スーパーコンピュータ「京」の中の宇宙

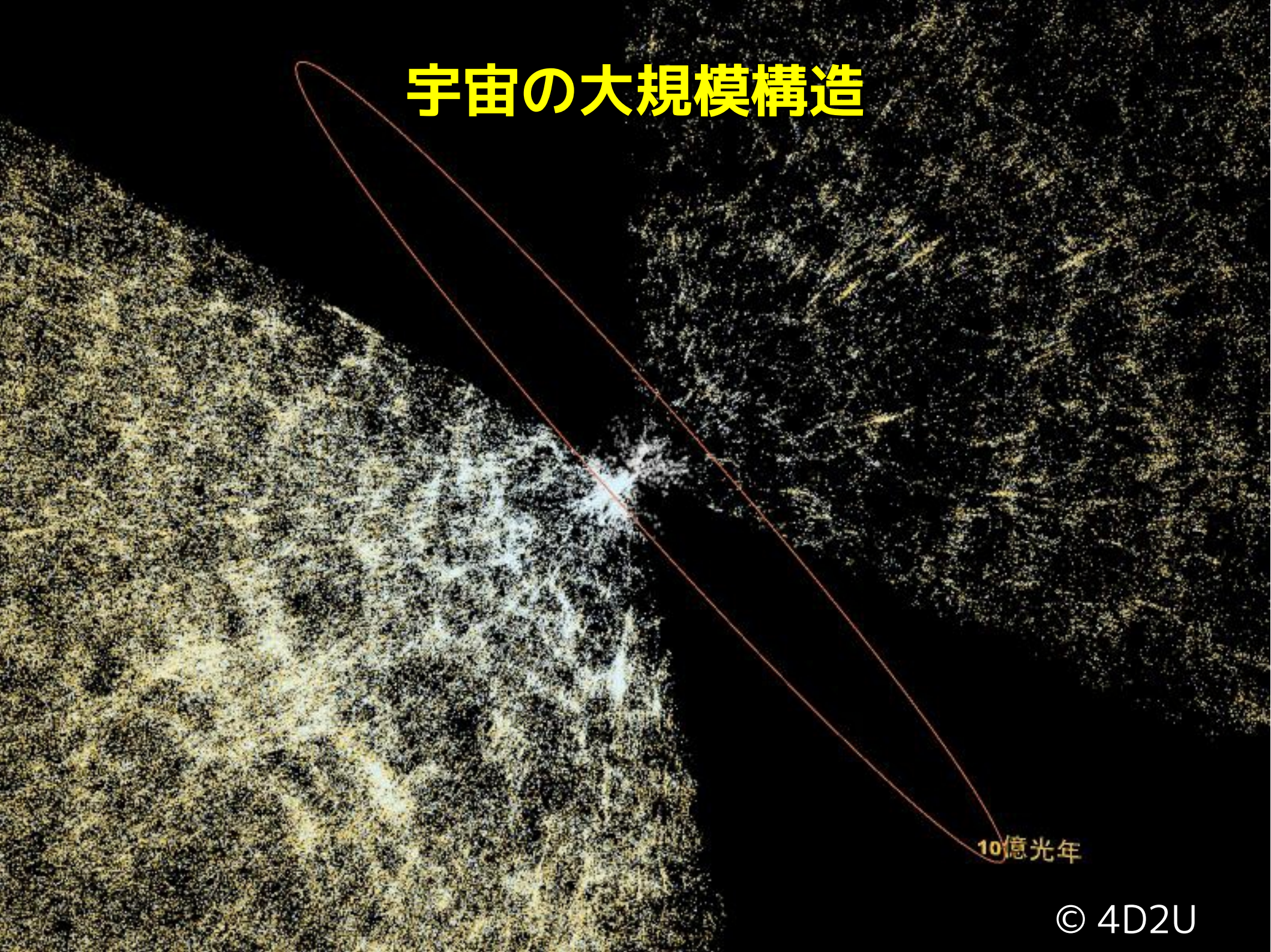
戦略分野5「物質と宇宙の起源と構造」
筑波大学計算科学研究センター (AICS内神戸分室)

石山 智明

さまざまな銀河


銀河団
Abel 1689
© NASA

宇宙の大規模構造

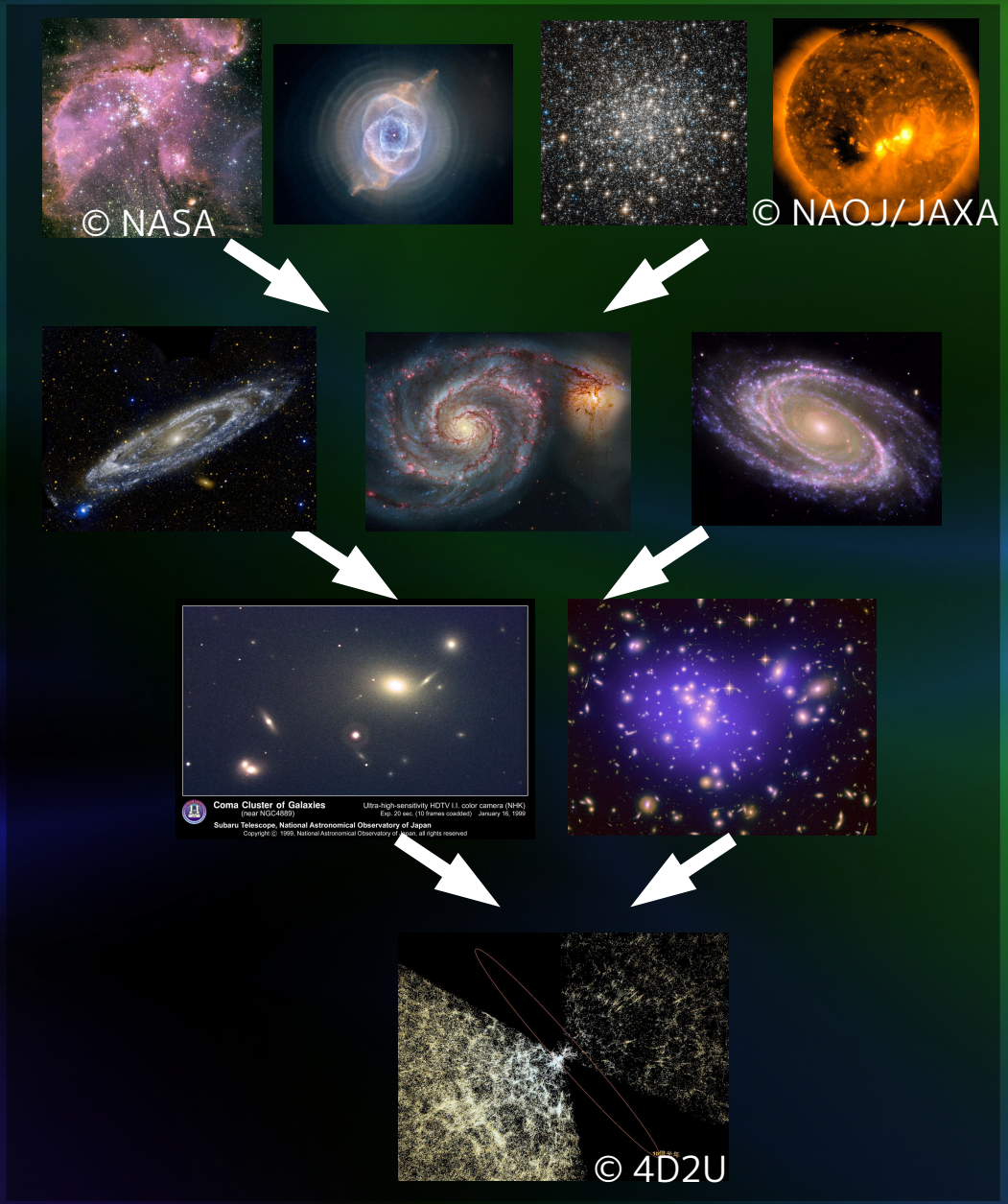


宇宙の階層構造

- 星があつまって銀河に
- 銀河があつまって銀河団に
- 銀河団があつまって
  宇宙の大規模構造に

  というように宇宙は
  階層的な構造を示している

- これらの構造形成には重力の
  みはたらく**ダークマター**が
  必要不可欠 (CDM理論)



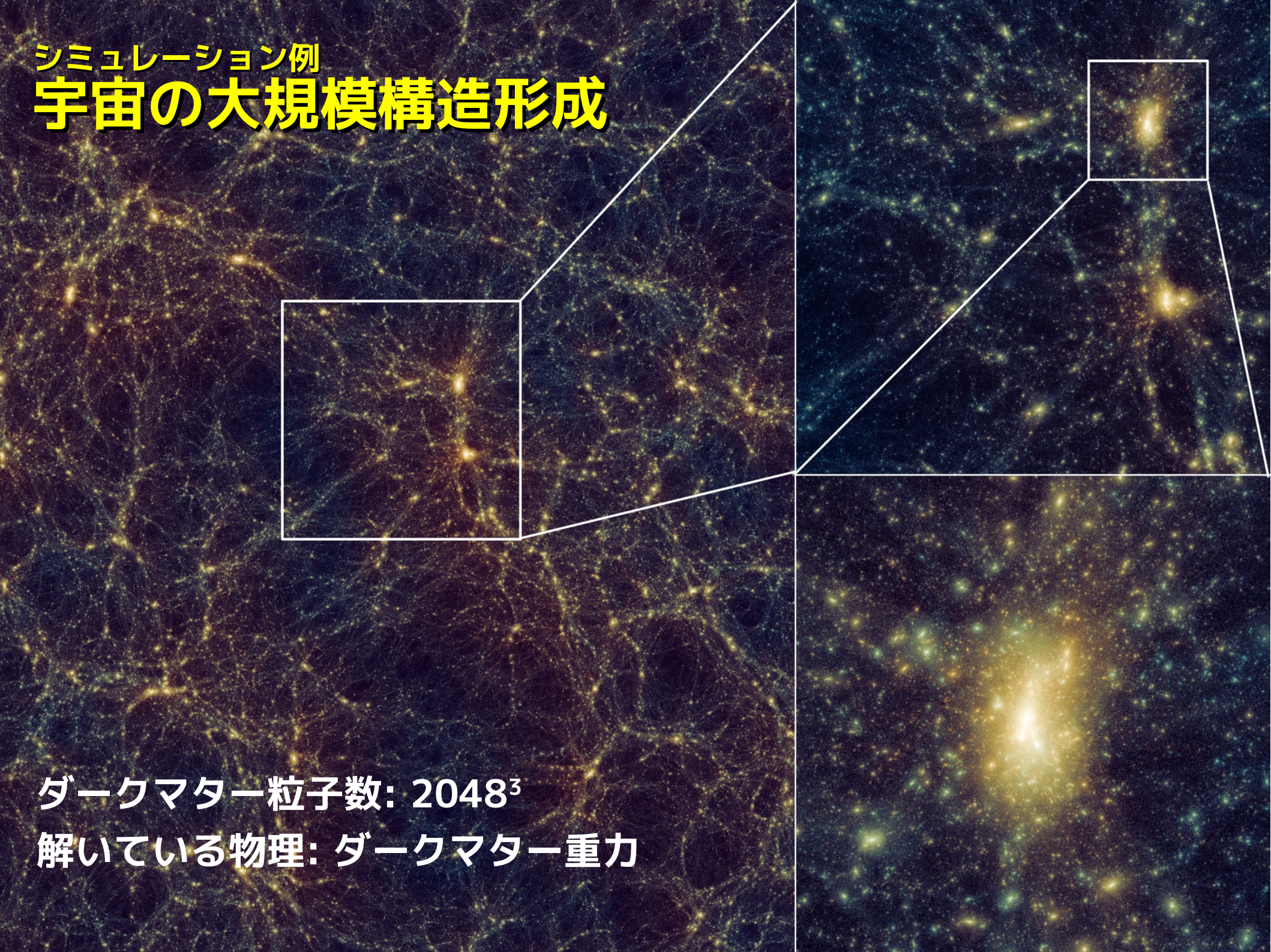
ダークマターの密度揺らぎによる宇宙の構造形成

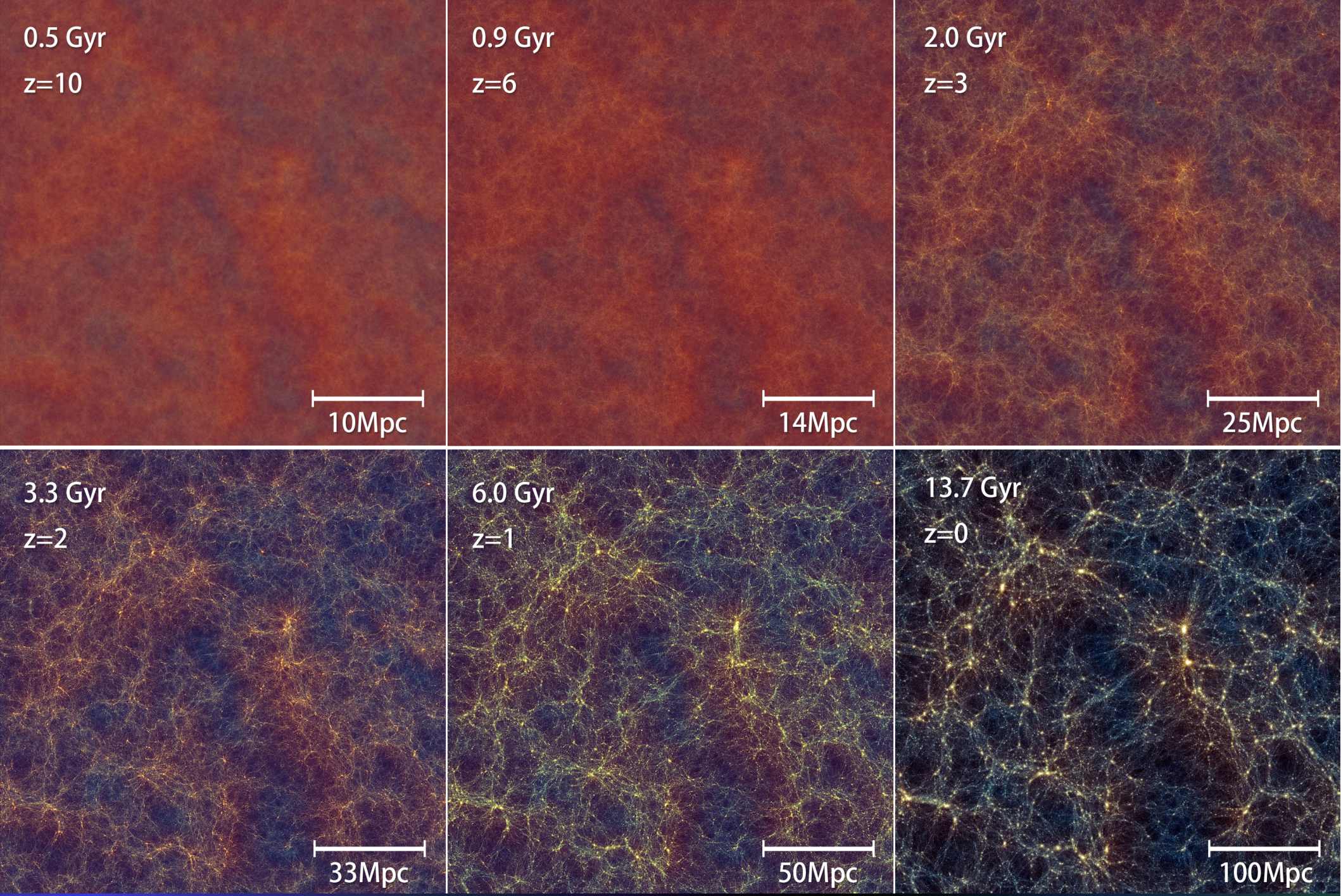
ダークマター

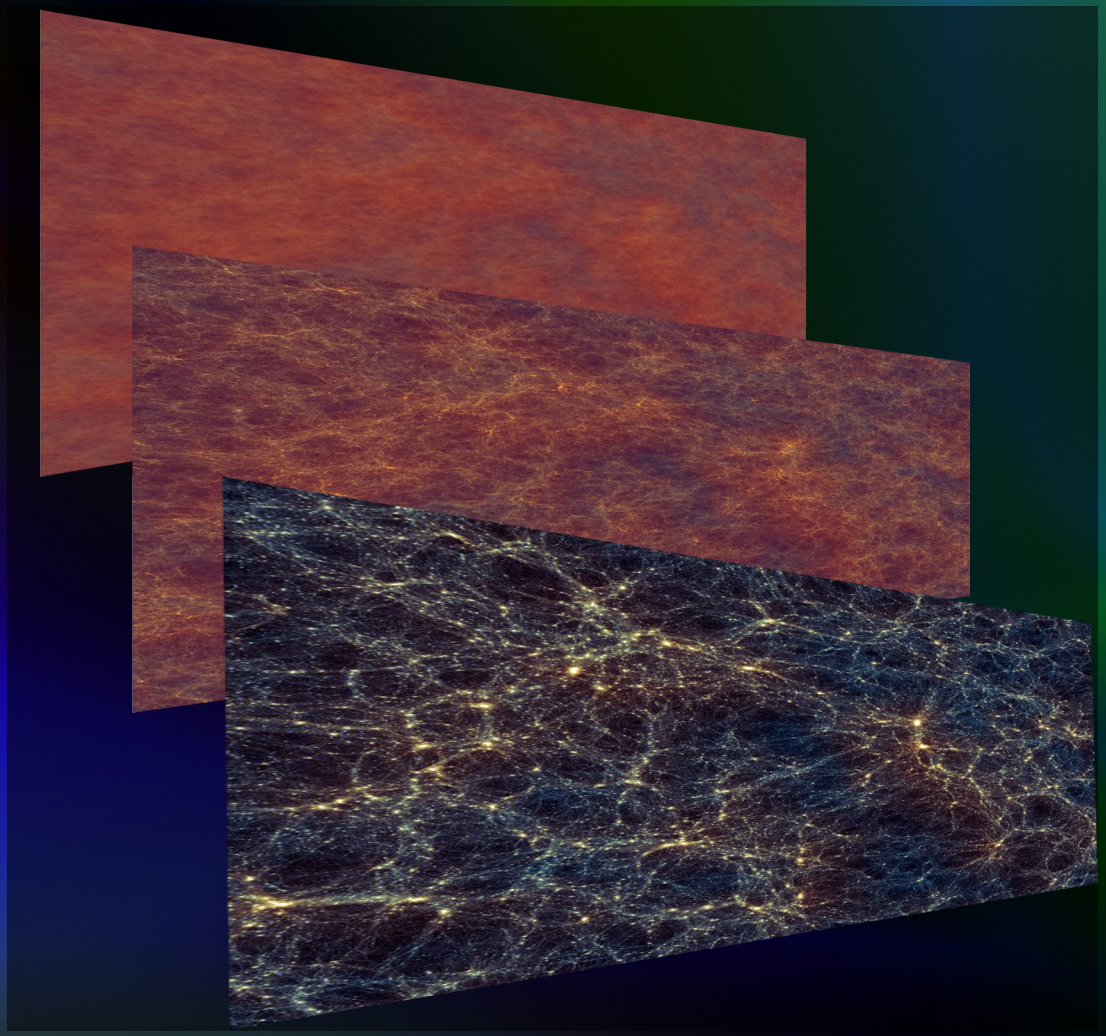
- 宇宙の全物質の85%を占める
- ビッグバン直後はほとんど一様に存在
- 重力のみはたらく

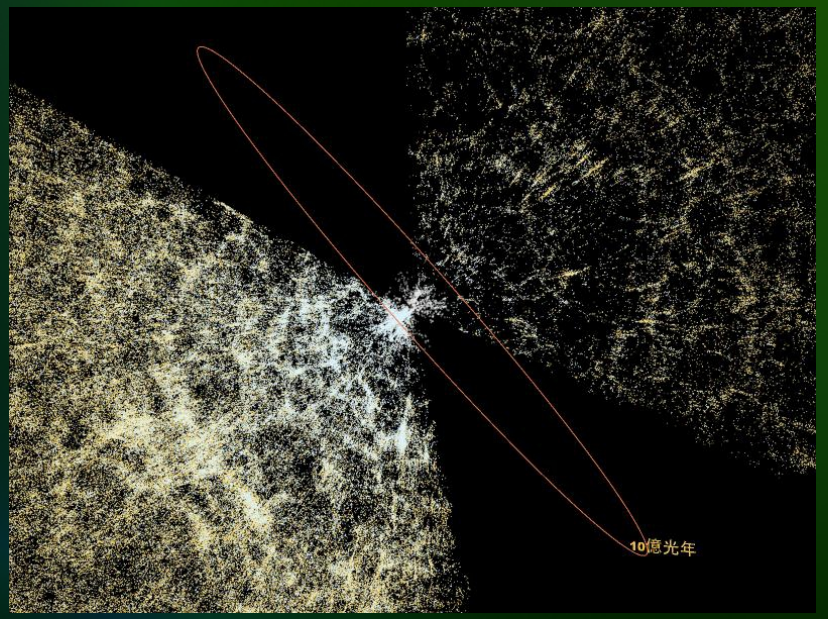
構造形成

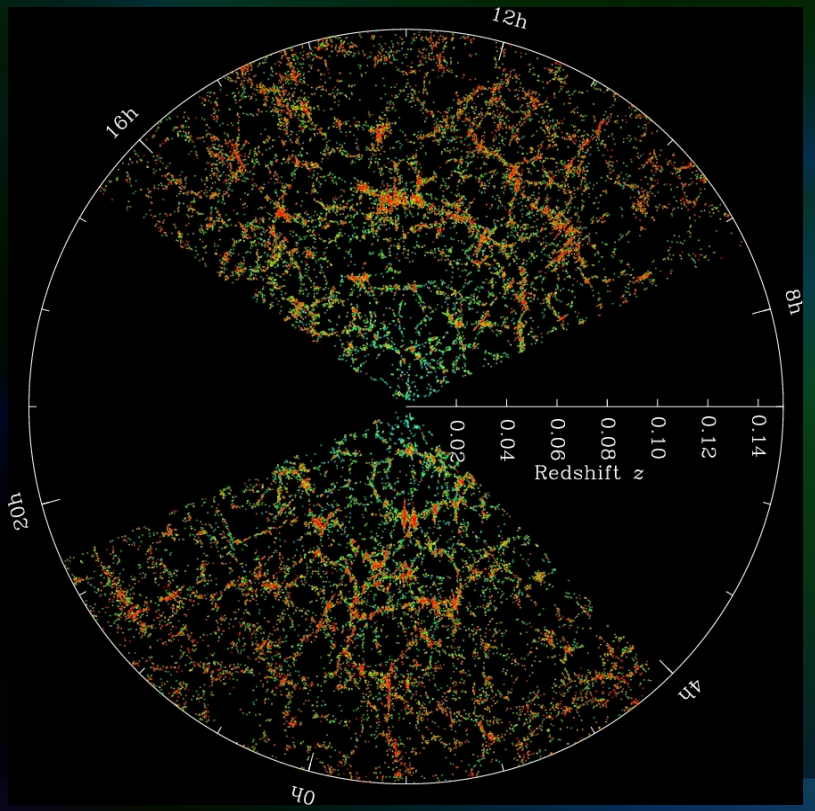
- 重力によってわずかなダークマターのむら(密度揺らぎ)が成長。中心が高密度な天体を形成(ダークマターハロー)
- ハローが合体して大きいハローを形成。ガスが集って、星や銀河ができた


シミュレーション例
宇宙の大規模構造形成
ダークマター粒子数: 2048^3
解いている物理: ダークマター重力


0.5 Gyr
z=10
10Mpc
0.9 Gyr
z=6
14Mpc
2.0 Gyr
z=3
25Mpc
3.3 Gyr
z=2
33Mpc
6.0 Gyr
z=1
50Mpc
13.7 Gyr
z=0
100Mpc



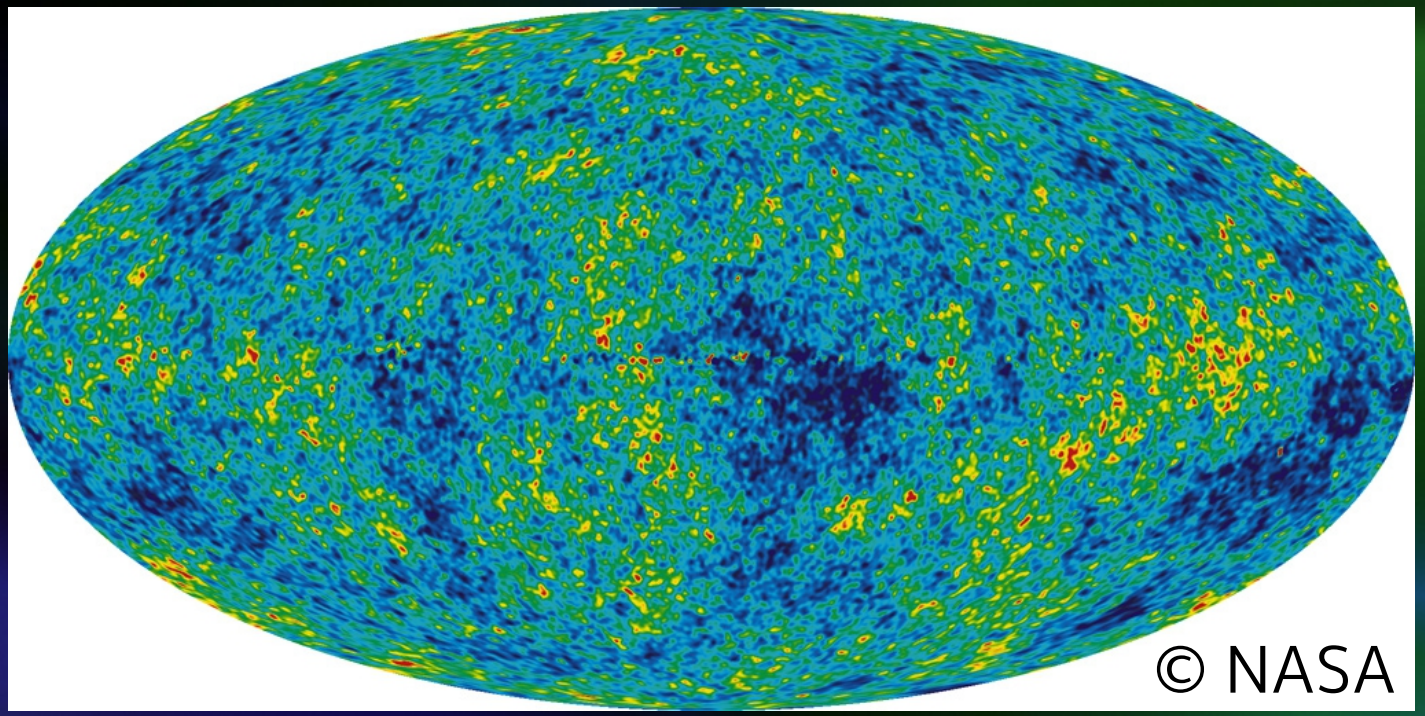




シミュレーション（左）と
観測（右）

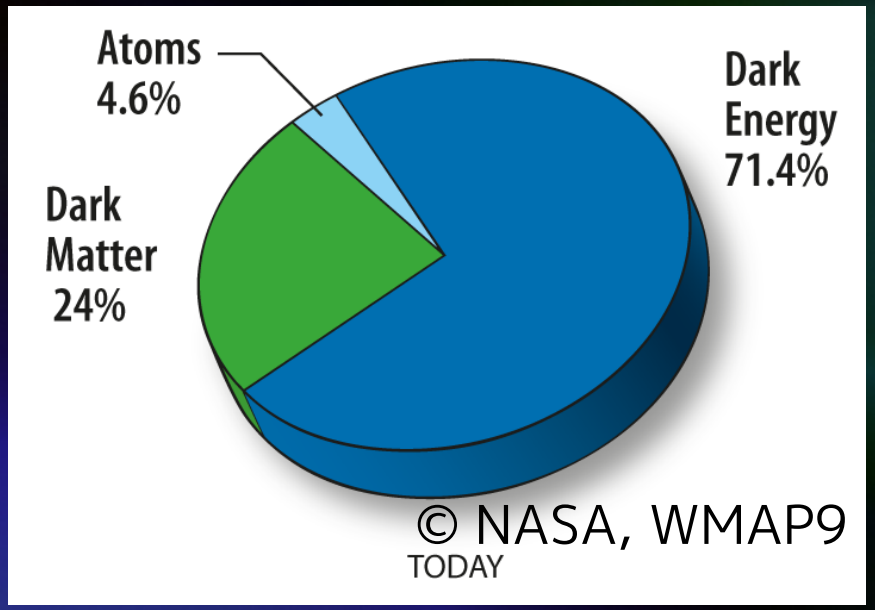
初期条件

WMAPによる密度揺らぎの精密測定(2001~)



- ビッグバンの観測的証拠
- 宇宙全体でほとんど等方で、そのスペクトルは平均温度2.725Kの黒体輻射で良く近似できる
- 10^{-5}程度の等方性が存在 (密度揺らぎそのもの)
- 宇宙初期のさまざまな情報を得ることができ、宇宙年齢などを記述する宇宙モデルに大きな制限をつけることができる

宇宙のエネルギー成分

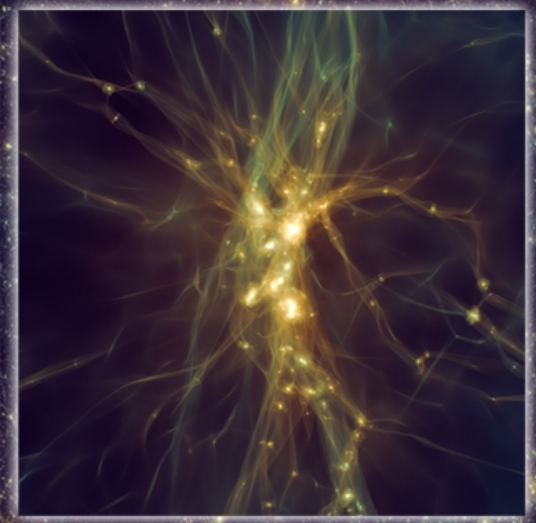


- **バリオン 4.6%**
  - 原子など目に見えるもの
- **ダークマター (暗黒物質) 24%**
  - 重力のみはたらく
  - 宇宙の構造形成に寄与
  - コールド(速度分散が小さい)
- **ダークエネルギー 71.4%**
  - 宇宙膨張に寄与


ダークマターハロー形成
ダークマター粒子数 1億
Visualization:
Takaaki Takeda
(4D2U, National Astronomical Observatory of Japan)

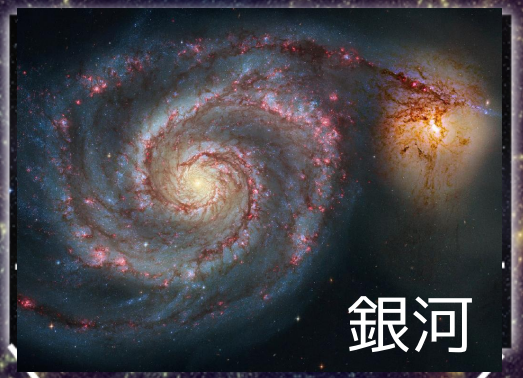
ダークマターは、星や銀河などの天体形成に必要不可欠

背景：宇宙の大規模構造



初代ハロー

初代星

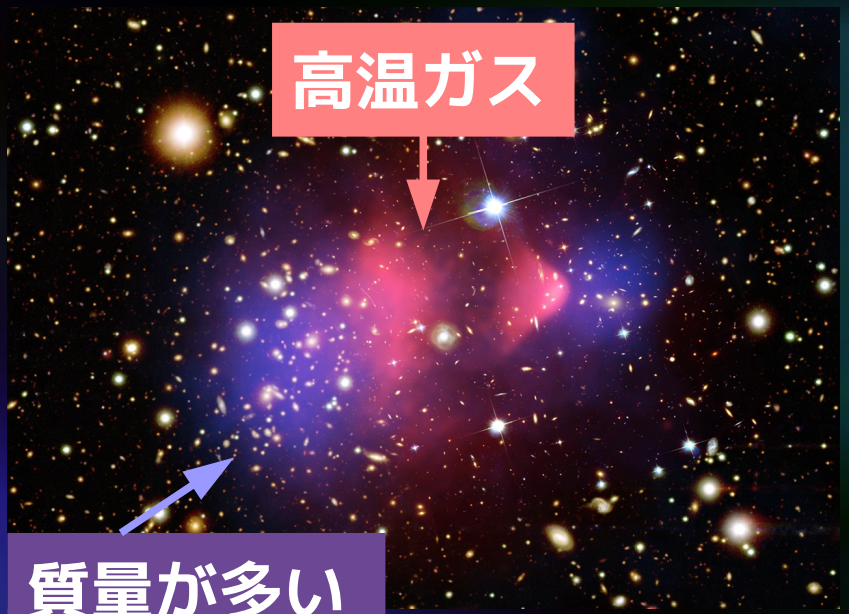


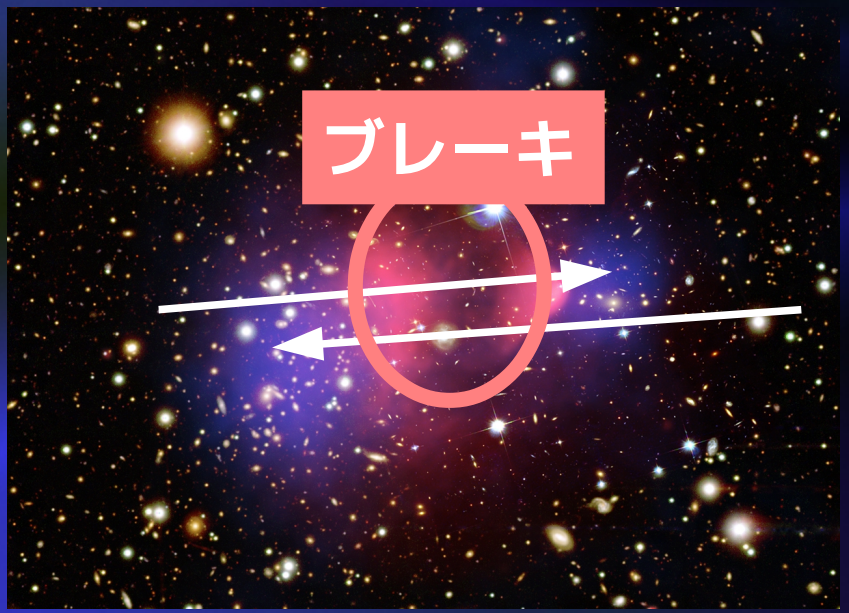
銀河ハロー

銀河団

<10　　<1万　　100万　　1000万　　光年

ダークマターが存在する証拠: 弾丸銀河団





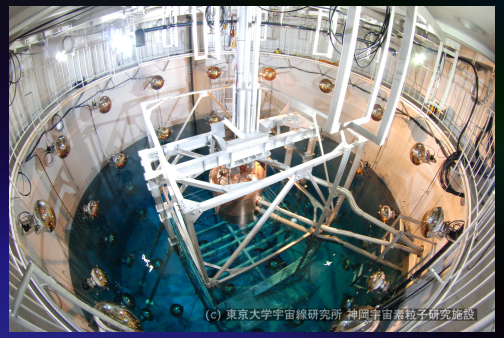
- 高温ガスの分布と、質量の分布が異なる
- 高温ガスは重力の他に、摩擦等により減速する
- 高温ガスが質量の大部分なら、両者の分布は同じはず

重力のみで相互作用するダークマターの存在の決定的な証拠

宇宙最大の謎の一つ

ダークマター粒子の正体

- ダークマター素粒子の正体は不明。ダークマターの直接・間接検出は宇宙物理学、素粒子物理学のグランドチャレンジ



- 一つの有力候補：質量100GeV~TeVの超対称性粒子 ニュートラリーノ
  - 対消滅してガンマ線を放出する (ダークマター間接検出実験)
- 太陽近傍でガンマ線フラックスが最大なのはどこか?

- 対消滅の回数はダークマター密度の2乗に比例。
  銀河系ダークマターの微細構造を解明する必要がある

ダークマターハローの構造

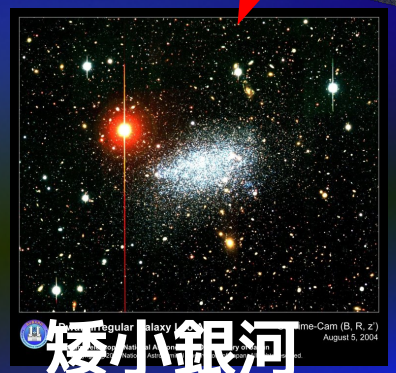
大規模数値
シミュレーション
ではじめて解明
されたこと

- 中心が高密度
- 楕円状
- 無数のサブハローが存在
  (小さいもの程数が多い)

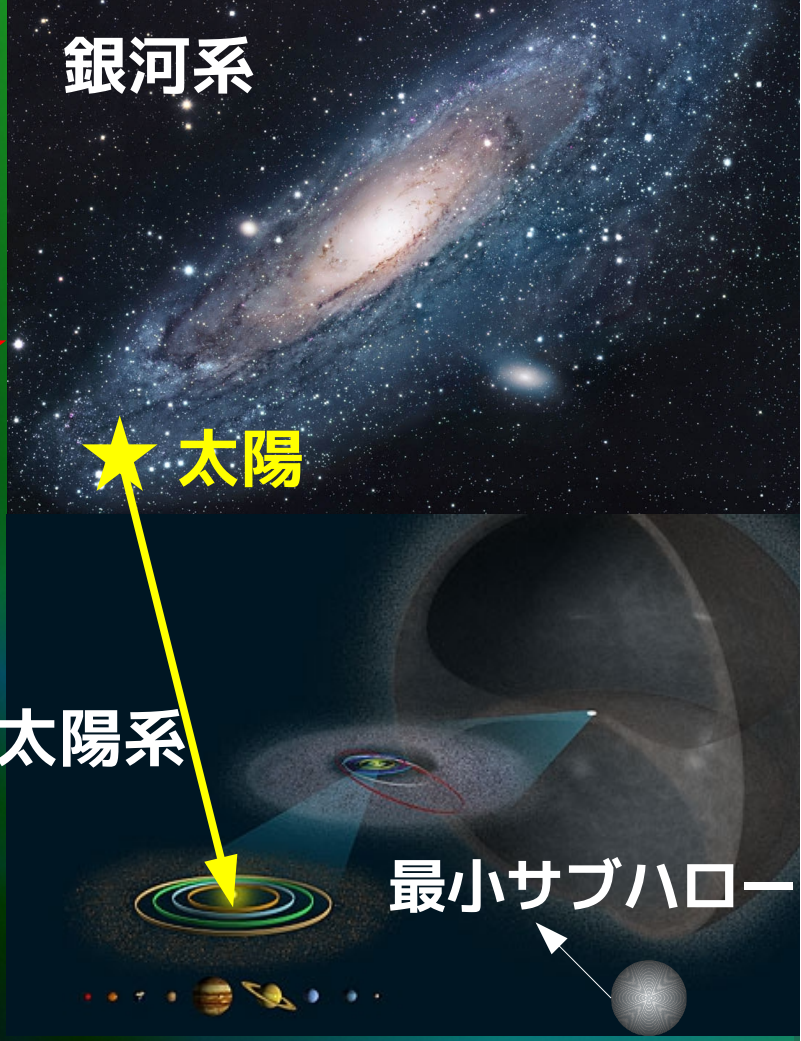
銀河系の構造

銀河系のダークマターハロー

100万光年



矮小銀河



- 星やガスなどのバリオン成分は、1兆太陽質量程度のハローの中心に存在している
- 無数のサブハローが存在 (10^{-6}~10^{10}太陽質量)
  - サブハローの質量が1/10 → 数は10倍

ダークマターの正体解明へ向けて

何を明らかにしなければならないのか?

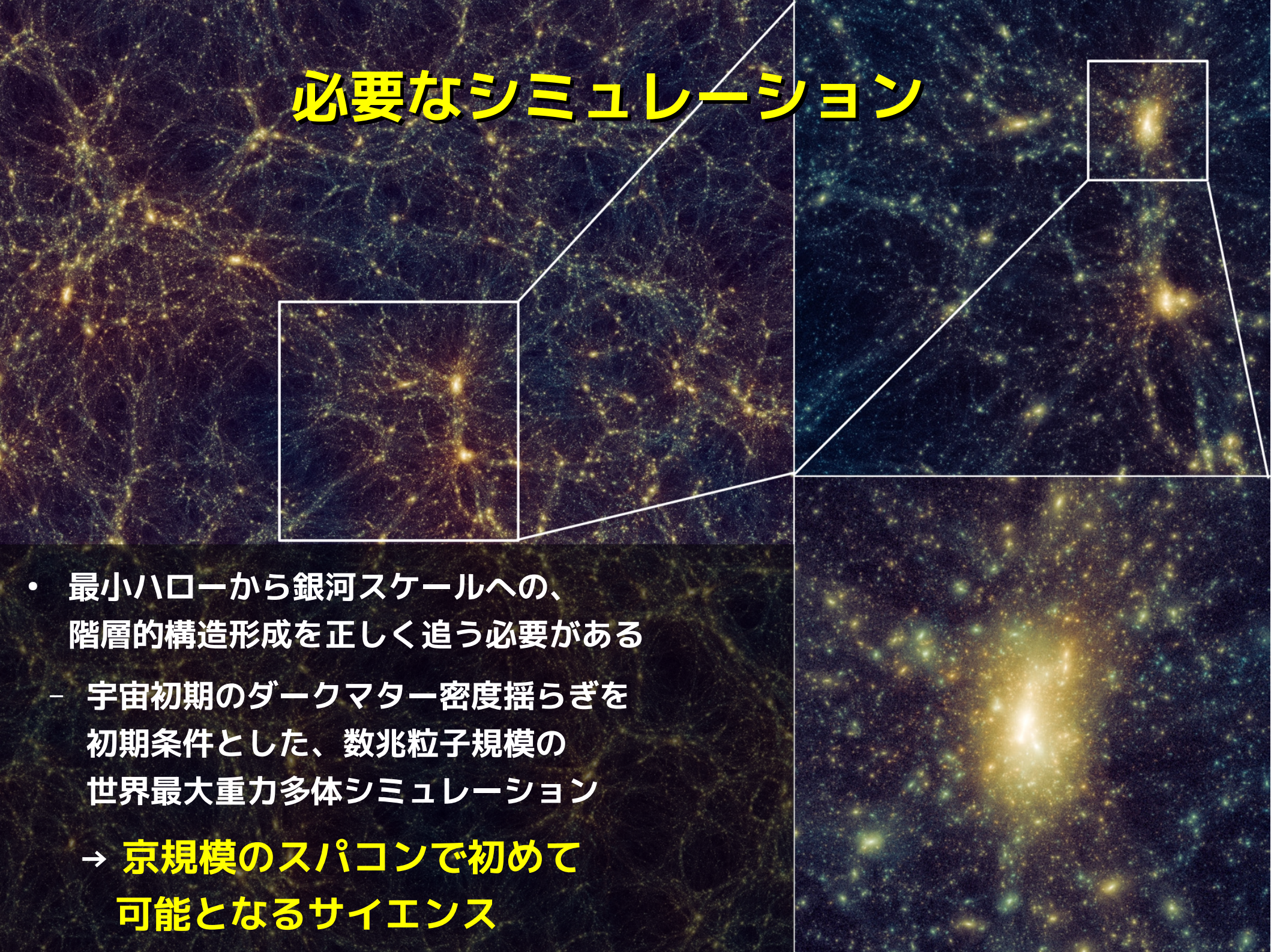
- 最小サブハローが太陽近傍に存在すれば、ガンマ線源として検出可能 (Diemand et al. 2005, Nature)
- 最小サブハローは大きいものと比べ**高密度**で、銀河系内に数多く存在し得る (Ishiyama et al. 2010)

↓

- 銀河系ダークマターハローの**微細構造**を明らかにする必要がある
  - 最小サブハローが太陽近傍にどれくらい生き残れるか

  → ダークマター検出に最も適した場所が明らかになる

必要なシミュレーション



- 最小ハローから銀河スケールへの、
  階層的構造形成を正しく追う必要がある
  - 宇宙初期のダークマター密度揺らぎを
    初期条件とした、数兆粒子規模の
    世界最大重力多体シミュレーション

→ **京規模のスパコンで初めて**
可能となるサイエンス


テスト計算 (粒子数168億)
Visualization:
Takaaki Takeda
(4D2U, National Astronomical Observatory of Japan)

シミュレーションコードの準備状況

- 京の全システムを使って、 ~**5.8Pflops** の実行性能 (~**55**% の対ピーク性能比)を達成
- 外国製 (米アルゴンヌ)の同様のコードと比べ、**5**倍近い速度でシミュレーションを実行可能

- ピーク性能 20Pflops のセコイアを退け、SC12で**ゴードンベル賞を単独受賞**
- 大規模システム向けの新しい並列アルゴリズム (Ishiyama et al, 2012)

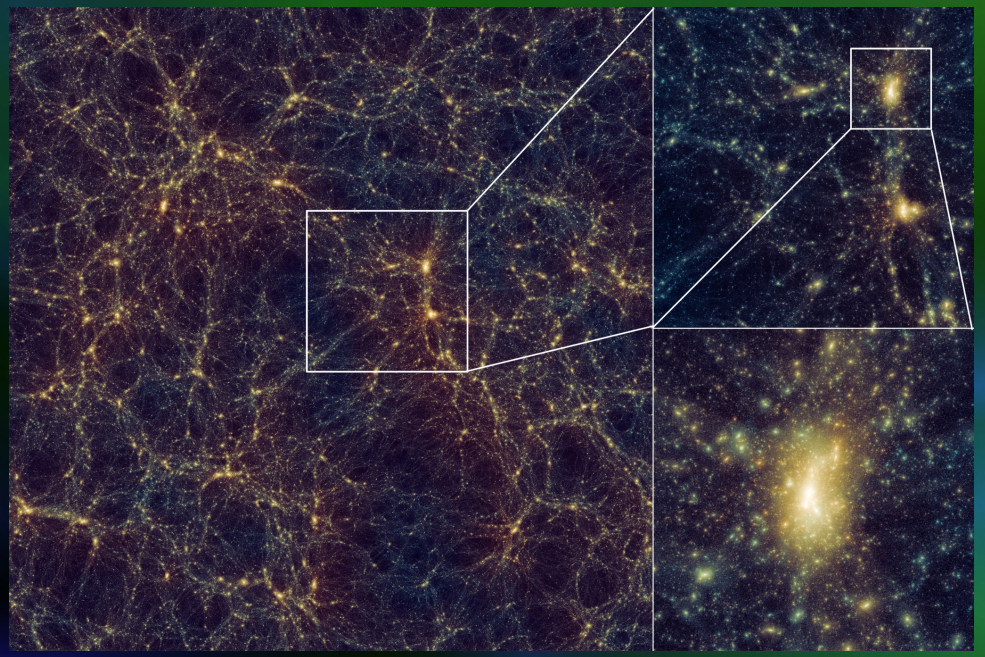
シミュレーション

コードの共同開発者
似鳥啓吾 (AICS)

重力多体 (N体) シミュレーション

- 重力多体系をN個の質点で表現。粒子間の相互作用重力を計算し、運動方程式を時間積分、系の時間発展を追う

$$\frac{d^2 r_i}{dt^2} = \sum_{j \neq i}^{N} G m_j \frac{r_j - r_i}{|r_j - r_i|^3}$$



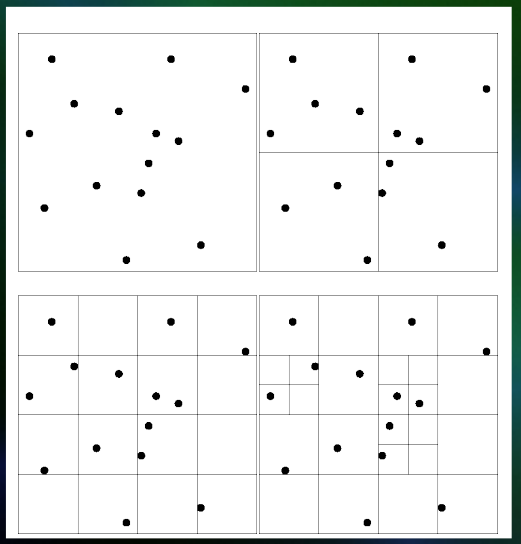
- 対象の例
  - **ダークマター構造形成**
  - 銀河・銀河団
  - 星団
  - 惑星系
  - ブラックホールの運動

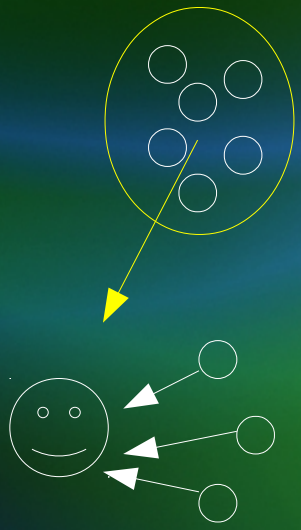
重力計算アルゴリズム

- **直接計算**：全粒子対全粒子の力を計算する
  - **O(N²)**

$$\frac{d^2 r_i}{dt^2} = \sum_{j \neq i}^{N} G m_j \frac{r_j - r_i}{|r_j - r_i|^3}$$

- **ツリー法**：近傍の粒子との相互作用は直接、
  遠方の粒子群との相互作用はまとめて
  多重極展開で計算する
  - **O(N²) から O(*N* log*N*)へ**
  - **Modified algorithm**:
    相互作用リストを粒子グループで共有する



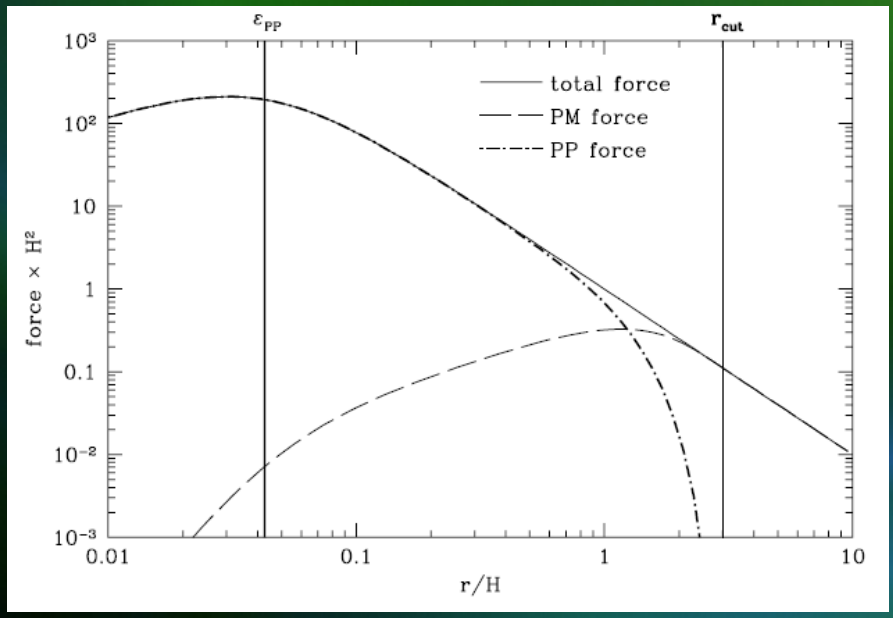


- **PM法**：一様格子上の密度場を計算し、FFTを用いてポアソン方程式を解く
  - 周期境界条件を自然に解ける

宇宙論N体での主流

TreePM法

- 近距離力(カットオフつき)をTree法、遠距離力をPM法で解く
  - $O(N^2) \rightarrow O(N \log N)$
  - 周期境界条件を実現
- 比較的新しいアルゴリズム
  - GOTPM (Dubinski+ 2004)
  - GADGET-2 (Springel 2005)
  - GreeM (Ishiyama+ 2009)
  - HACC (Habib+ 2012)

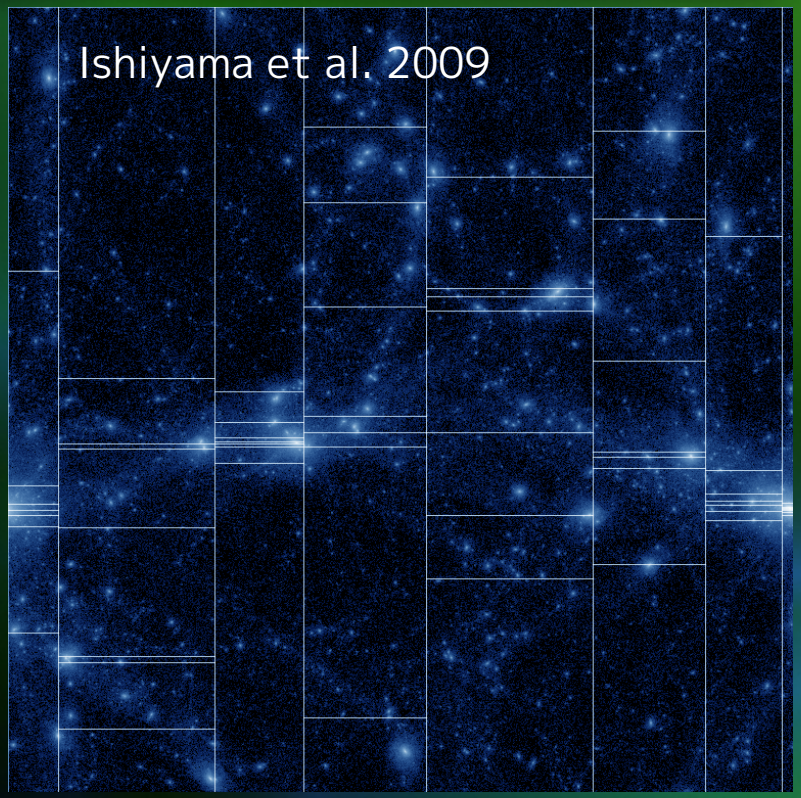


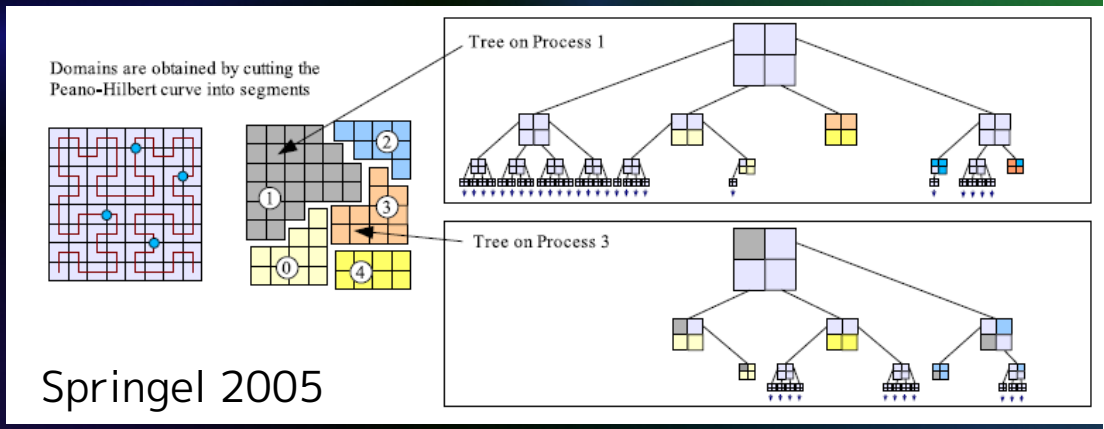
$$\boldsymbol{a}_i = \sum_{i \neq i} \frac{m_j(\boldsymbol{r}_j - \boldsymbol{r}_i)}{|\boldsymbol{r}_j - \boldsymbol{r}_i|^3} g_{\mathrm{P3M}}(|\boldsymbol{r}_j - \boldsymbol{r}_i|)/\eta),$$

$$g_{\mathrm{P3M}}(R) = \begin{cases} 1 - \frac{1}{140}\left(224R^3 - 224R^5 + 70R^6 + 48R^7 - 21R^8\right) & (0 \leq R \leq 1) \\ 1 - \frac{1}{140}\left(12 - 224R^2 + 869R^3 - 840R^4 + 224R^5 + 70R^6 - 48R^7 + 7R^8\right) & (1 \leq R \leq 2) \\ 0 & (2 \leq R) \end{cases},$$

並列化

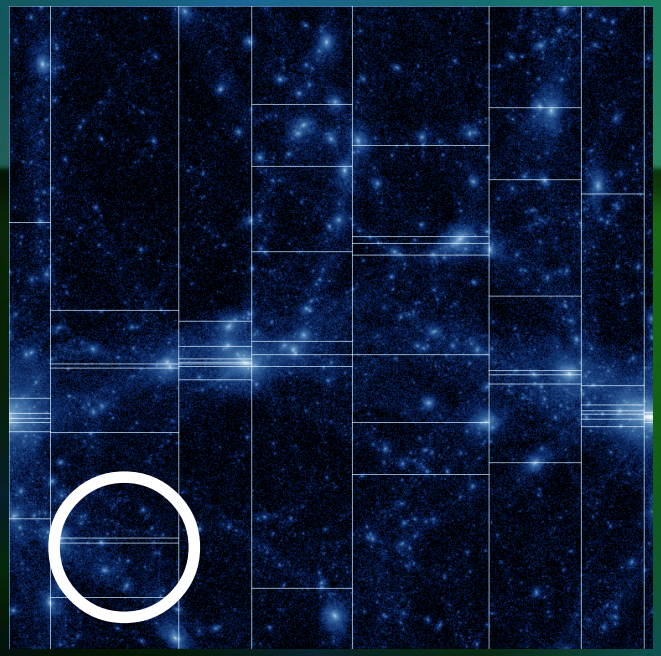
1. 全空間を分割し計算ノードに割り当て、
   粒子を再配分する
   - サンプリング粒子の集約 (全対通信)
   - 一部粒子情報を通信 (隣接通信)
1. 長距離重力の計算 (PM 法)
   - 全メッシュ情報を通信
     (全対通信、粒子よりは通信量小)
2. 短・中距離重力の計算 (Tree法)
   - 一部ツリー情報を通信 (隣接通信)
3. 粒子の時間積分
4. 1に戻る

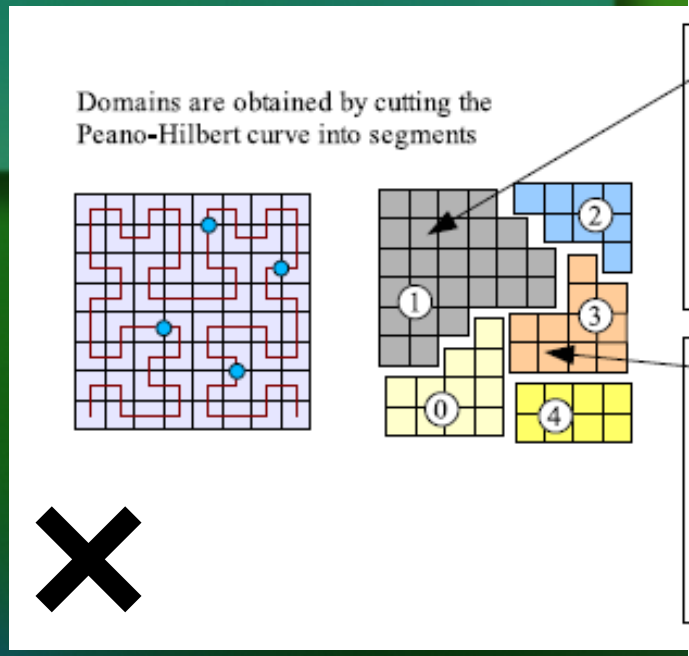




動的領域分割

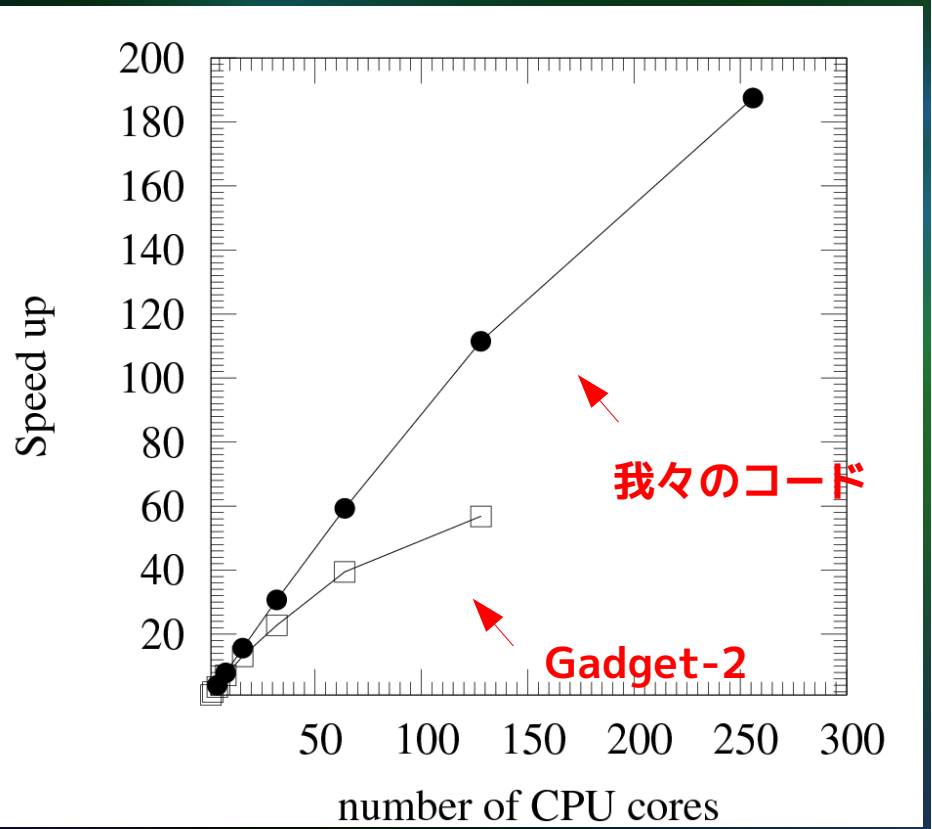
- 領域分割の仕方が重要





- × 粒子数均等
- ○ 相互作用数均等
- **◎ 相互作用均等 + 補正**
  - 計算時間が一定になるようにする

- △ 空間充填曲線
- ○ **再帰的多段分割**
  - 隣接ノードが自明



Load balancer

- Sampling method to determine the geometry
  - Each node samples particles and shares with all nodes
  - Sample frequency depends on the local calculation cost -> realizes near-ideal load balance

$$n_{\mathrm{samp},i} = N R_{\mathrm{samp}} f_{\mathrm{samp},i}$$

R~10^{-3} ~ 10^{-5}

$$f_{\mathrm{samp},i} = \frac{t_{\mathrm{PP},i} + t_{\mathrm{PM},i}}{\sum_j \left(t_{\mathrm{PP},j} + t_{\mathrm{PM},j} \right)}$$

  - New geometry is adjusted so that all domains have the same number of sampled particles
  - Linear weighted moving average for last 5 steps

ネットワーク性能について使ってみた印象

- (コードによってはそんなことはないだろうが) 数千ノードまでは、
  他のスパコンと大差はない印象

- 数千ノード以上

  長距離通信の性能低下が見え始めてくる

- 数万ノード以上

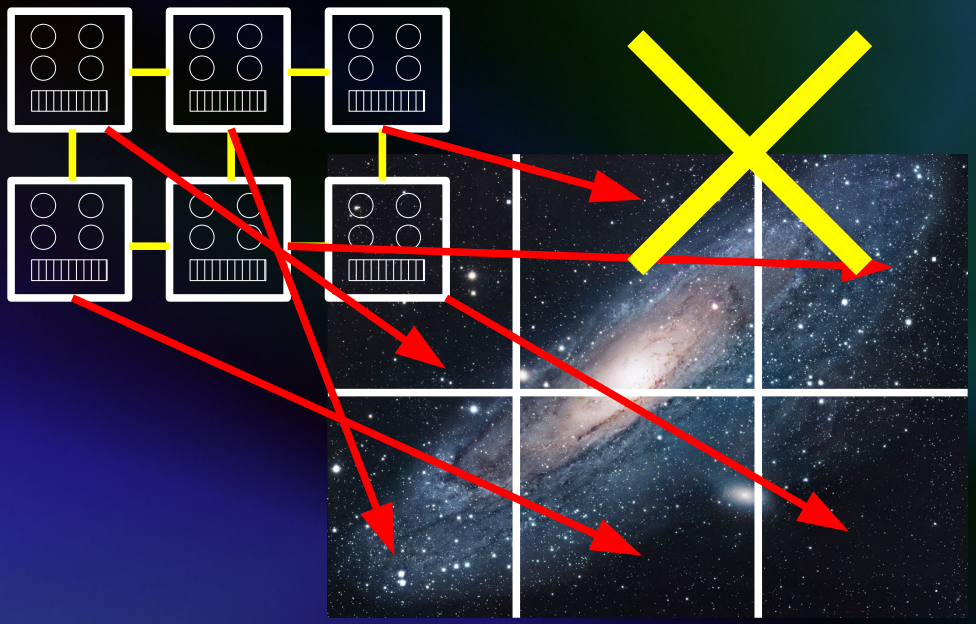
  ボトルネックに

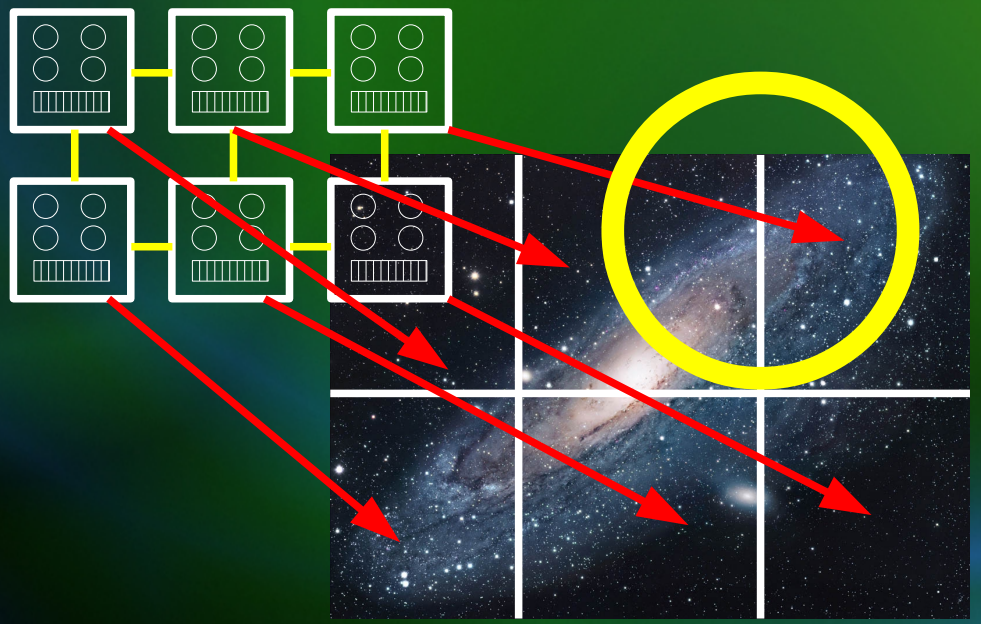
トーラスマッピングと、
MPI_COMM_WORLD の分割で改善

トーラスマッピングによる通信最適化

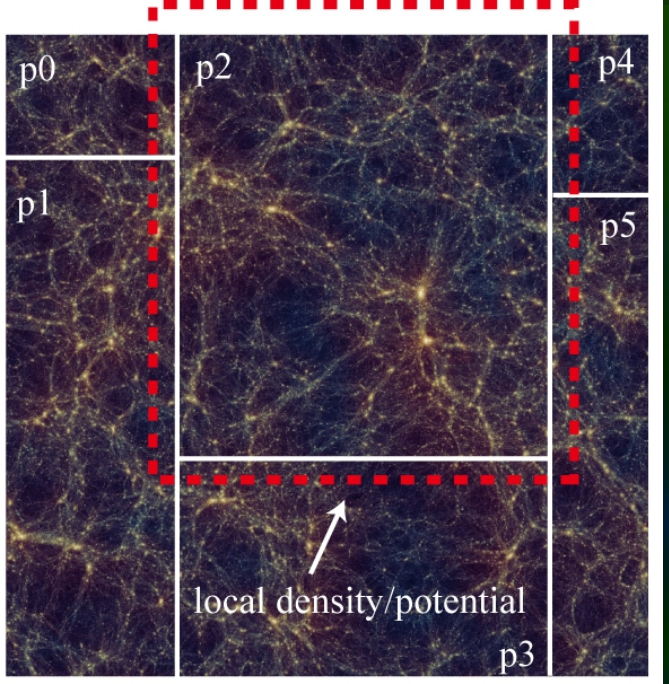
- 特に何も指定しないと、どういった形状でノードが確保されるかわからない
- 32ノードジョブで 32x1x1 のような形状になる可能性もある
  - 長距離通信のホップ数が大きい
  - シミュレーション空間を 4x4x2 で切ってしまうと、シミュレーション空間上の隣接通信が、ノード上は隣接していない通信になることも
- ノード形状を指定
  - ただしジョブは流れにくくなる
- 各領域を担当するノードの物理配置を、実際の領域分割にあわせる
  - 仮想3次元トーラス上の座標を取得するAPIが存在する

トーラスマッピングによる通信最適化



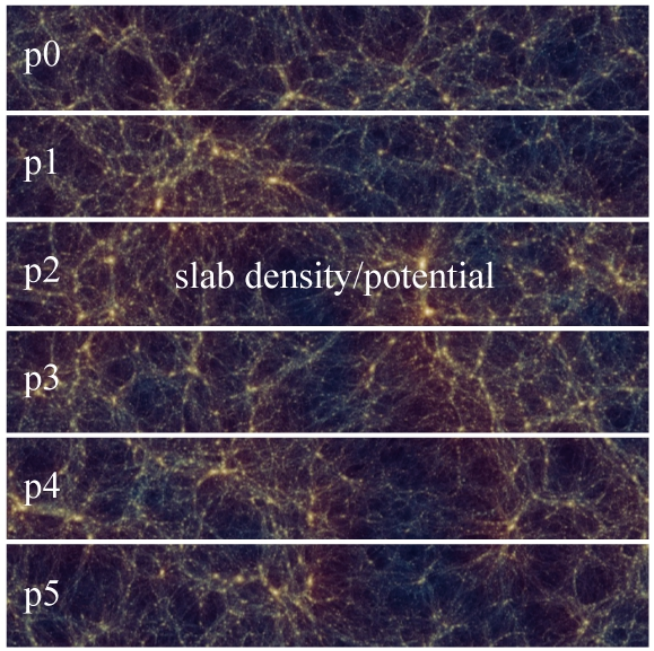


- 4096 ノード、16x16x16の空間分割で合計20MBの Allgather をした場合
  - ノード形状 10x15x28 0.06 sec
  - 16x16x16 0.04 sec



2) density comm

4) potential comm



3) FFT & convolution

MPI_COMM_WORLDの分割による改善例1

並列PM

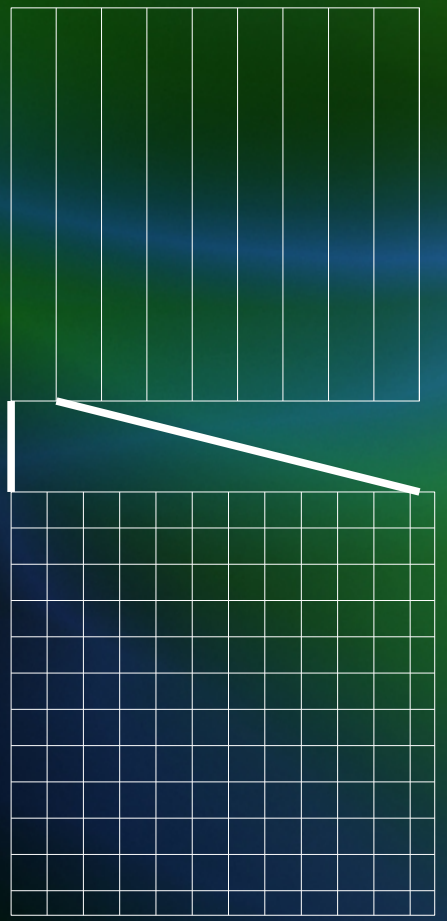
1. density assignment
2. local density → slab density
   - **長距離通信**
3. FFT (density → potential)
   - 長距離通信 (library 任せ)
4. slab potential → local potential
   - **長距離通信**
5. force interpolation

空間分割は不規則三次元分割だが、FFT は規則的な 1or2or3 次元分割のデータ構造になっている必要があるため

MPI_COMM_WORLDの分割による改善例1：並列PM

もう少し具体的に

- ~10000^3粒子、4096^3 PMメッシュのシミュレーションを京のフルシステム(48x54x32=82944 ノード)で行う場合
  - 最適な領域分割は 48x54x32 (network topology 依存)
  - 1D並列FFTW を使うと FFT は 4096 並列で可能



1FFTプロセスは **数千ノード** から
メッシュデータを受け取る必要有!!!

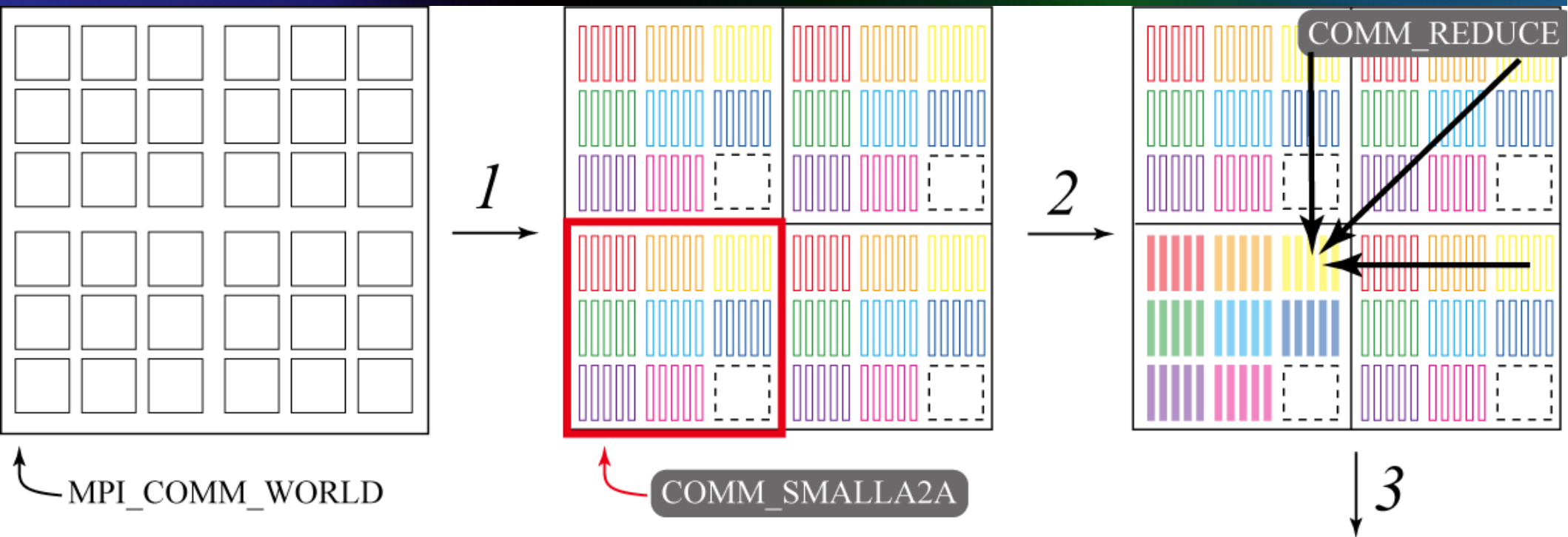
- MPI_Isend、MPI_Irecv は無理
- MPI_Alltoallv を使うと簡単に実装できるが、、、
- 2D、3D FFT なら幾分ましだが、通信競合は発生する

MPI_COMM_WORLDの分割による改善例1：並列PM

通信の階層化 (MPI_Comm_split を使う)

- MPI_Alltoallv (..., MPI_COMM_WORLD)
  - 1.MPI_Alltoallv (..., COMM_SMALLA2A)
  - 2.MPI_Reduce(..., COMM_REDUCE)

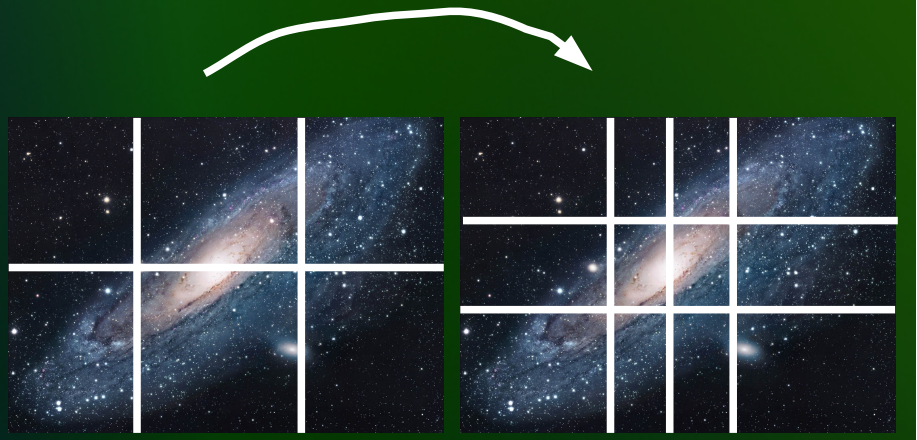
3~4 倍の通信高速化に成功!

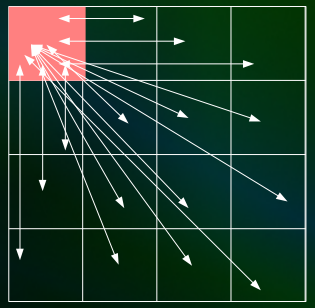


MPI_COMM_WORLDの分割による改善例2

全データのシャッフル

- 全ノードのデータをシャッフルしたくなることもあるかもしれない
  - restart時にノード数が変わる時など
- 実装は Alltoallv が楽。だが数千〜数万並列になると色々問題が……
  - 通信回数は O(p)、p は並列数
- 通信を階層的にすることで解決
  - 斜め通信を避け、通信競合を防ぐ
  - 近接通信に使えることも



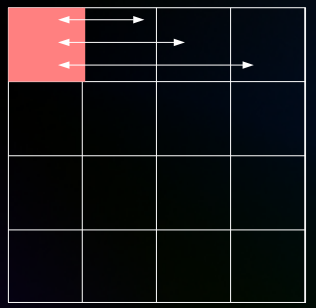


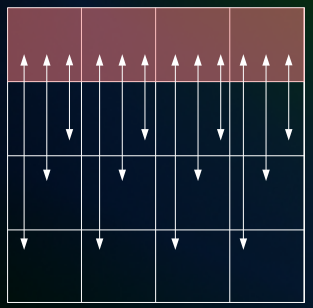
MPI_Alltoallv
(…, COMM_WORLD)

O (p)

MPI_Alltoallv (…, COMM_X) +
MPI_Alltoallv (…, COMM_Y) +
MPI_Alltoallv (…, COMM_Z)

O ($p^{1/3}$)





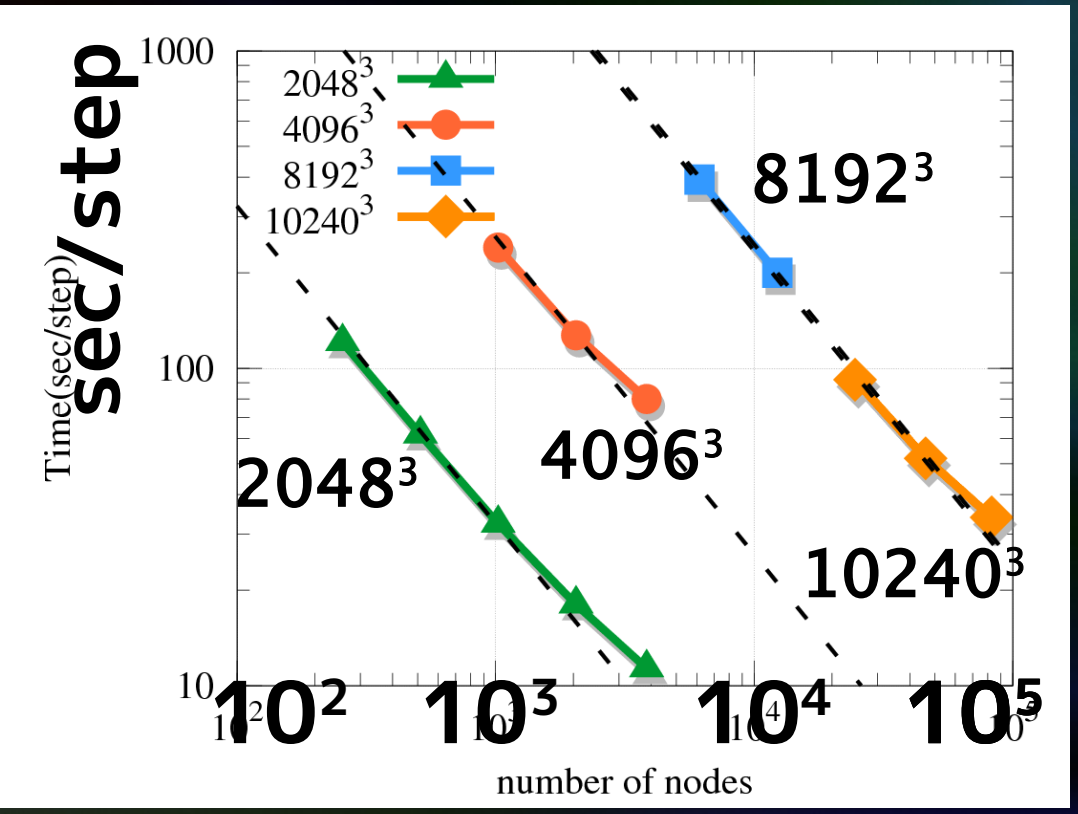
MPI_COMM_WORLDの分割による改善例2

全データのシャッフル

- 粒子数2048^3、1024分割、合計384GByte
  - 4096ノード MPI_COMM_WORLD → **17.7 sec**
  - 4096ノード 階層的通信 → **4.2 sec**

- たいした違いはないと思われるかもしれないが、全ノード (82944) では、前者は1時間以上かかりそうな勢いだった記憶が
(データがなくてすいません)
- 後者はそのために急遽実装するはめになった

Performance results on K computer



- Scalability (2048^3 – 10240^3)
  - Excellent strong scaling
  - 10240^3 simulation is well scaled from 24576 to 82944 (full) nodes of K computer
- Performance (12600^3)
  - The average performance on full system is ~**5.8Pflops**,

~**55**% of the peak speed

Ishiyama, Nitadori, and Makino, 2012 (arXiv: 1211:4406), SC12 Gordon Bell Prize Winner

まとめ

- 世界ではじめて、最小スケールからのダークマターハロー構造形成過程を追い、ハローの微細構造を明らかにする
  - 京規模のスパコンによる大規模シミュレーションが必要不可欠
- ダークマターを"観測"するのに最も適した場所を明らかにする
  (ダークマター検出は、世界中の研究機関が一番乗りを目指す超激戦区)

- そのためのシミュレーションコードを開発し、京のフルシステムを用い、**5.8**Pflops (**55%** の対ピーク性能比)を達成
  - **SC12でゴードンベル賞を単独受賞**
- プロダクトランを始める準備は整った
  - 後は計算リソースの問題
  - 大規模ジョブが流れにくい方が深刻か