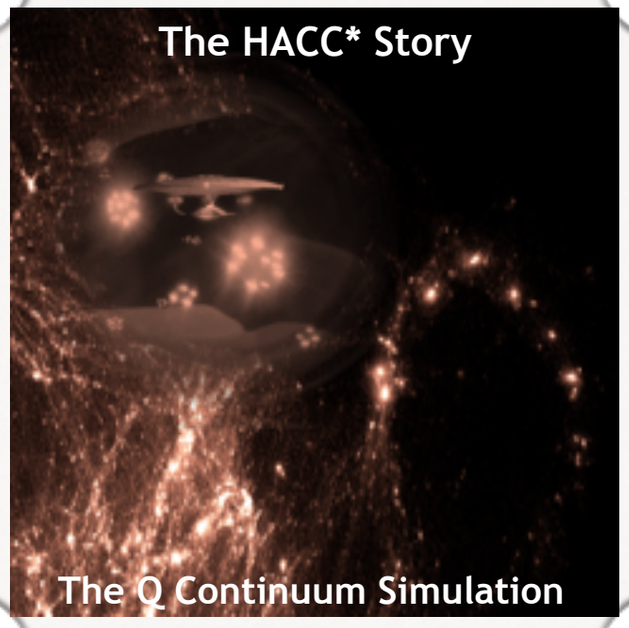
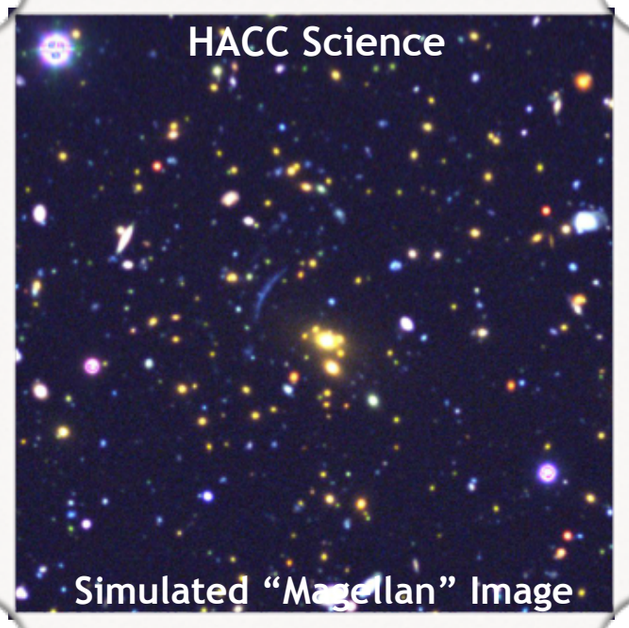
HACC: Simulating Sky Surveys on State-of-the-Art Supercomputer Architectures

Katrin Heitmann

Davos, February 17, 2017





* HACC = Hardware/Hybrid Accelerated Cosmology Code

Thanks to many collaborators!

● Physics/Astrophysics (Cosmology, Field Theory, Lattice QCD) ● Computer Science ● Statistics

Thanks to many collaborators!

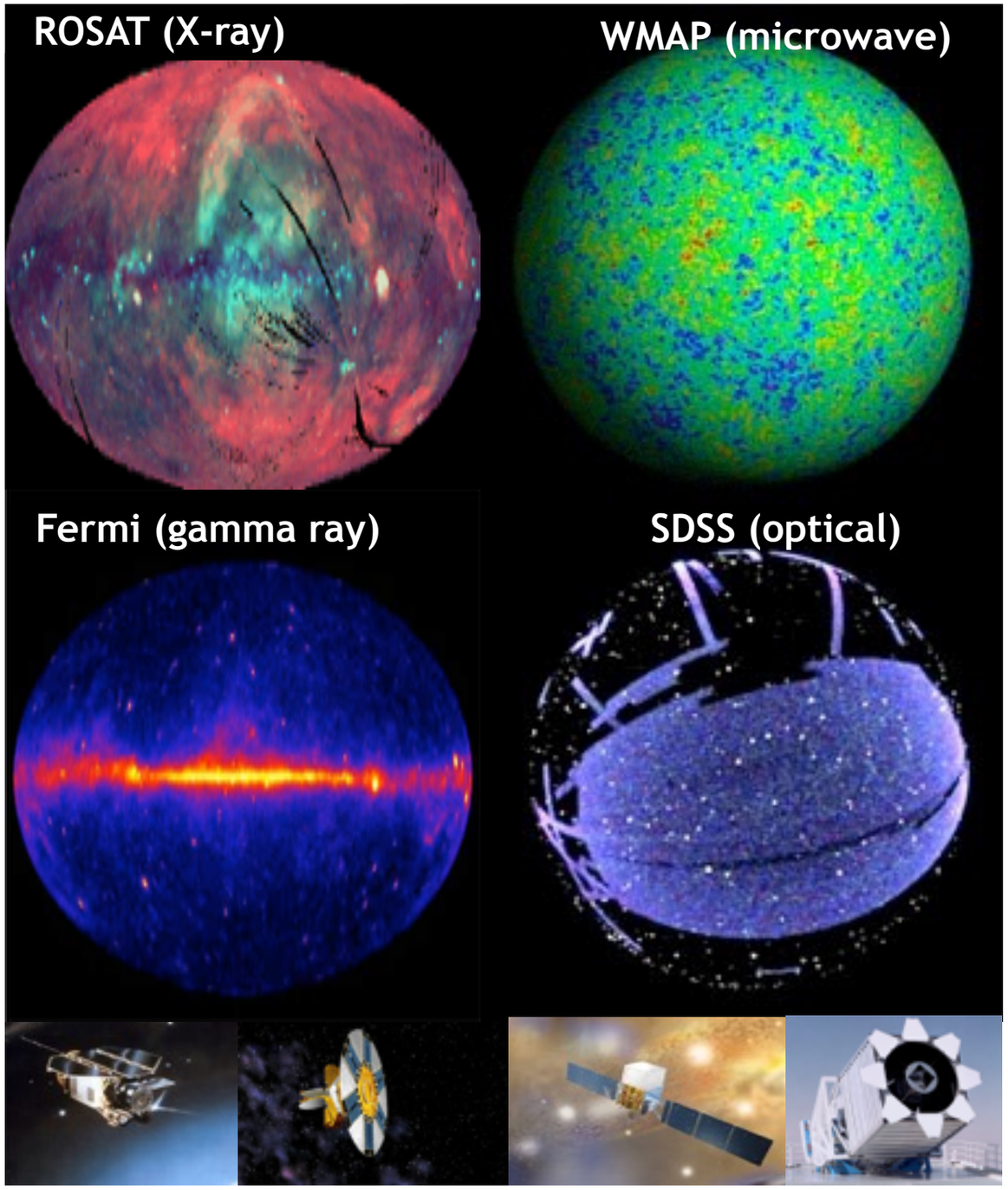
- Physics/A
- mputer Science
- Statistics

Thanks to many collaborators!

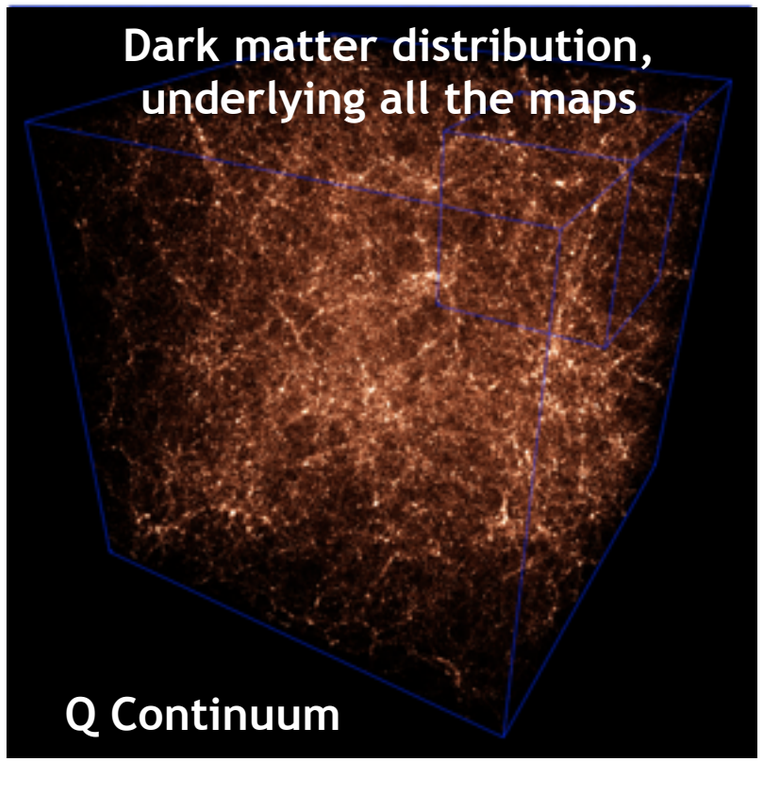
● Physics/Astrophysics (Cosmology, Field Theory, Lattice QCD) ● Computer Science ● Statistics

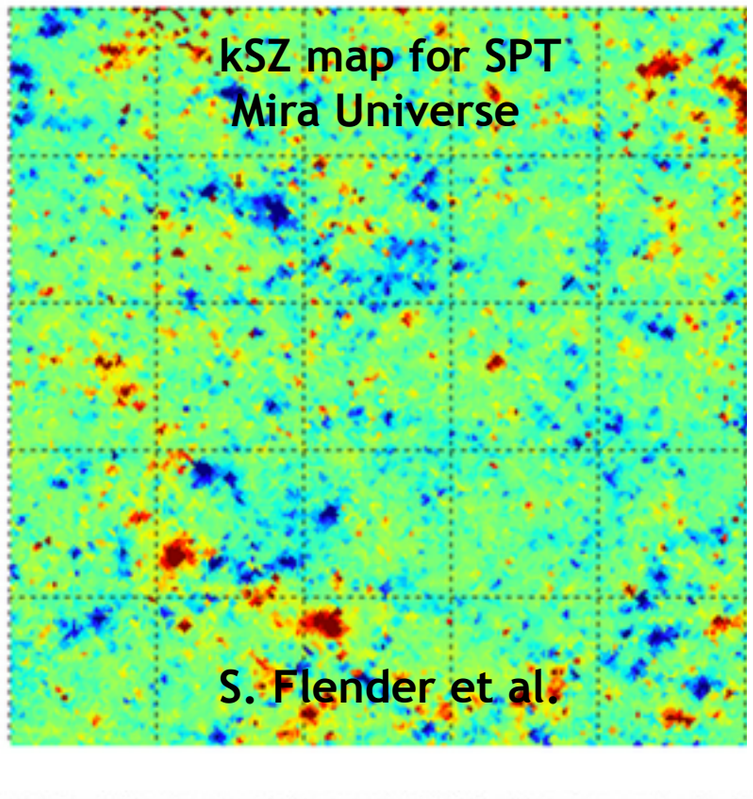
Modern Cosmology and Sky Maps

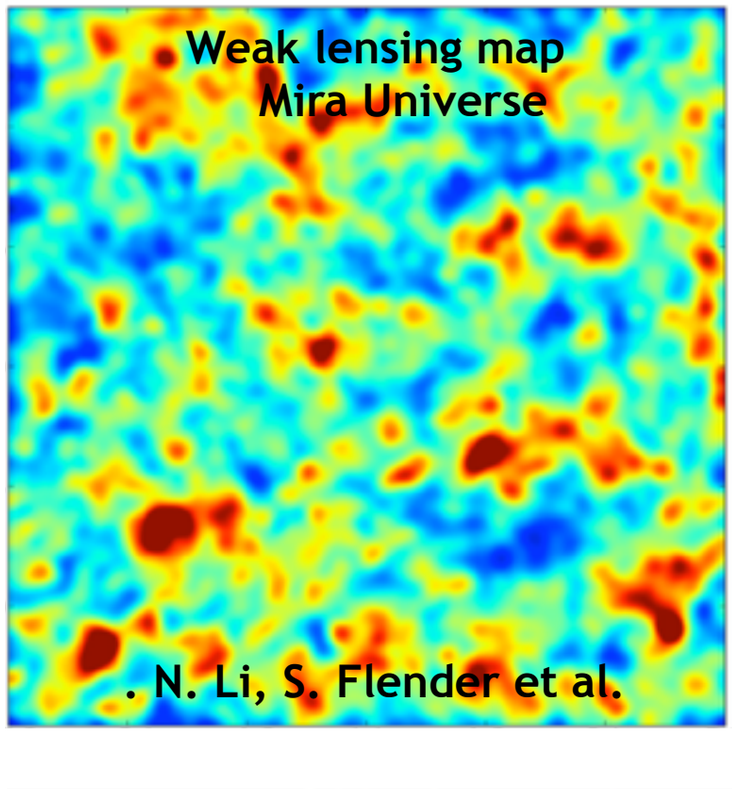
- Modern cosmology is the story of mapping the sky in multiple wavebands
- Maps cover measurements of objects (stars, galaxies) and fields (temperature) and can be very large (SDSS>500 million photometric objects, many billions for planned surveys)
- All precision cosmological analyses constitute a statistical inverse problem: **from sky maps to scientific inference**
- Therefore: *No* cosmology without (large-scale) computing
- **Large-scale simulations used for:**
  - Exploring signatures of new physics
  - Controlled investigation of systematics
  - High-accuracy predictions
  - Testbeds for analysis pipelines

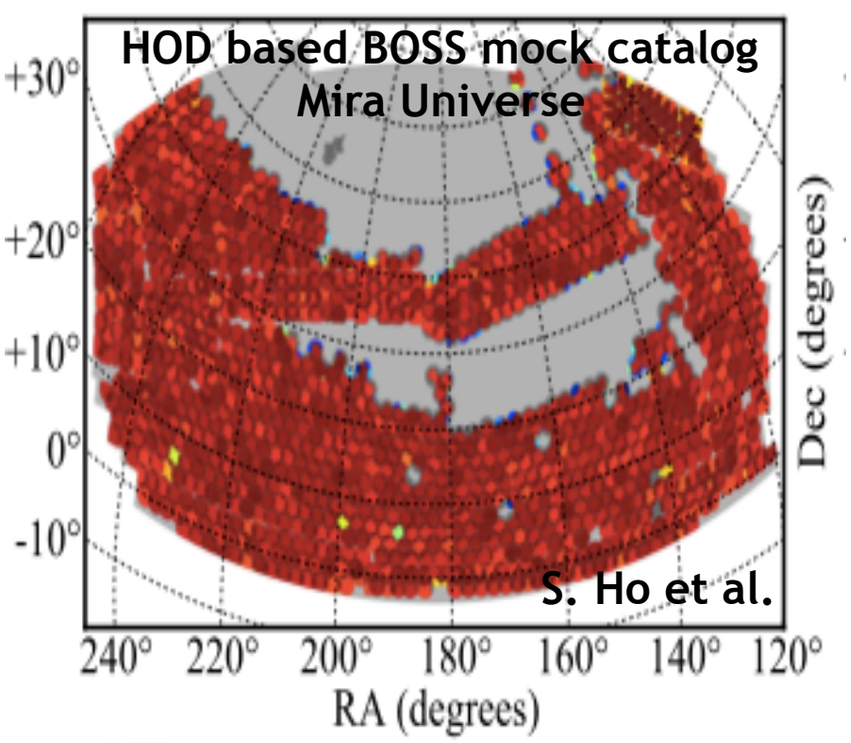


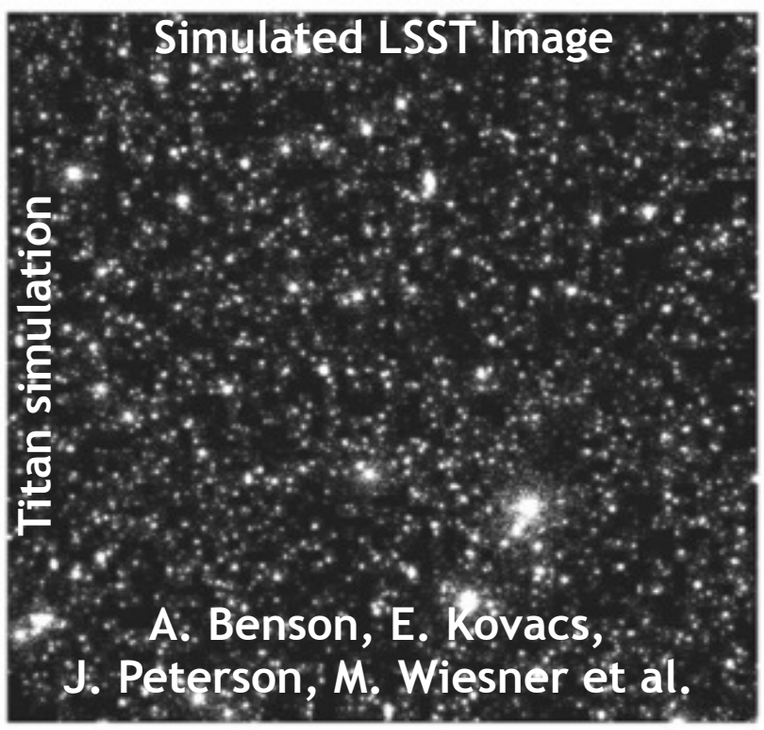
Virtual Skies for Cosmological Surveys

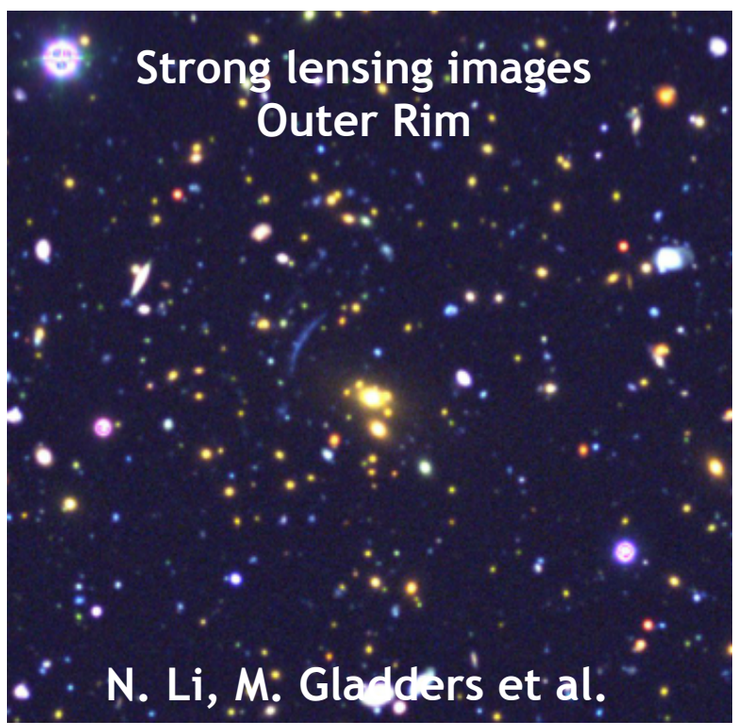












Computing the Universe

- Gravity dominates on large scales, use Monte Carlo sampling of density with tracer particles
- Particles are tracers of the dark matter in the Universe, mass typically at least ~10^9 $M_\odot$

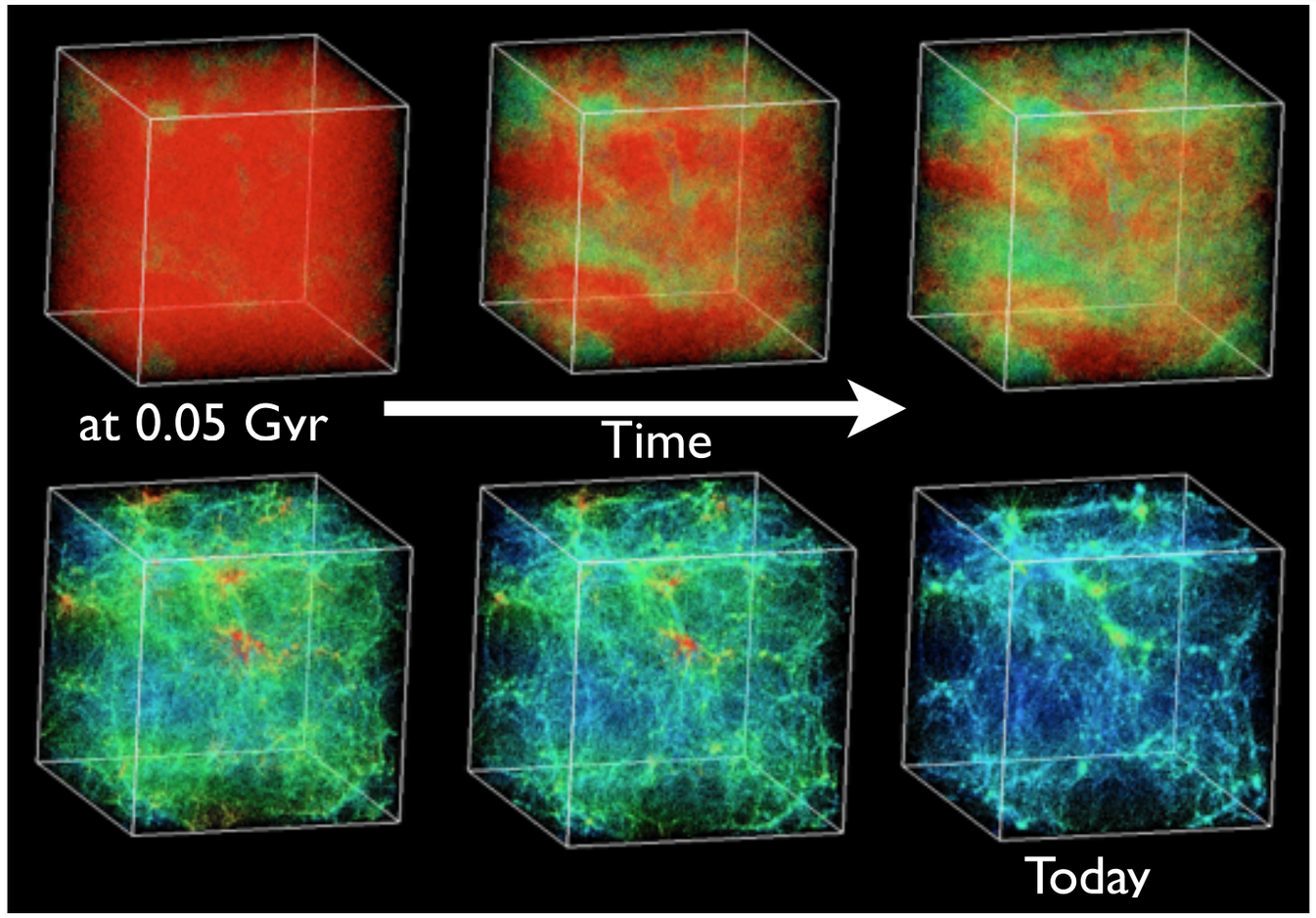
$$m_p \sim V/n_p$$

- Simulate galaxy size objects ($v^2/c^2 << 1$), Newtonian description accurate
- Different approaches: particle-mesh (PM), AMR, tree, combination of PM+tree (PM on large scales, tree on small scales)
- At smaller scales, add gas physics, feedback etc., sub-grid modeling inevitable

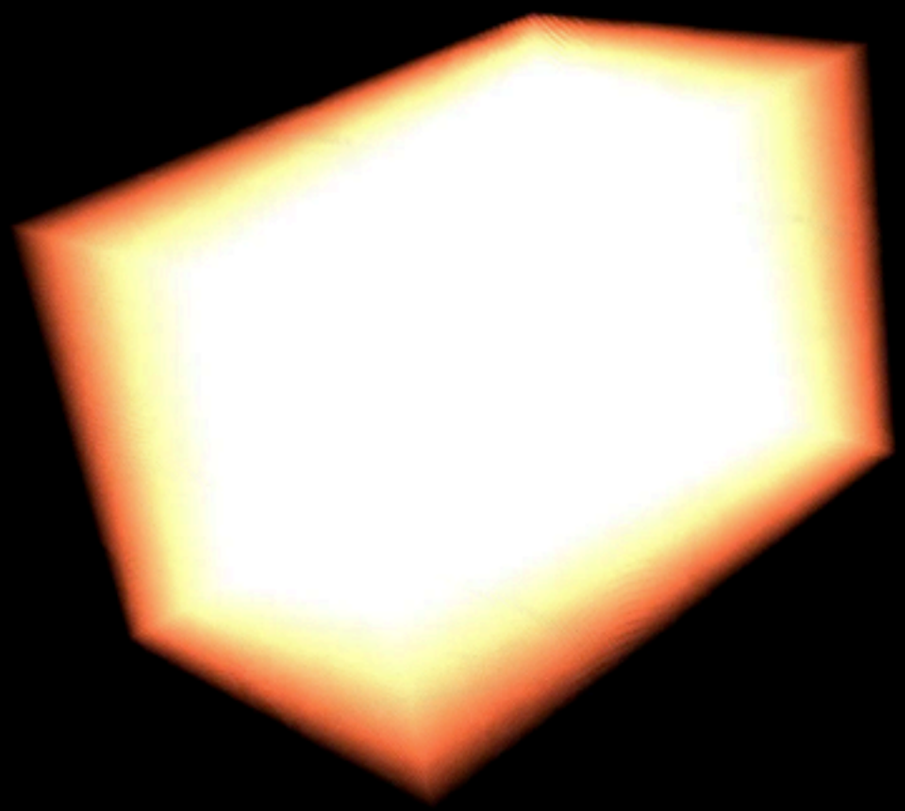
$$\ddot{\mathbf{x}} + 2\frac{\dot{a}}{a}\dot{\mathbf{x}} = -\frac{\nabla\Phi}{a^2}$$ Equation of motion for tracer particles in expanding Universe

$$\frac{\dot{a}}{a} = H = \frac{H_0}{a^{3/2}}\sqrt{\Omega_{tot} + a^3\Omega_\Lambda}$$ CDM + baryons + DE

$$\nabla^2\Phi(\mathbf{x}) = 4\pi G a^2[\rho(\mathbf{x},t) - \rho_b(t)]$$ Poisson equation

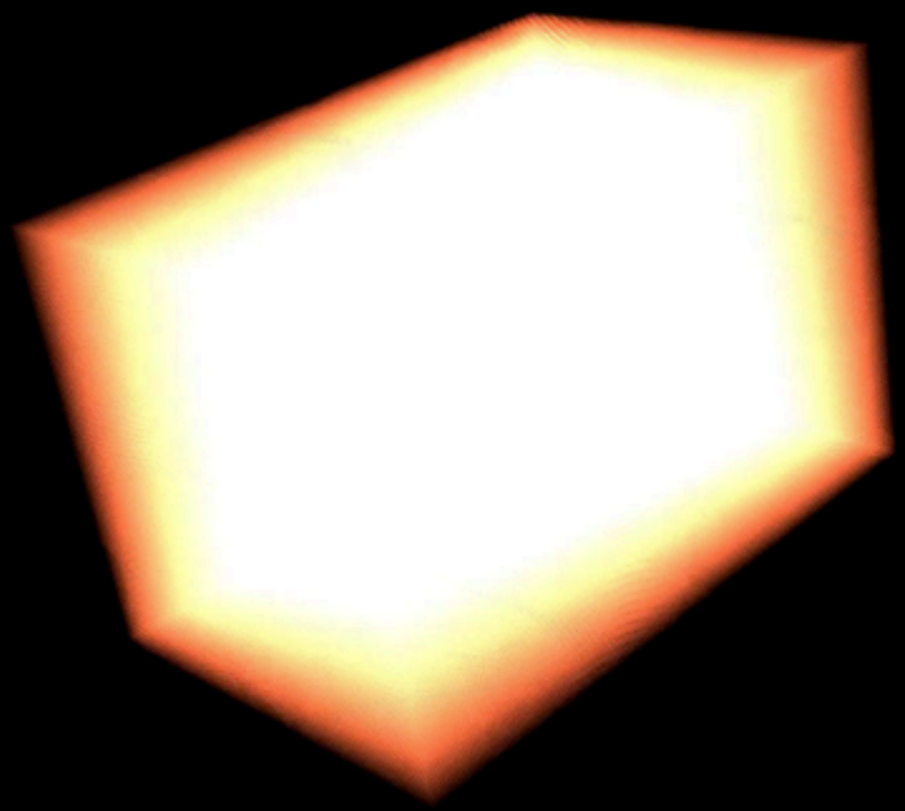


And now for the pretty movie!



(500 Mpc/h)^3 box, 3072^3 particles, saved 750 snapshots, ~800TB of data,
Movie from one rank (out of 1152 ranks), ~(41Mpc/h x 41Mpc/h x 62.5Mpc/h)

And now for the pretty movie!



(500 Mpc/h)^3 box, 3072^3 particles, saved 750 snapshots, ~800TB of data,
Movie from one rank (out of 1152 ranks), ~(41Mpc/h x 41Mpc/h x 62.5Mpc/h)

The HACC Story

The Story Begins ...

☐ Andrew White	Dec 7, 2007	+	What if you had a petaflop/s

- **... with an email:** Los Alamos National Lab offers the opportunity to run open science projects on the fastest supercomputer in the world for the first six months of the machine's existence: **Roadrunner**
- **Roadrunner:** First machine to achieve Petaflop performance via Cell-acceleration, CPU/Cell hybrid architecture (more details later) (equivalent to ~200,000 laptops)
- **The Challenges:**
  - The machine has a "crazy" architecture, requiring major code re-designs and rewrites (we ended up writing a brand **new code**)
  - Roadrunner probably **one of a kind**, code-design needs to be flexible and **portable** to other future architectures
- **Cosmologists are poor -- so we took on the challenge!**
- **Outcome: MC3 (Mesh-based Cosmology Code on the Cell, based on MC2, a PM code written in High-Performance Fortran) which later morphed into HACC, N-body code to simulate large-scale structure formation in the Universe**

The Story Begins ...

☐ Andrew White	Dec 7, 2007	+	What if you had a petaflop/s

- **... with an email:** Los Alamos National Lab offers the opportunity to run open science projects on the fastest supercomputer in the world for the first six months of the machine's existence: **Roadrunner**
- **Roadrunner:** First machine to achieve Petaflop performance via Cell-acceleration, CPU/Cell hybrid architecture (more details later) (equivalent to ~200,000 laptops)
- **The Challenges:**
  - The machine has a "~~crazy~~" architecture, requiring major code re-designs and rewrites (we ended up writing a brand **new code**)
  - Roadrunner probably **one of a kind**, code-design needs to be flexible and **portable** to other future architectures

Andy White: "forward-looking"

- **Cosmologists are poor -- so we took on the challenge!**
- **Outcome: MC3 (Mesh-based Cosmology Code on the Cell, based on MC2, a PM code written in High-Performance Fortran) which later morphed into HACC, N-body code to simulate large-scale structure formation in the Universe**

The Story Continues ... and makes it into “Die Süddeutsche”

Newspaper I read every morning

sueddeutsche.de

Politik | Wirtschaft | Geld | Kultur | Sport | Leben | Karriere | München & Region | Bayern | M

Home > Digital › Supercomputer - Rasend schnell

Supercomputer

Rasend schnell

“He is the fastest calculator in the world:”

Er ist der schnellste Rechner der Welt: der amerikanische Supercomputer "Roadrunner" hat die Petaflop-Grenze geknackt. Sein Job: die Simulation von Atombombenexplosionen.

Twittern 0 | Empfehlen | Senden | +1 0

Ein Rechner der US-Regierung schafft erstmals mehr als eine Billiarde Operationen in der Sekunde (Petaflops) und ist damit nun der schnellste Computer der Welt. Das berichten das US-Energieministerium und der Hersteller IBM am Montag.

Der Computer namens Roadrunner wurde am Los Alamos National Laboratory (LANL) in New Mexico installiert. Er wird zuvorderst für die Forschung an US-Atomwaffen rechnen. Der neu konstruierte Roadrunner ist auf einen Schlag mehr als doppelt so schnell wie der bisherige Spitzenreiter der "Top 500"-Liste der Supercomputer.

Anfangs soll der Roadrunner aber vor allem wissenschaftliche Probleme lösen. Beispielsweise sind Tests von Klimamodellen vorgesehen, doch rechnet das LANL mit Anwendungen in diversen Bereichen, darunter die Kosmologie, die Entwicklung von Antibiotika oder die Astrophysik. Danach wird der Supercomputer laut LANL militärischen Aufgaben zugeteilt und unter Geheimhaltung Explosionen nuklearer Waffen simulieren, um physikalische Modelle zu verbessern und das Vertrauen in das nukleare Arsenal der USA ohne tatsächliche Atomtests zu erhalten.

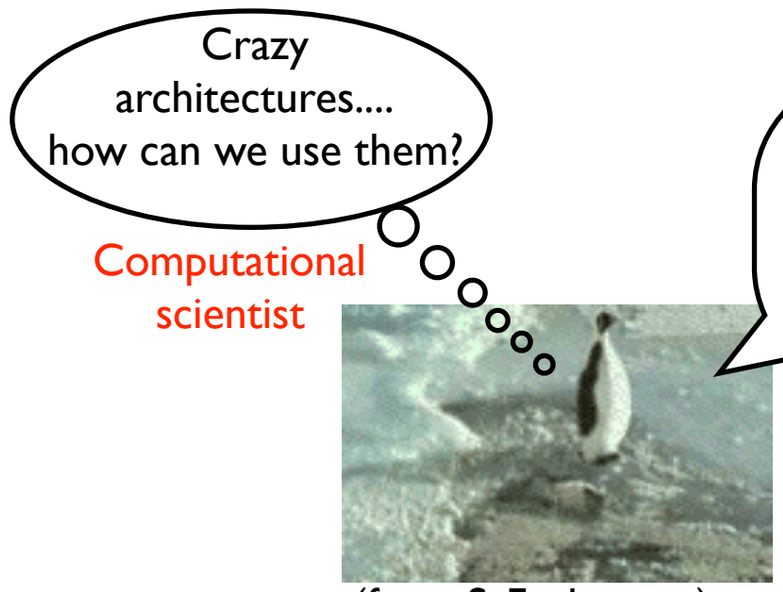
Supercomputer mit Vorbildfunktion “leads by example”

"Für uns und die HPC-Community ist es hoch erfreulich, dass es ein System gibt, das diese Marke geknackt hat", sagt Thomas Lippert, Leiter des Jülich Supercomputing Centre. Dadurch werde dem Supercomputing berechtigte Aufmerksamkeit zuteil.

Technologisch dürfte Roadrunner Vorbildwirkung haben. "Es zeichnet sich ab, dass Hybrid-Technologie auf jeden Fall Zukunft haben", meint Lippert. Damit sind Systeme gemeint, die klassische CPUs mit Beschleunigern wie beispielsweise den Cell-Chips oder Grafikprozessoren kombinieren.

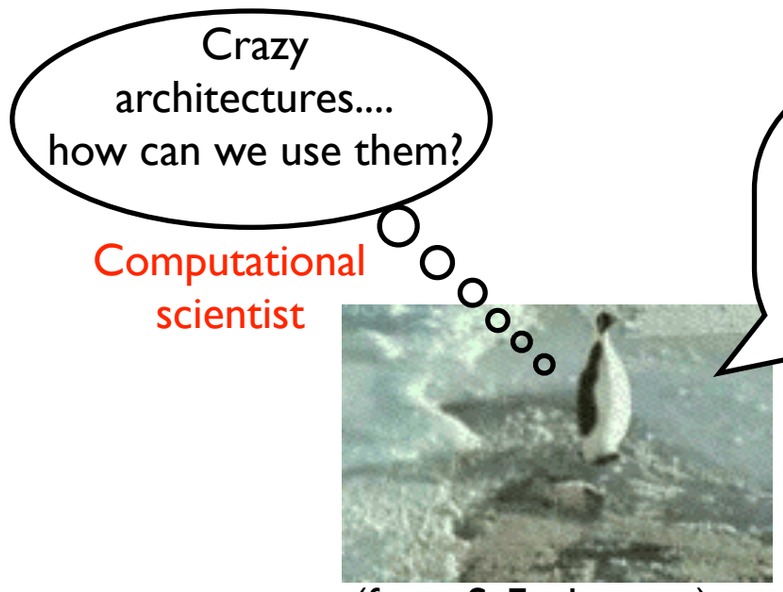
“technology of the future”

So we started thinking --



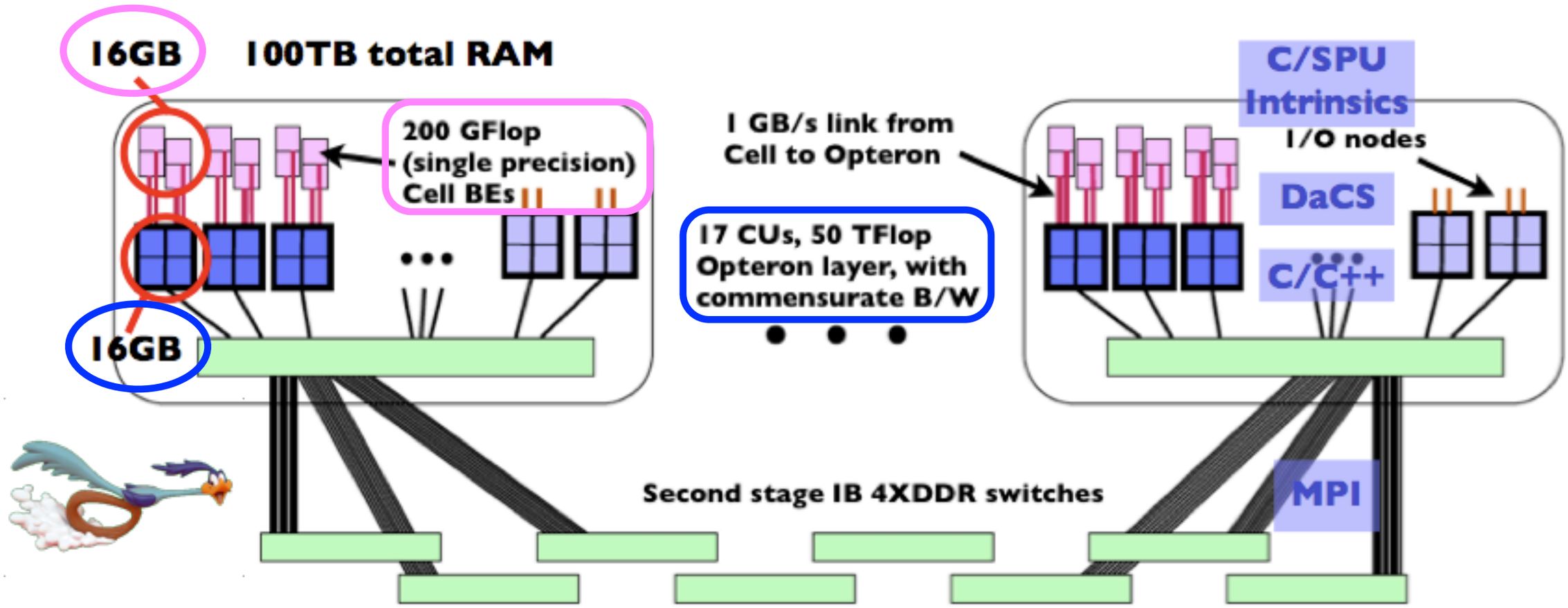
(from S. Furlanetto)

So we started thinking --



(from S. Furlanetto)

The Roadrunner Architecture



- **Opterons** have little compute (5% of total compute) but half the memory and balanced communication, for N-body codes, memory is limiting factor, so want to make best use of CPU layer
- **Cells dominate** the compute but communication is poor, 50-100 times out of balance (also true for CPU/GPU hybrid systems)
- **Multi-layer programming model:** C/C++/MPI (Message Passing Interface) for Opterons, C/Cell-intrinsics for Cells (OpenCL or Cuda for GPUs)

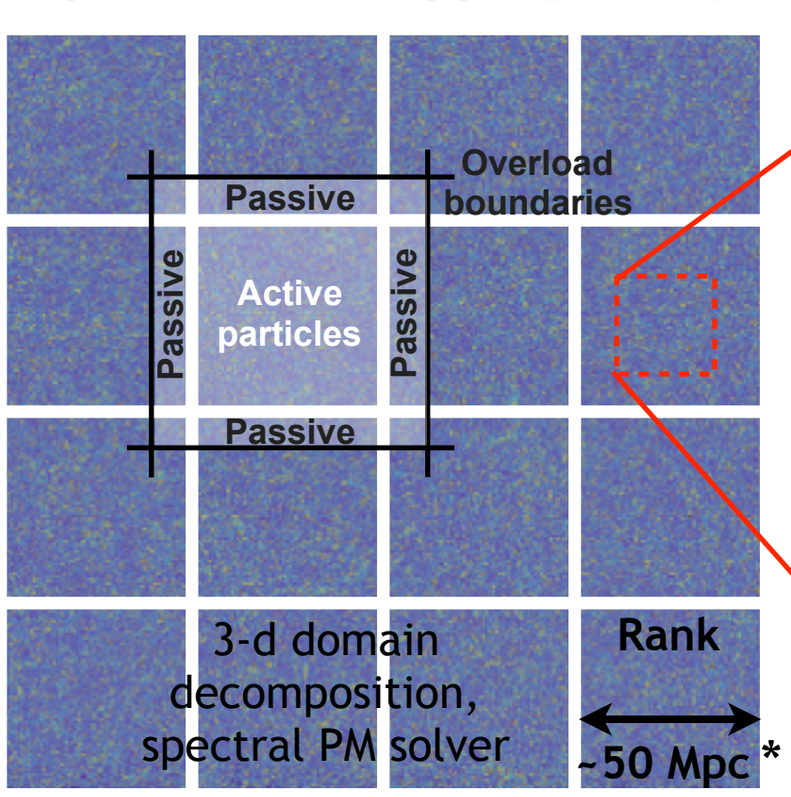
Design Challenges and Solutions

- **Challenges (summarized from last slide):**
  - Opterons have half of the machine's memory, balanced communication, but not much compute, standard programming paradigm, C/C++/MPI
  - Cells have other half of machine's memory, slow communication, **lots** of compute (95% of machine's compute power), new language required
- **Design desiderata:**
  - Distribute memory requirements on both parts of the machine (different on GPUs!)
  - Give the Cell lots of (communication limited) work to do, make sure that Cell part is easy to code and later on easy to replace by different programming paradigm

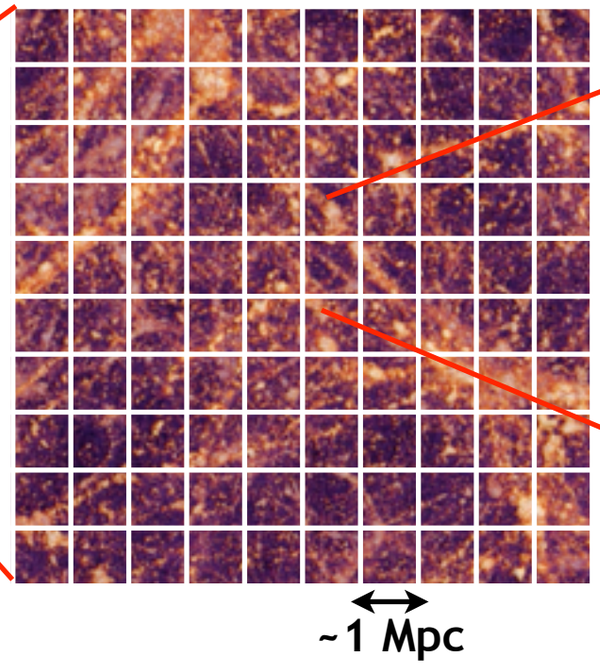
HACC in a Nutshell

S. Habib et al. 2016, New Astronomy

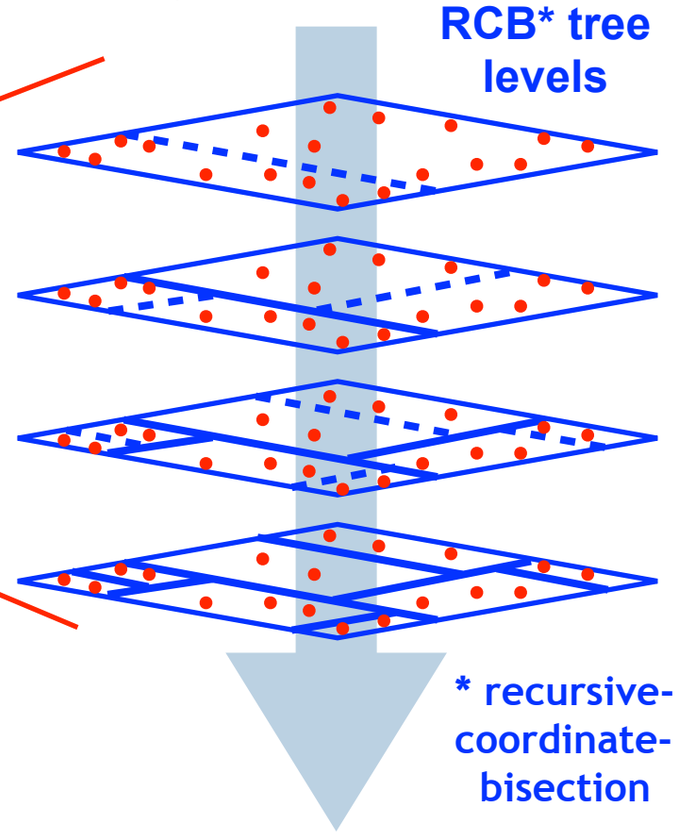
- **Long-range/short range force splitting:**
  - **Long-range:** Particle-Mesh solver, C/C++/MPI, **unchanged for different architectures,** FFT performance dictates scaling (custom pencil decomposed FFT)
  - **Short-range: Depending on node architecture** switch between tree and particle-particle algorithm; tree needs "thinking" (building, walking) but computationally less demanding (BG/Q, X86, KNL), PP easier but computationally more expensive (GPU)
- **Overload concept to allow for easy swap of short-range solver and minimization of communication (reassignment of passive/active in regular intervals)**
- **Adaptive time stepping, analysis on the fly, mixed precision, custom I/O, ...**



HACC Top Layer



HACC 'Nodal' Layer



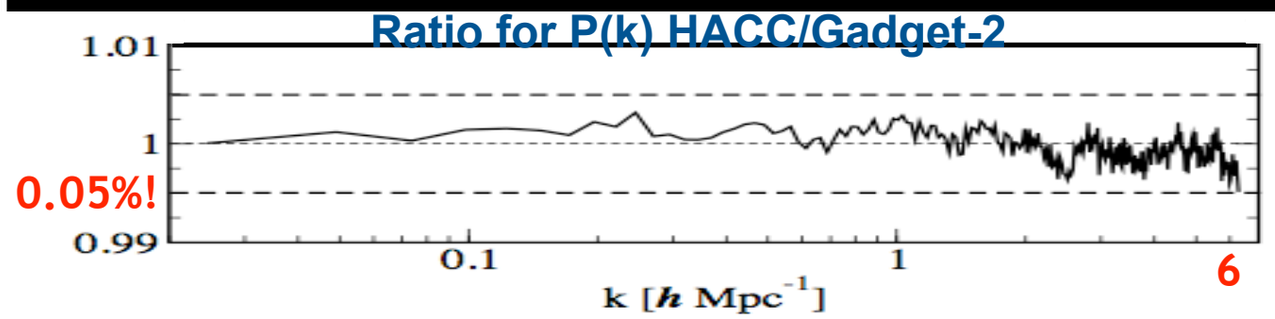
* Mpc~3M light years, distance between 2 galaxies

MC³

Gadget-2

Average halo velocity
400 800 1200 1600
27.7964 1875.3

Snapshot from Code Comparison simulation, ~25 Mpc region; halos with > 200 particles, b=0.15
Differences in runs: P³M vs. TPM, force kernels, time stepper: MC³: a; Gadget-2: log(a)
Power spectra agree at sub-percent level



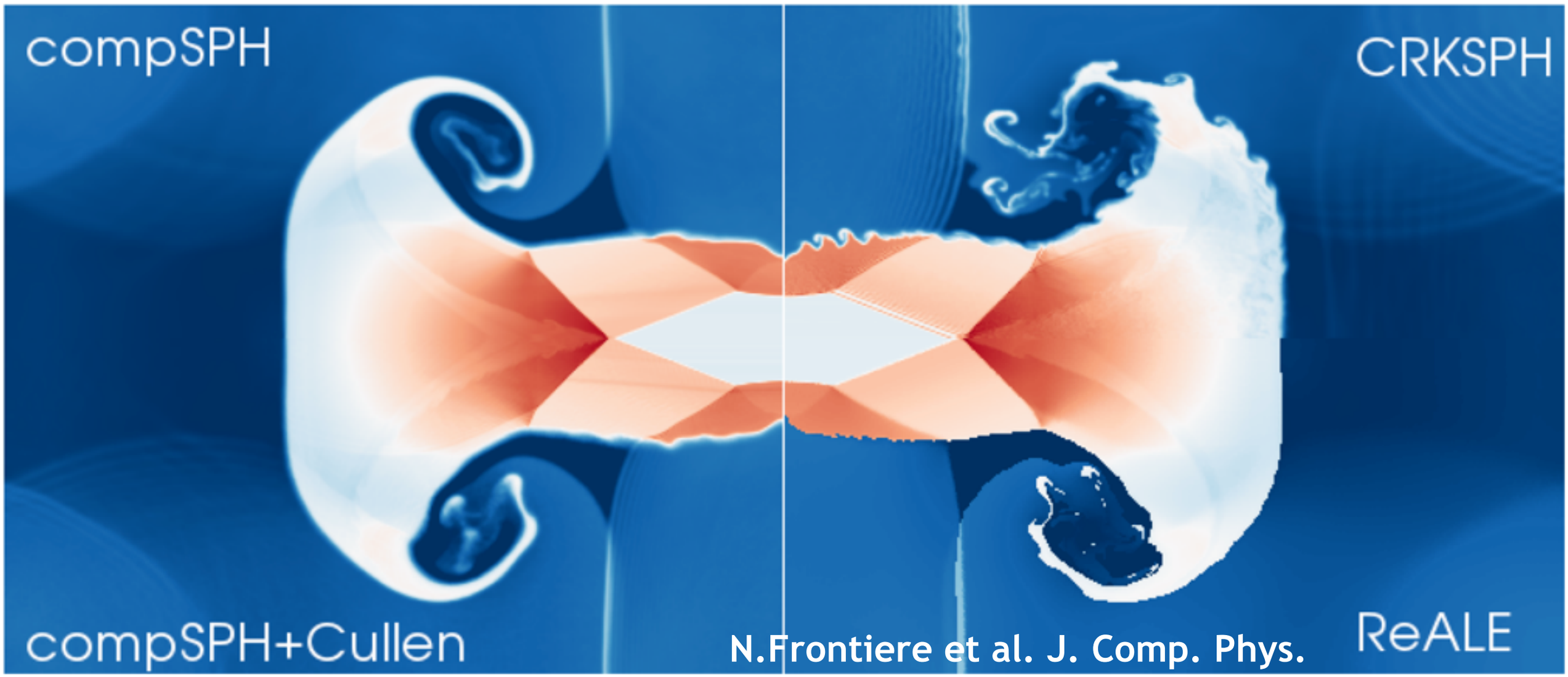
HACC on Different Architectures

- **IBM Blue Gene (BG) systems**
  - Machines: BG/Q Mira at Argonne: 10 PFlops, arrived in 2012, 750,000 cores, 16GB per node; Sequoia at Livermore: twice as large
  - New challenges: BG/Q systems have many cores but no accelerators; slab-decomposed FFT does not scale well on very large number of cores
  - Solutions: Particle-particle interaction now replaced by tree, OpenMP node parallel, pencil decomposed FFT, adaptive time stepping
  - Performance: Achieved 13.94PFlops on Sequoia, 90% parallel efficiency on 1,572,864 core, 3.6 trillion particle benchmark runs (1.1 trillion particle science run on Mira)
- **GPU systems**
  - Machine: Titan at Oak Ridge: 17.6 PFlops (double precision), arrived in 2013, 18.688 AMD Opterons + Nvidia Tesla K20X GPU (= new challenge)
  - Solutions: Force kernel implementation in OpenCL, later CUDA, back to particle-particle interaction (have tree as well), new load-balancing scheme
  - Performance: 20.54 Pflops peak in test run on ~75% of machine

HACC on Different Architectures and Future

- **New systems, just arrived**
  - HACC runs on KNL: Theta at Argonne, 10PFlops machine, ~3400 nodes; Cori at NERSC, ~9000 nodes (were able to take advantage of BG/Q work)
  - HACC runs on SummitDev: Oak Ridge test bed machine, 54 IBM S822LC nodes each with 2 IBM POWER8 CPUs and 4 NVIDIA Tesla P100 GPUs (we were able to take advantage of Titan work)
- **Future for HACC**
  - HACC on Summit: IBM9 + NVIDIA Volta GPUs, to arrive in 2017, 200+ PFlops, 3400 nodes, Early Science Project, 10 MW power consumption (Titan: 9MW)
  - HACC on Aurora: KNH system, to arrive in 2019, 200+ PFlops, ~50,000 nodes, Early Science Project, 13 MW power consumption (Mira: 4.8MW)
  - ECP: HACC is part of the DOE-sponsored Exascale Project (ECP), prepare applications and software for exa-scale machines
    - **CRK-HACC! HACC now with baryonic physics, feedback, …** (conservative reproducing kernel SPH, improved first order method that conserves mass, momentum, and energy, combined with a new, less diffuse artificial viscosity model)

HACC on Different Architectures and Future



- **CRK-HACC! HACC now with baryonic physics, feedback, …** (conservative reproducing kernel SPH, improved first order method that conserves mass, momentum, and energy, combined with a new, less diffuse artificial viscosity model)

MC2


Aurora
SUMMIT
Edison
MIRA
5
HACC

HACC Science

HACC Science

Mock galaxy catalogues: one simulation every year with 10T particles. Galaxy population on the light cone with HOD/AM/SAM techniques with lensing maps.

Resources: 2 million node-hours (with GPU)

Emulators: 50 such simulations (one per cosmological parameter set)

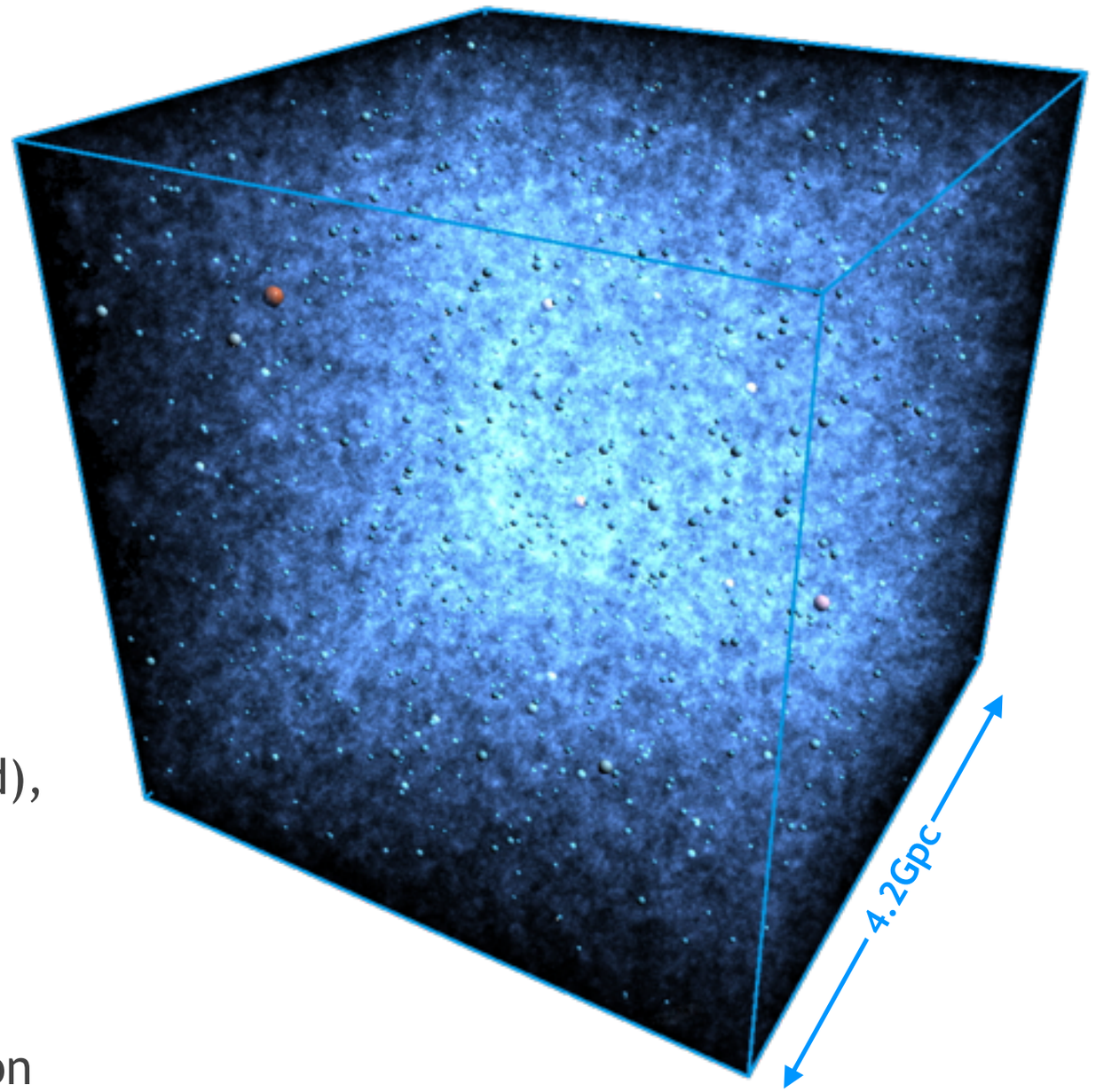
Coyote: Heitmann et al. 2014, Mira-Titan; Heitmann et al. 2016

Covariance matrices: 3000 simulations with 8B particles every year.

Resources: 2 million node-hours (with GPU)

HACC Extreme Scale Simulations

- **The Outer Rim Simulation**
  - HACC simulation with > 1 trillion particles, 100 snapshots, 4.5PB of data; 4.2Gpc volume with $2x10^9 M_{sun}$ mass resolution
  - Carried out on 2/3 of Mira (BGQ), the 6th fastest supercomputer in the world
- **The Q Continuum Simulation**
  - More than 0.5 trillion particles, 2.5PB of data; 1.3Gpc volume, $10^8 M_{sun}$ mass resolution
  - Carried out on 90% of Titan (GPU enhanced), 3rd fastest supercomputer in the world
- **The Mira-Titan Universe Suite**
  - World-wide unique suite of different models, each simulation: 2.1 Gpc, 33 billion particles (72 finished, 38 more in prep)



Large halos in the Outer Rim simulation, volume large enough to hold the DESI survey

K. Heitmann, et al. ApJS, 2015; K. Heitmann et al. ApJ, 2016

http://chicagotonight.wttw.com/2015/12/15/argonne-national-lab-simulation-tracks-evolution-universe

HACC Extreme Scale Simulations

- **The Outer Rim Simulation**
  - HACC simulation with > 1 trillion particles, 100 snapshots, 4.5PB of data; 4.2Gpc volume with $2x10^9 M_{sun}$ mass resolution
  - Carried out on 2/3 of Mira (BGQ), the 6th fastest supercomputer in the world
- **The Q Continuum Simulation**
  - More than 0.5 trillion particles, 2.5PB of data; 1.3Gpc volume, $10^8 M_{sun}$ mass resolution
  - Carried out on 90% of Titan (GPU enhanced), 3rd fastest supercomputer in the world
- **The Mira-Titan Universe Suite**
  - World-wide unique suite of different models, each simulation: 2.1 Gpc, 33 billion particles (72 finished, 38 more in prep)

The Cosmic Emu

K. Heitmann, et al. ApJS, 2015; K. Heitmann et al. ApJ, 2016

http://chicagotonight.wttw.com/2015/12/15/argonne-national-lab-simulation-tracks-evolution-universe

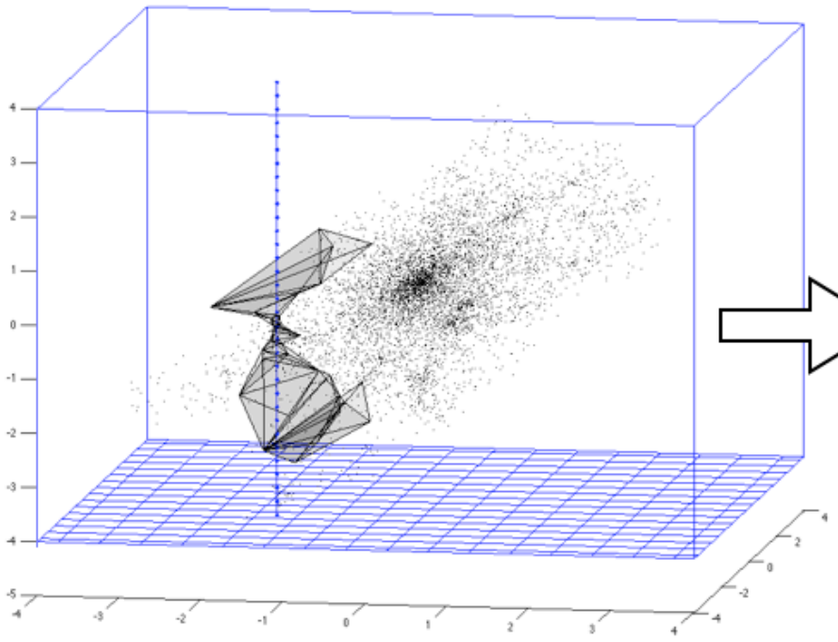
Science with HACC: Strong Lensing

SDSS Images

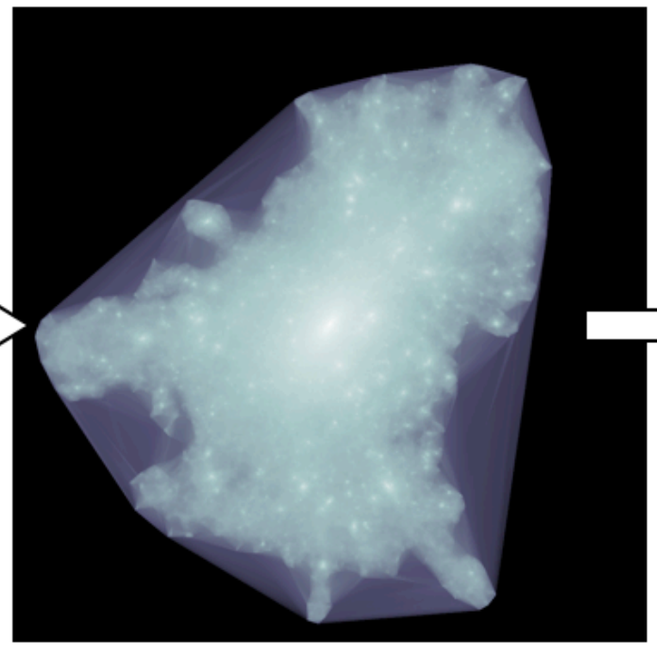
Images courtesy of M. Gladders

Cosmology with HACC: Strong Lensing

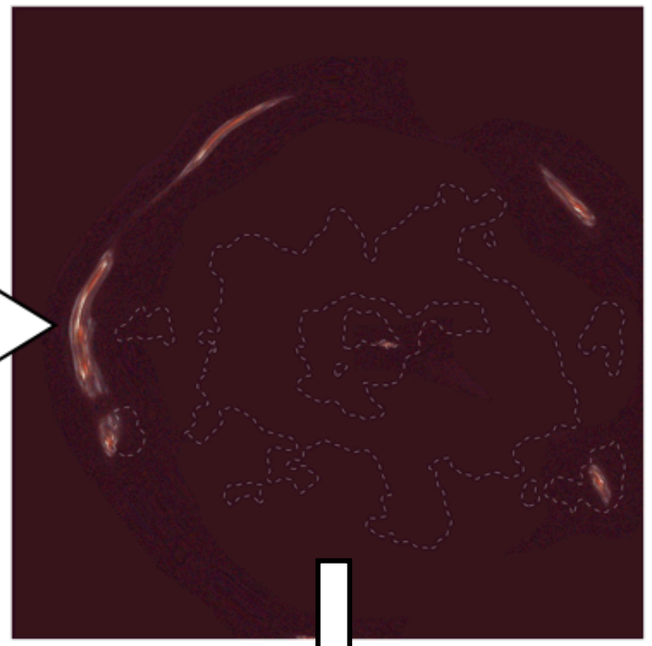
Halo particle data

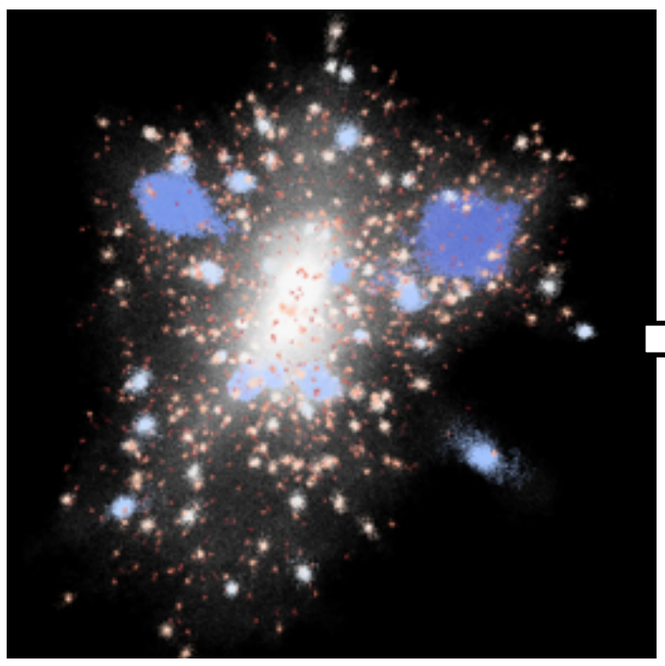


Density

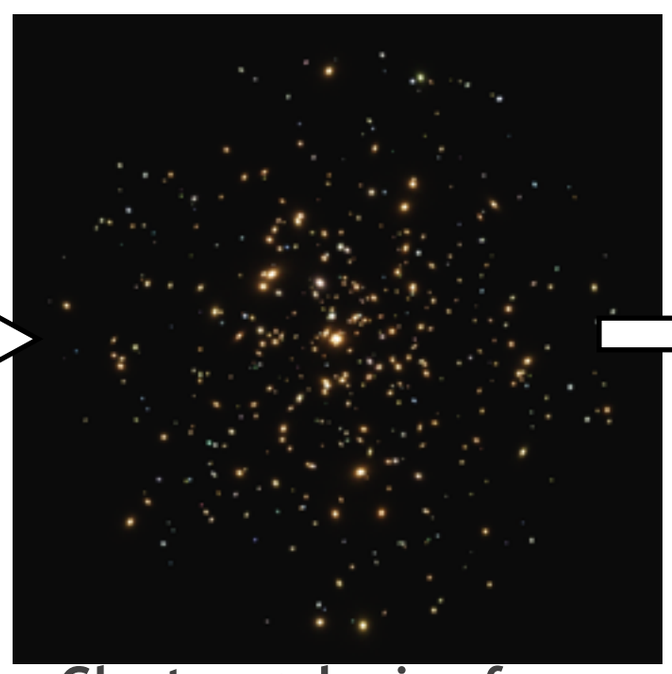


Strong lensing arc

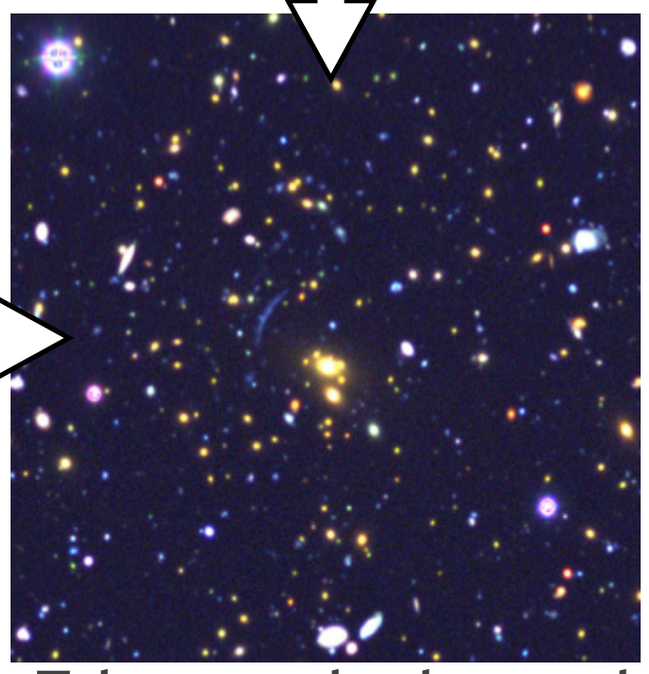




Subhalo particle data



Cluster galaxies from Galsim, optical clusters

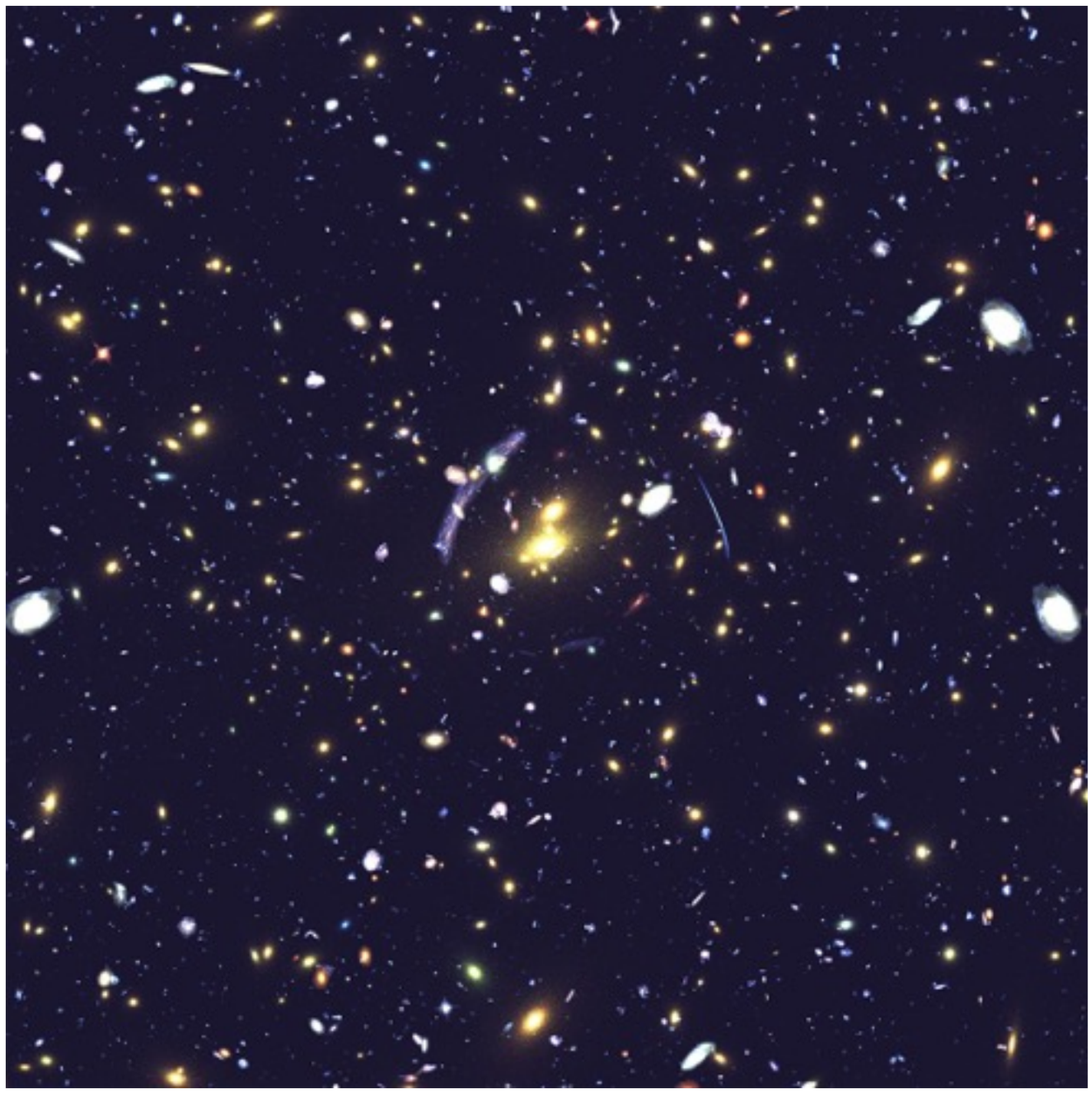


Telescope, background galaxies, cluster, arc ...

PICS: Generating Strong Lensing “Observations”

PICS: Pipeline for Images of Cosmological Strong lensing

Simulated



Real

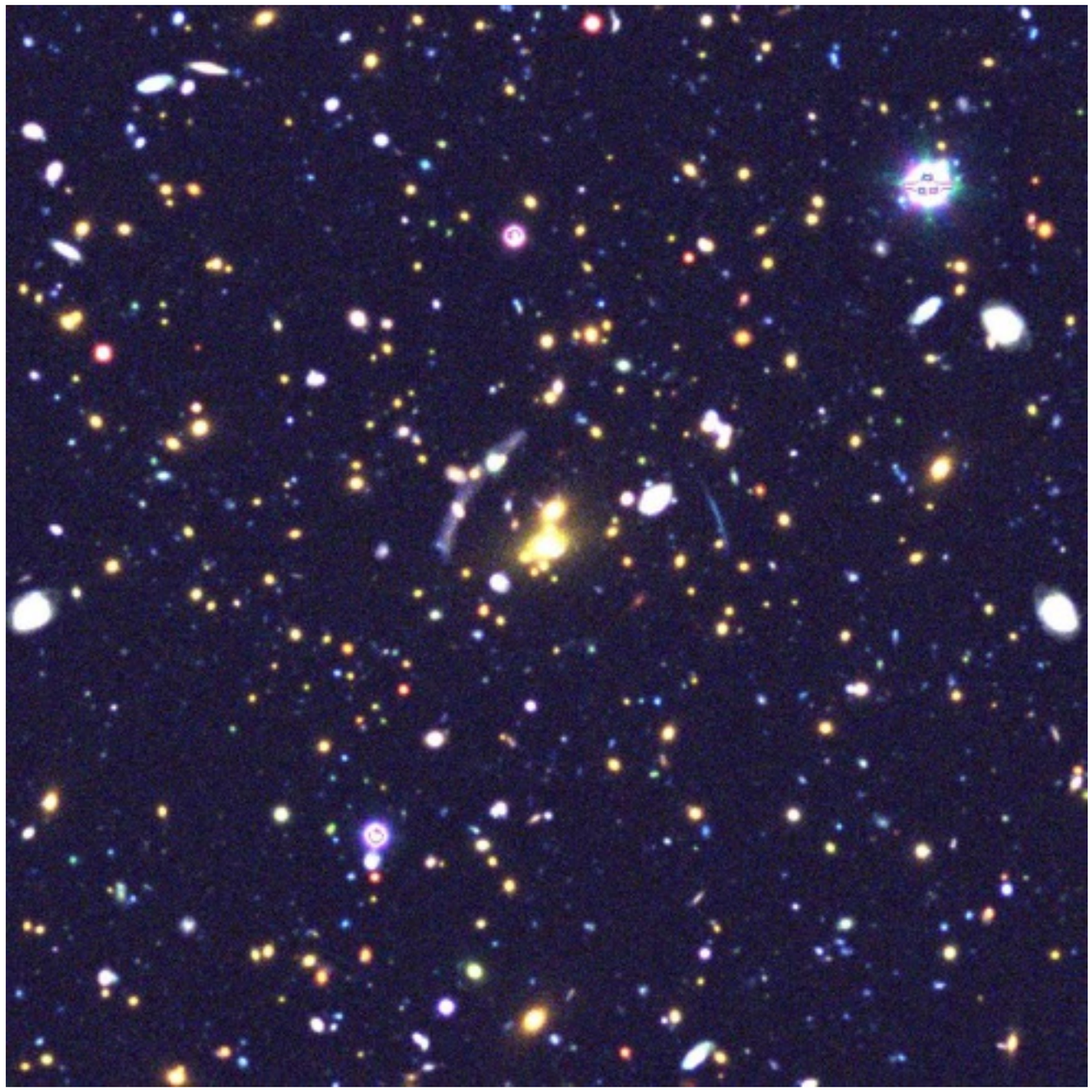
Simulated strong lens image to match SPT cluster observations taken with the MegaCAM camera on Magellan, in collaboration L. Bleem, M. Florian, S. Habib, M. Gladders, N. Li, S. Rangel

N. Li et al., ApJ, arXiv:1511.03673

PICS: Generating Strong Lensing “Observations”

PICS: Pipeline for Images of Cosmological Strong lensing

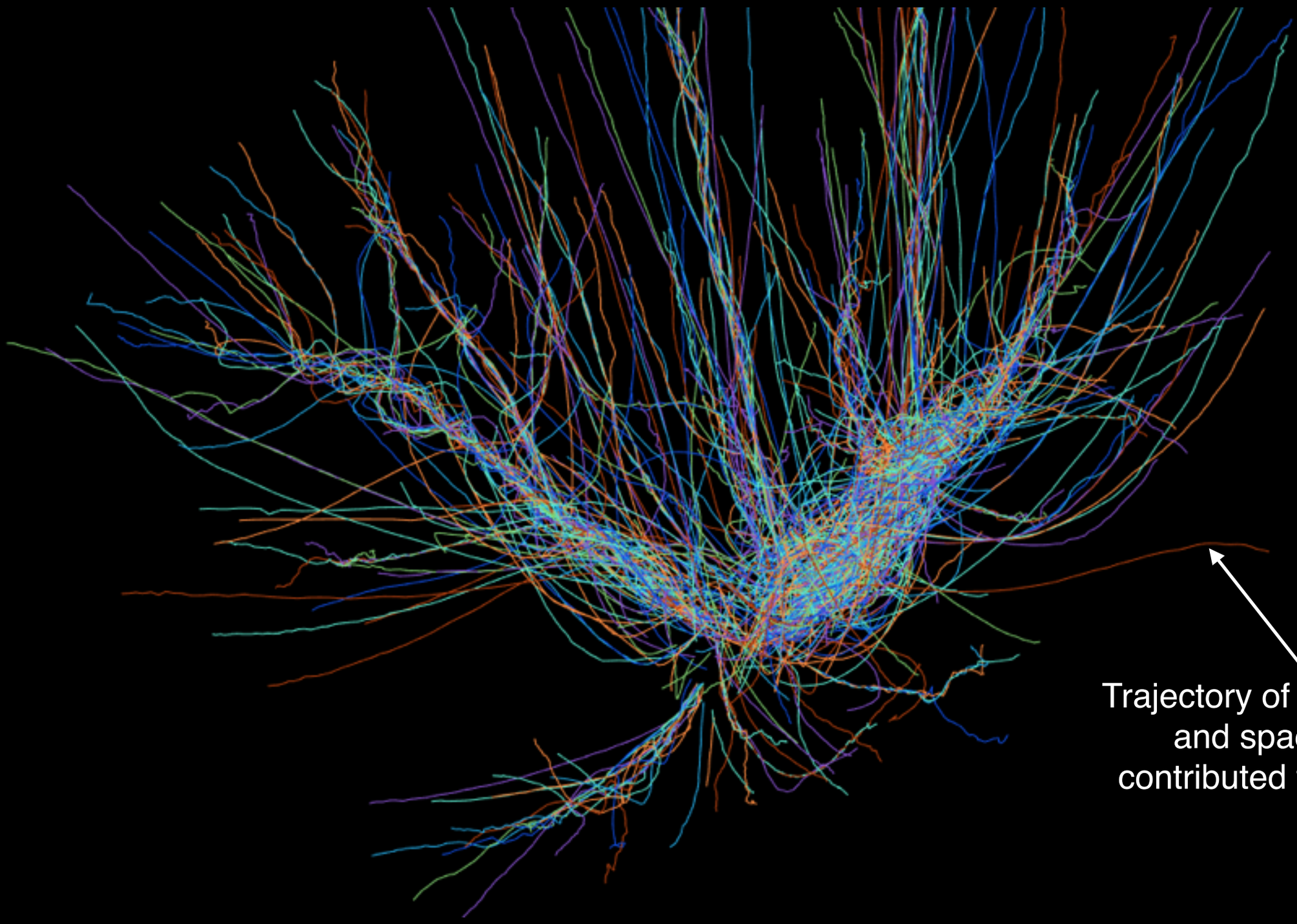
Simulated | **Real**



Simulated strong lens image to match SPT cluster observations taken with the MegaCAM camera on Magellan, in collaboration L. Bleem, M. Florian, S. Habib, M. Gladders, N. Li, S. Rangel

N. Li et al., ApJ, arXiv:1511.03673

Formation History of One Halo

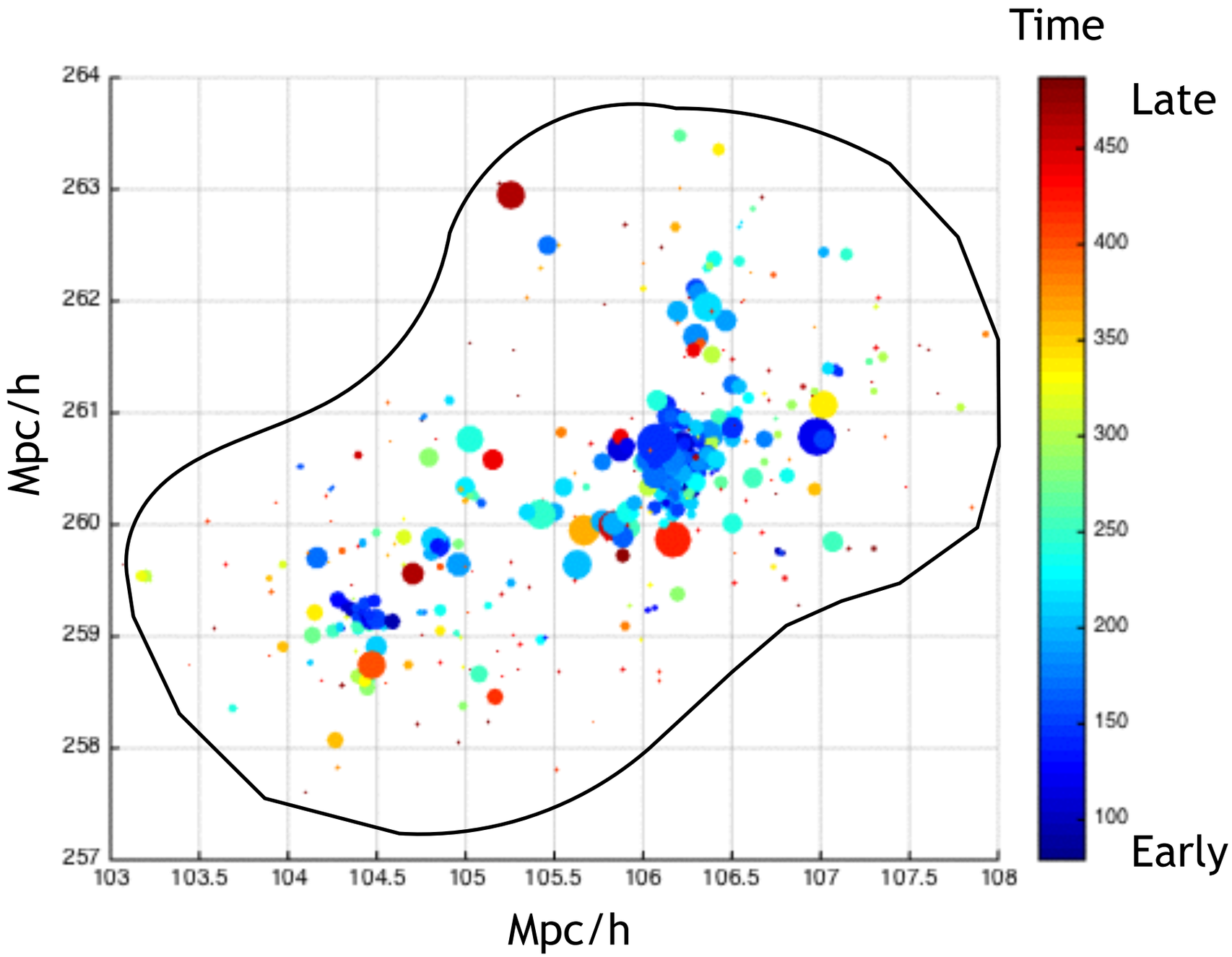


Trajectory of a halo over time and space halo that contributed to the final halo

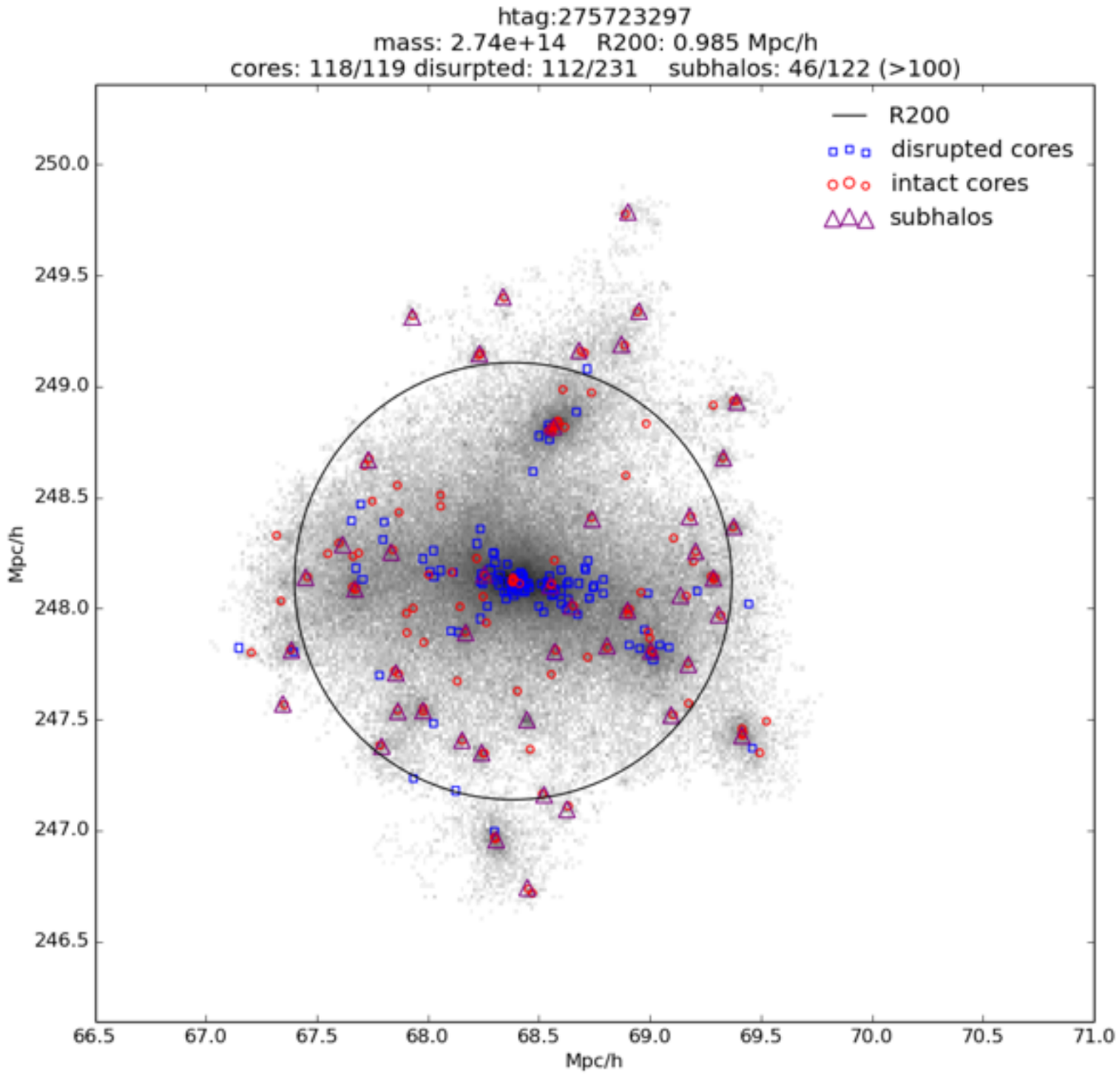
Subhalos give static view, but we would like to take into account full history of halo and continue follow it after it merged into bigger halos, “core” tracking

Cores of One Halo Tracked to z=0

- Identify halos starting at high redshift
- Once halo above certain mass is found (100 particles here) we identify 20 particles closest to the center
- These particles are now tracked until z=0
- Merger trees include information about cores to enable relating cores back to main halo



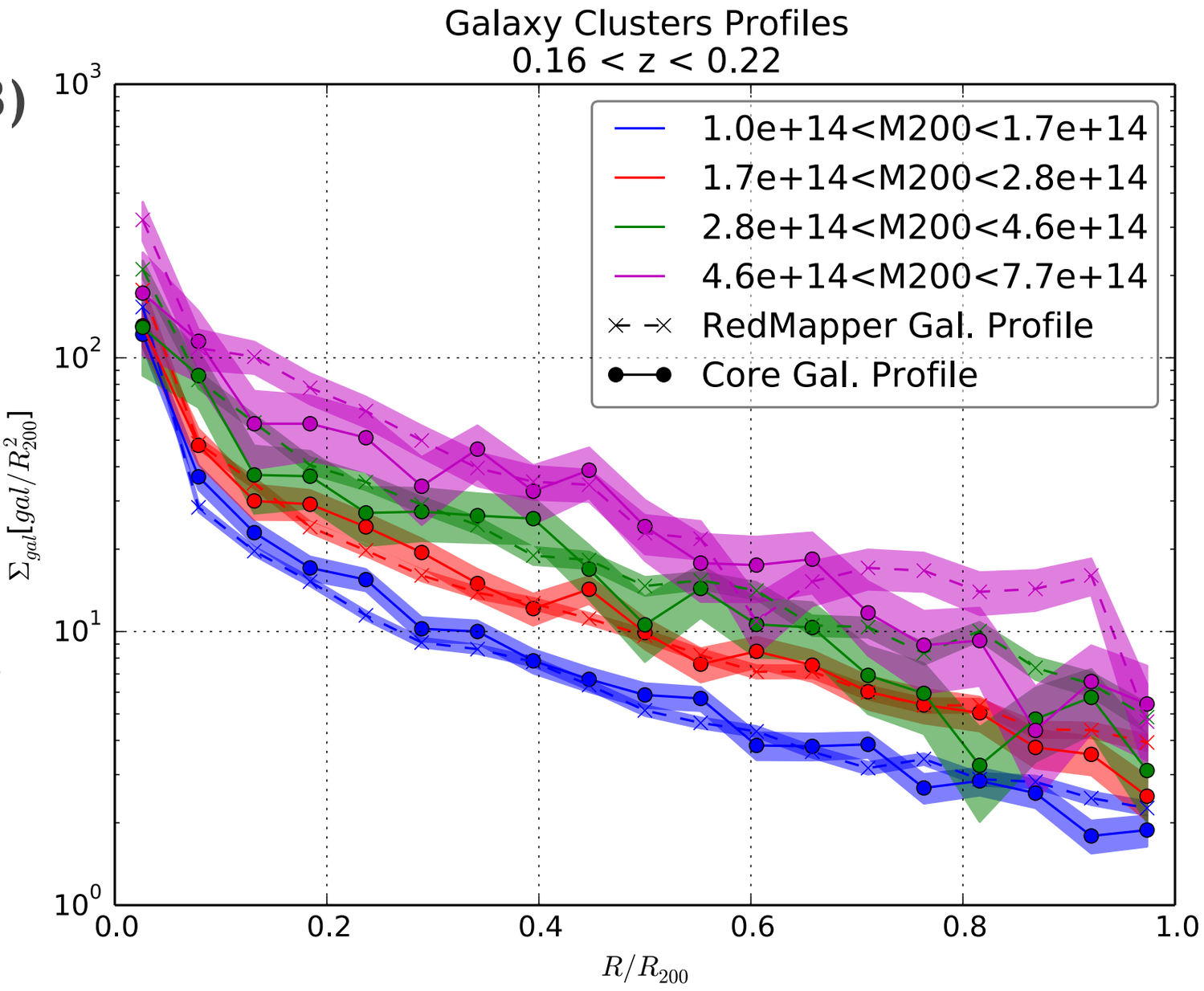
Cores vs. Subhalos



D. Korytov et al., in prep.

Galaxy Cluster Profiles

- Optical galaxy cluster finder & Catalog (Rykoff et al. 2013)
- Construct galaxy density within R_{200}
  - $0.4L_*$ or brighter
  - $10^{14}\ M_{sun} < M_{200} < 10^{15}\ M_{sun}$
  - $0.1 < z < 0.35$
  - Background subtracted
- 3 parameters for core fitting
  - Minimum halo size for tracking
  - Core disruption
  - Core merging



Data from Rykoff et al. 2013

D. Korytov et al., in prep.

Summary and Outlook

- Exciting times for cosmology!
- Analysis of current and future cosmological data requires large-scale precision simulations, exploring different cosmological models
- Current and future supercomputers allow for such simulations but are challenging to use at the same time!
- HACC is a large scale computational tool to address both challenges
  - Designed for high performance and ease of portability
  - Runs on all currently available supercomputing architectures
  - In addition to the code itself, extensive work on analysis tools (in-situ as well as post-processing) to create data as close to observations as possible
- Future surveys will explore smaller and smaller scales, accurate modeling/simulating of astrophysics (for cosmology: systematics) will become crucial if we want to extract cosmological information from these scales at high accuracy
- 2 Postdoc positions available! If you want to join the HACC team, please contact me —