFLOPs sustained over the entire application
- CRK-HACC processes forces of **46.6 billion particles per second**

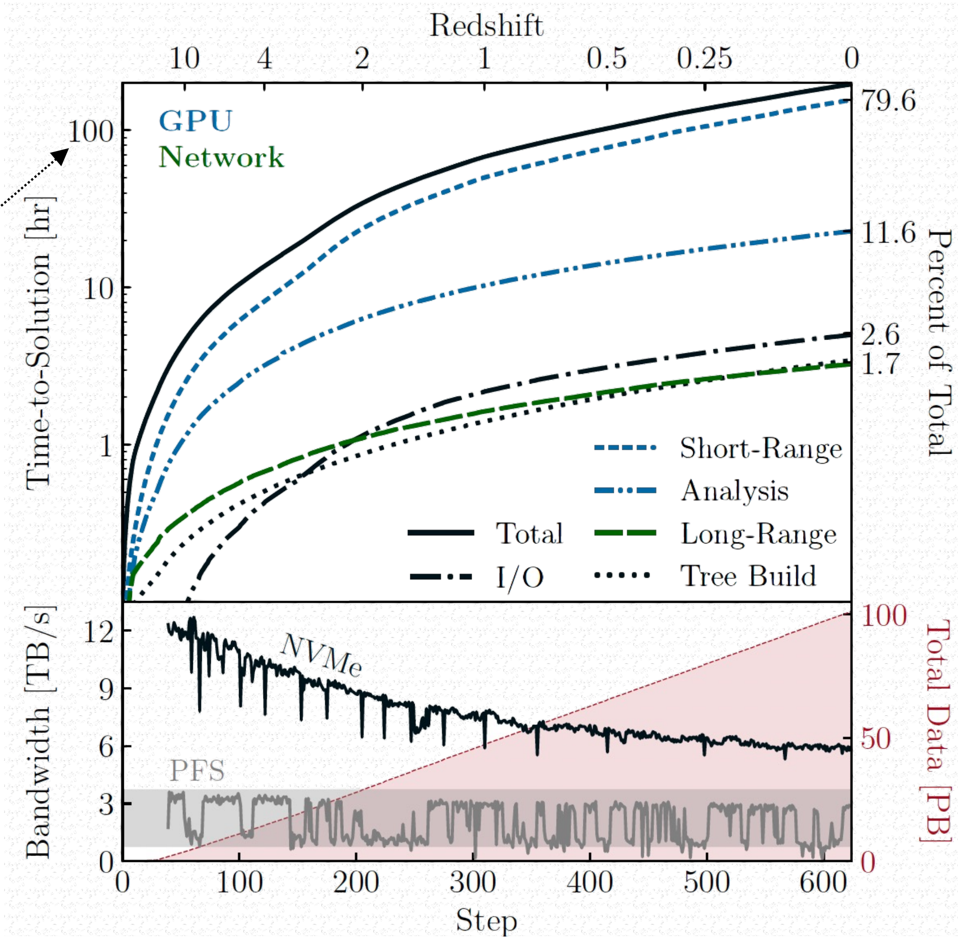


Frontier-E Cumulative Time to Solution



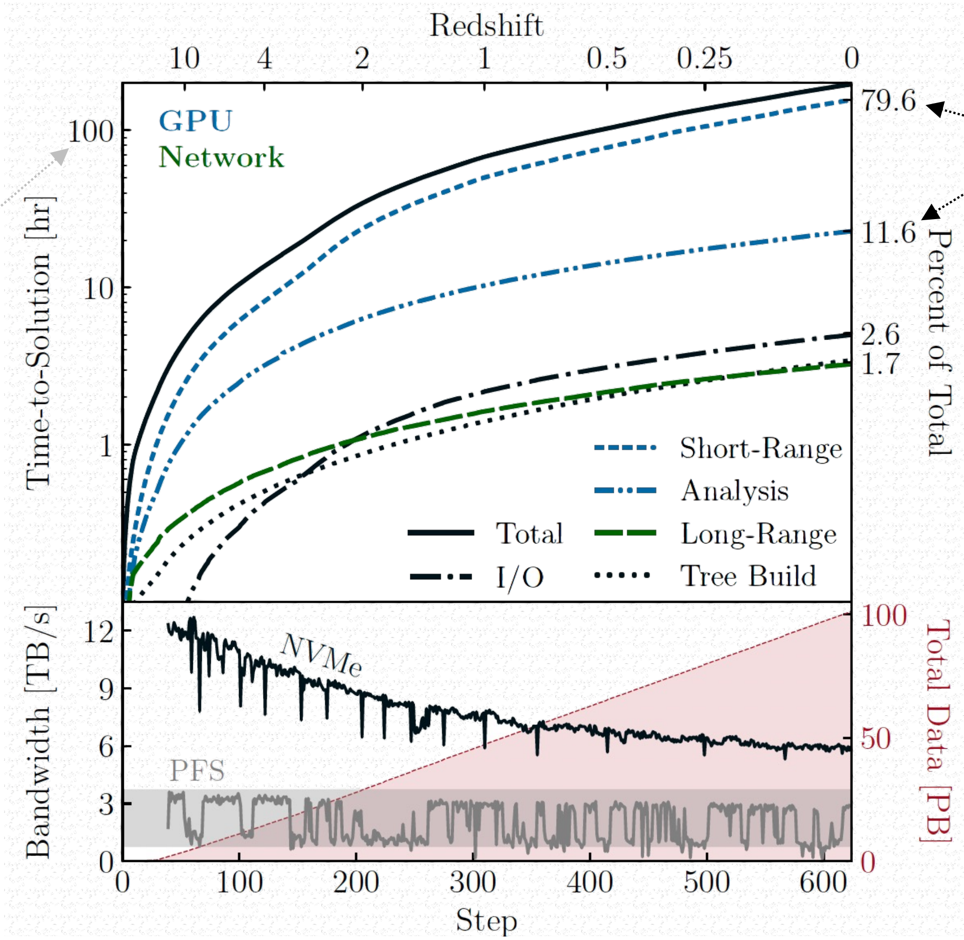
Frontier-E Cumulative Time to Solution

Completed in 1 week of machine time.
On the CPUs of Frontier, the same simulation would take **~1 year** to run



Frontier-E Cumulative Time to Solution

Completed in 1 week of machine time.
On the CPUs of Frontier, the same simulation would take **~1 year** to run

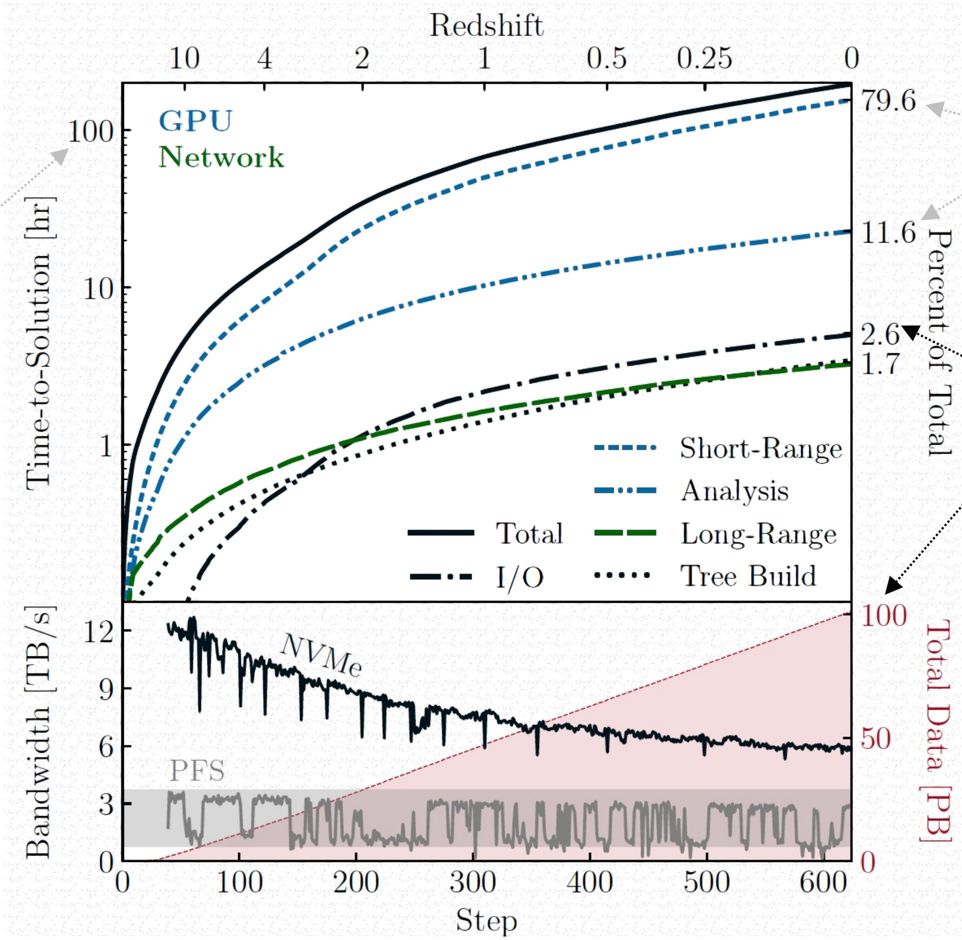


Over 90% TTS was performed on the GPU

Frontier-E Cumulative Time to Solution

Completed in 1 week of machine time.

On the CPUs of Frontier, the same simulation would take **~1 year** to run



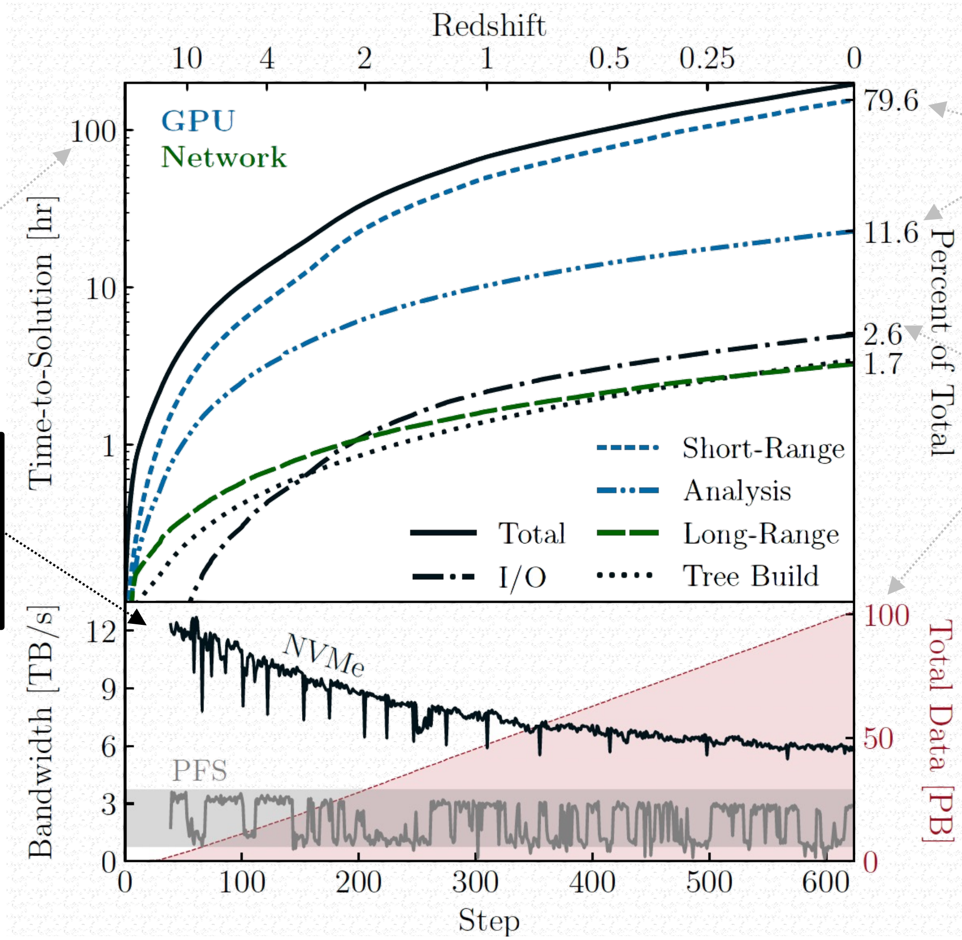
Over 90% TTS was performed on the GPU

I/O was 2.6% despite writing over 100 PB of data

Frontier-E Cumulative Time to Solution

Completed in 1 week of machine time.
On the CPUs of Frontier, the same simulation would take **~1 year** to run

Asynchronous writes using NVMe provides high-bandwidth I/O



Over 90% TTS was performed on the GPU

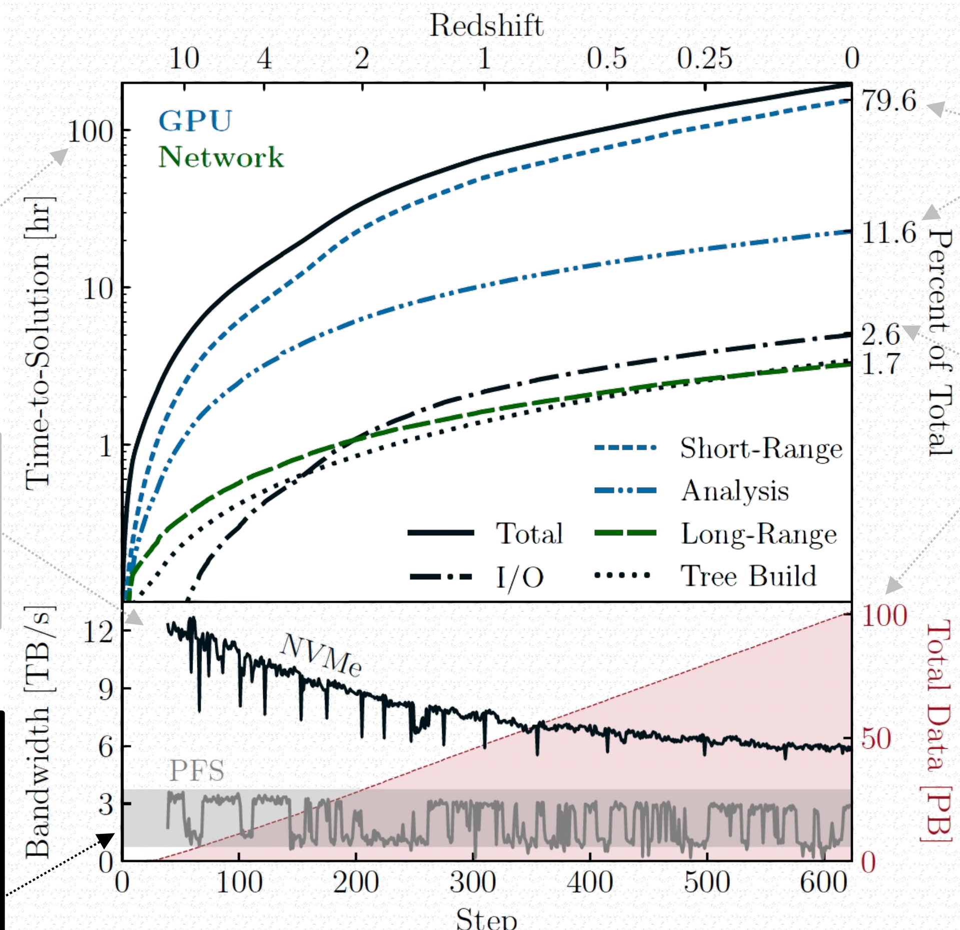
I/O was 2.6% despite writing over 100 PB of data

Frontier-E Cumulative Time to Solution

Completed in 1 week of machine time.
On the CPUs of Frontier, the same simulation would take **~1 year** to run

Asynchronous writes using NVMe provides high-bandwidth I/O

Continuous bleeds mitigate PFS variability without impacting simulation



Over 90% TTS was performed on the GPU

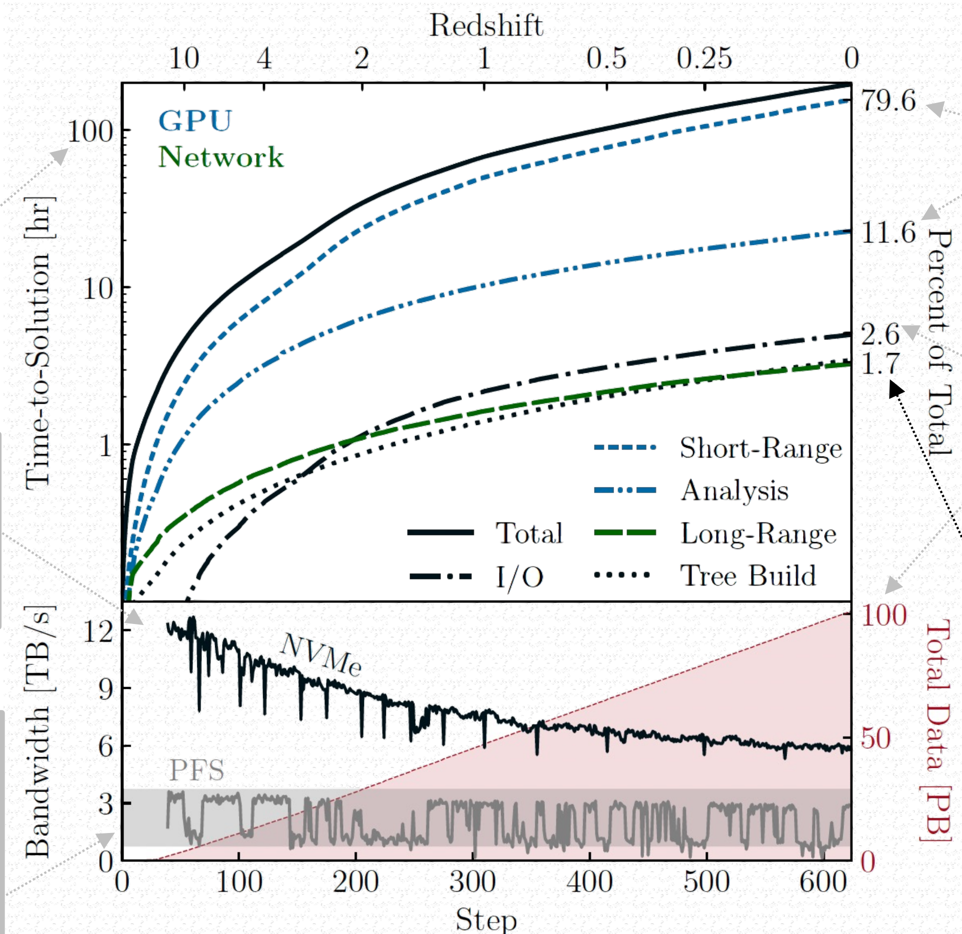
I/O was 2.6% despite writing over 100 PB of data

Frontier-E Cumulative Time to Solution

Completed in 1 week of machine time.
On the CPUs of Frontier, the same simulation would take **~1 year** to run

Asynchronous writes using NVMe provides high-bandwidth I/O

Continuous bleeds mitigate PFS variability without impacting simulation



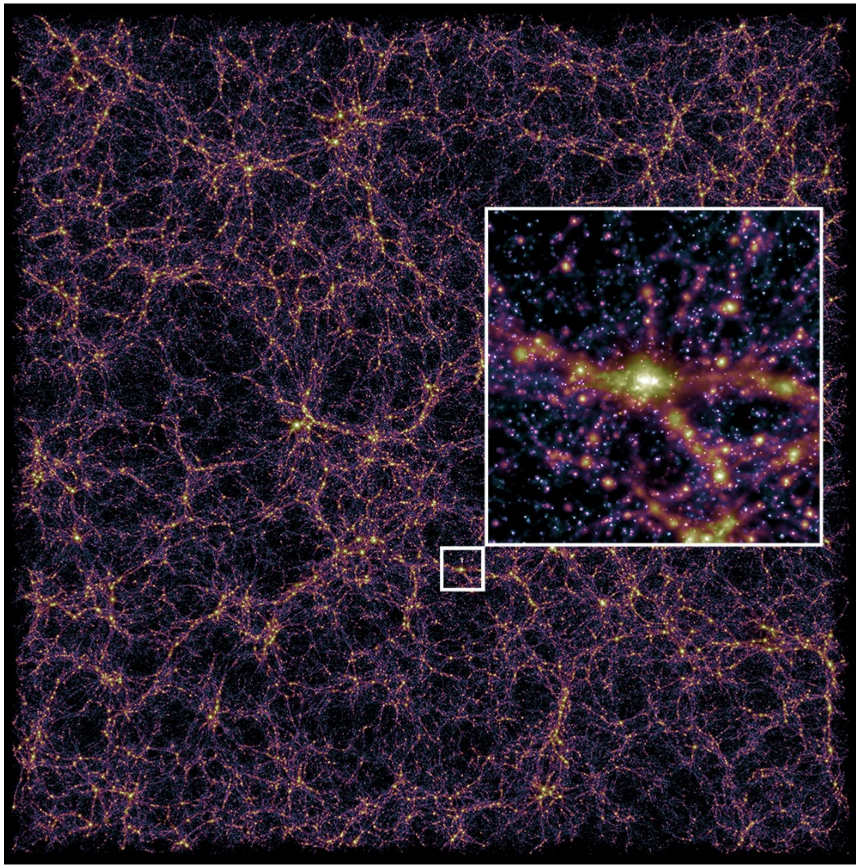
Over 90% TTS was performed on the GPU

I/O was 2.6% despite writing over 100 PB of data

Long-range solver subdominant despite running FFTs on more than two trillion cells

The Exascale Future is Bright

- Frontier-E is the first leadership-scale, GPU-accelerated hydrodynamic cosmology simulation, delivering a **15× leap in capability**
- CRK-HACC **innovations** – separation of scales, warp-splitting, in situ cluster finding, and staged I/O – **generalize** broadly across particle-based science domains
- Strong/weak (**92/95%**) scaling to **9,000 nodes** and **>500 PFLOPs** peak performance generating **100 PB** of data in a **week** of simulation time
- Exascale unlocks a **new era** of predictive cosmology with simulations positioned for next generation observational surveys



slice through downscaled Frontier-E simulation raytracing visualization

The CRK-HACC Team

Nick Frontiere

J.D. Emberson

Michael Buehlmann

Esteban Rangel

Salman Habib

Katrin Heitmann

Patricia Larsen

Vitali Morozov

Adrian Pope

Claude-André Faucher-Giguère

Antigoni Georgiadou

Damien Lebrun-Grandié

Andrey Prokopenko

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

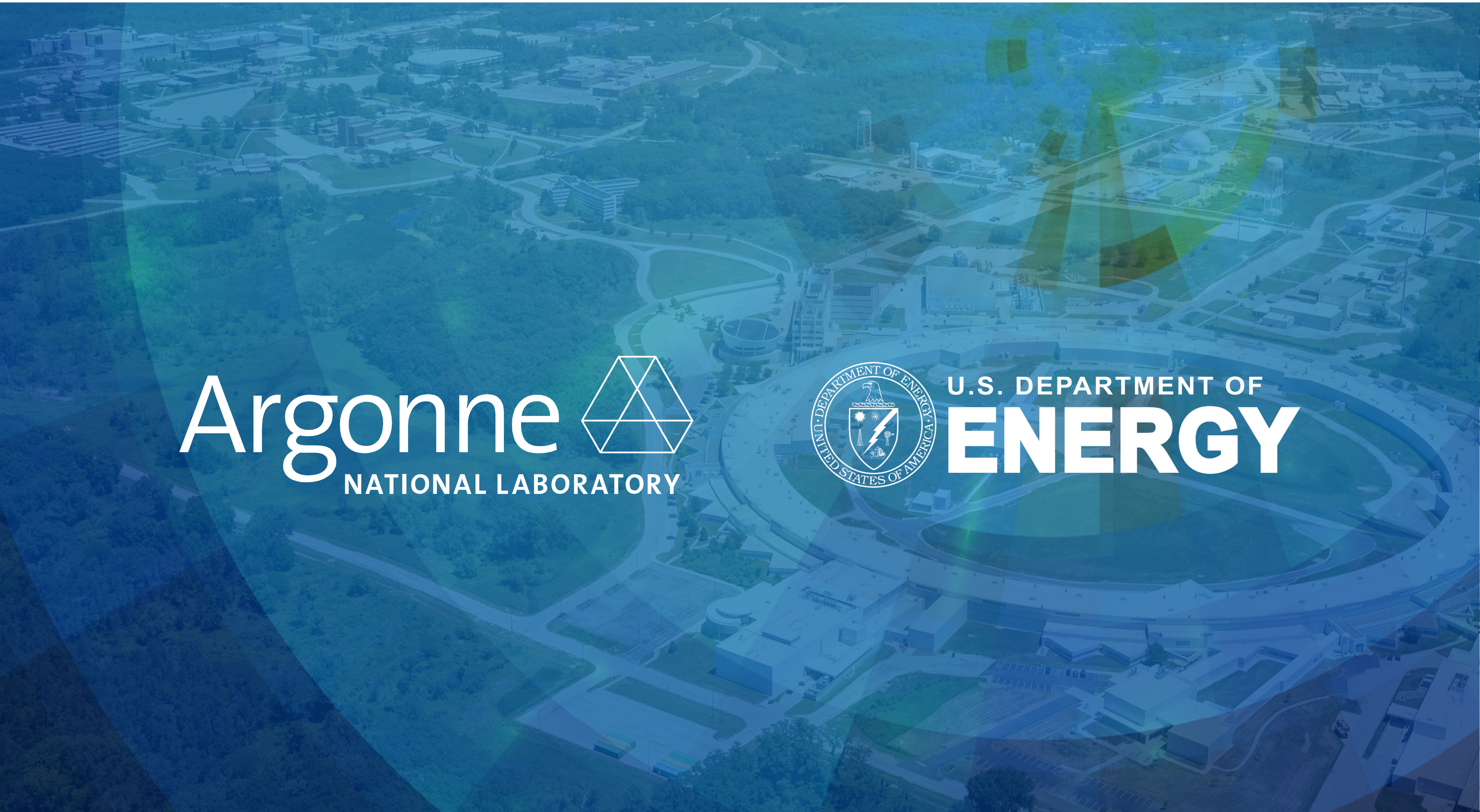
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Northwestern University

OAK RIDGE National Laboratory

Argonne NATIONAL LABORATORY

Questions?


Argonne
NATIONAL LABORATORY
U.S. DEPARTMENT OF
ENERGY