The Astrophysical Journal, 724:244–266, 2010 November 20 doi:10.1088/0004-637X/724/1/244

REIONIZATION SIMULATIONS POWERED BY GRAPHICS PROCESSING UNITS. I. ON THE STRUCTURE OF THE ULTRAVIOLET RADIATION FIELD

Dominique Aubert[1] and Romain Teyssier[2,3]
[1] Observatoire Astronomique de Strasbourg, Universite de Strasbourg, CNRS UMR 7550, 11 rue de l'Universite, F-67000 Strasbourg, France
[2] IRFU, CEA Saclay, Batiment 709, F-91191 Gif-sur-Yvette Cedex, France
[3] Institüt fur Theoretische Physik, Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
Received 2010 March 31; accepted 2010 September 13; published 2010 November 1

ABSTRACT

We present a set of cosmological simulations with radiative transfer in order to model the reionization history of the universe from $z = 18$ down to $z = 6$. Galaxy formation and the associated star formation are followed self-consistently with gas and dark matter dynamics using the RAMSES code, while radiative transfer is performed as a post-processing step using a moment-based method with the M1 closure relation in the ATON code. The latter has been ported to a multiple Graphics Processing Unit (GPU) architecture using the CUDA language together with the MPI library, resulting in an overall acceleration that allows us to tackle radiative transfer problems at a significantly higher resolution than previously reported: $1024^3 + 2$ levels of refinement for the hydrodynamic adaptive grid and 1024^3 for the radiative transfer Cartesian grid. We reach a typical acceleration factor close to $100\times$ when compared to the CPU version, allowing us to perform $1/4$ million time steps in less than 3000 GPU hr. We observe good convergence properties between our different resolution runs for various volume- and mass-averaged quantities such as neutral fraction, UV background, and Thomson optical depth, as long as the effects of finite resolution on the star formation history are properly taken into account. We also show that the neutral fraction depends on the total mass density, in a way close to the predictions of photoionization equilibrium, as long as the effect of self-shielding are included in the background radiation model. Although our simulation suite has reached unprecedented mass and spatial resolution, we still fail in reproducing the $z \sim 6$ constraints on the neutral fraction of hydrogen and the intensity of the UV background. In order to account for unresolved density fluctuations, we have modified our chemistry solver with a simple clumping factor model. Using our most spatially resolved simulation (12.5 Mpc h^{-1} with 1024^3 particles) to calibrate our subgrid model, we have resimulated our largest box (100 Mpc h^{-1} with 1024^3 particles) with the modified chemistry, successfully reproducing the observed level of neutral hydrogen in the spectra of high-redshift quasars. We however did not reproduce the average photoionization rate inferred from the same observations. We argue that this discrepancy could be partly explained by the fact that the average radiation intensity and the average neutral fraction depend on different regions of the gas density distribution, so that one quantity cannot be simply deduced from the other.

Key words: dark ages, reionization, first stars – methods: numerical – radiative transfer

Online-only material: color figures

1. INTRODUCTION

After self-gravity, hydrodynamics, and radiative cooling (see, e.g., Efstathiou et al. 1985; Hernquist et al. 1991; Cen 1992; Katz et al. 1996; Bertschinger 1998, among other historical references), radiative transfer has been included only recently in cosmological simulations of the formation of large-scale structure in the universe (see, e.g., Abel et al. 1999; Gnedin & Abel 2001; Ciardi et al. 2001; Razoumov et al. 2002, and more recently Iliev et al. 2006b; Trac & Cen 2007; McQuinn et al. 2007; Baek et al. 2009; Pawlik & Schaye 2008; Maselli et al. 2003; Alvarez et al. 2006; Susa 2006; Altay et al. 2008; Petkova & Springel 2009). Among many different astrophysical problems that require a proper treatment of light propagation, cosmic reionization stands out as a particularly challenging one because the ionizing radiation field plays a key role in the transition from the "dark ages" to the era of galaxy formation: the chronometry and geometry of the process is entirely related to the way matter and radiation interact. The proper numerical modeling of cosmic reionization represents an additional challenge since it requires to capture a whole set of physical phenomena which are difficult to tackle on their own (see, e.g., the review by Barkana & Loeb 2001). In a nutshell, reionization can be described as "atoms being dissociated by UV photons emitted by stars formed in collapsed, self-gravitating halos." This requires to follow the dynamics of dark matter and gas on large scale, cooling and star formation on galactic scales, the emission of ionizing radiation at microscopic scales, and, finally, UV light propagation back to the cosmological scales. Because of this chain of causality involving many different cosmological fluids (dark matter, gas, stars, and photons), it is only recently that significant progress was made in the field of cosmological radiative transfer.

Computer simulations of radiative transfer cover a wide range of techniques, most of them reviewed in Trac & Gnedin (2009) and with most implementations gathered in two sets of comparison papers (Iliev et al. 2006a; Iliev & the Cosmological Radiative Transfer Comparison Project Collaboration 2009). Current cosmological radiative transfer codes successfully pass these rather academic tests, but it should be noted that only a few observational tests can be used as a probe to calibrate these rather complicated numerical tools. The first major constraint comes from quasars with the detection of Gunn–Peterson (GP) troughs and a decrease of the flux transmission in spectra of objects at $z \sim 6$, which can be interpreted as the mark of the transition from a neutral universe to an ionized one (see, e.g.,

Songaila 2004; Fan et al. 2006). From the observed spectra and provided that some assumptions are made on the structure of the density field or the UV background, important quantities such as mean-free path, photoionization rate, or UV field intensity can be constrained (see, e.g., Fan et al. 2002; Fan et al. 2006; Bolton & Haehnelt 2007). These constraints provide anchor values at $z \sim 6$ for the calibration of cosmological simulations of reionization and track their ability to simulate the post-overlap era and the overlap itself (see, e.g., Gnedin & Fan 2006). However, this technique only provides upper/lower boundaries at higher redshifts as complete absorption can be reached with a neutral fraction as low as 0.001. Furthermore, since models are used to infer physical properties from flux transmission, any agreement or disagreement between calculations and quantities derived from observations should be taken with caution (as noted by, e.g., Trac & Cen 2007), and in a reversed role the simulations may happen to be informative about the proper way to interpret data. The second set of constraints comes from the scattering of cosmic microwave background (CMB) photons by electrons released during the reionization process. Usually expressed in terms of the Thomson optical depth τ, current constraints from the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) set $\tau = 0.084 \pm 0.016$ implying a redshift of (instantaneous) reionization of $z \sim 10.9 \pm 1.4$ (Komatsu et al. 2009). This constraint results from the integrated impact of the electrons on the CMB properties and is therefore more sensitive to the complete history of cosmic reionization.

In this paper, our goal is to confront our new radiative transfer code ATON to these observational constraints, using a set of hydrodynamical simulations at different resolutions. This code has already been presented and tested using a standard test suite in Aubert & Teyssier (2008). The dynamical simulations include gravity and gas physics with mesh refinement, as well as widely adopted and well-tested star formation recipe. The radiative transfer is performed as a post-processing step (full coupling of hydrodynamics with radiation is currently underway). It relies on a moment-based description of the propagation of light in the same spirit as, e.g., Gnedin & Abel (2001) or Finlator et al. (2009a). The original ATON code has since been fully ported on Graphics Processing Unit (GPU hereafter) architecture using CUDA. Thanks to the high acceleration rate ($\sim$100$\times$ compared to CPU) made possible by such hardware, we have been able to simulate the radiative transfer at the same resolution as the hydrodynamics base grid with 1024^3 cells. The current article aims at reaching two objectives: first, showing the ability of ATON to properly model the reionization process and second, to demonstrate the potential of GPU architecture for numerical cosmology. Regarding the ability to model the reionization, we partially recover the observational constraints at $z \sim 6$ if we include a simple clumping factor model. However, we also find that the properties of the radiation field and the neutral fraction distribution are driven by very different regions, making it difficult to relate the average UV intensity to the average fraction of neutral gas. Regarding the adaptation of our code on GPU, we describe in detail in the Appendix how such architecture can be used at full power for these type of problems.

This paper is organized as follows. First, we describe the methodology and the simulations. Second we describe a first set of fiducial simulations and assess in particular the issues related to resolution and numerical convergence. Third, we introduce a simple prescription for the subgrid clumping obtained from our most resolved simulation (12.5 Mpc h^{-1} with 1024^3 dark matter particles) and apply it to the largest simulation we have (100 Mpc h^{-1} with 1024^3 dark matter particles). Finally, we discuss our results, forthcoming applications, and possible improvements.

Table 1
Summary of the Parameters Used in Our Simulation Suite

Box Size (Mpc h^{-1})	a_b	k_b (Myr)	$f_{\rm esc}$	$m_{\rm dm}$ ($M_\odot$)	$m_{\rm bar}$ ($M_\odot$)	$m_{\rm star}$ ($M_\odot$)
12.5	0.7	300	0.055	1.52×10^5	2.54×10^4	5.81×10^4
25	1.0	650	0.030	1.22×10^6	2.03×10^5	4.65×10^5
50	1.2	1500	0.020	9.76×10^6	1.62×10^6	3.72×10^6
100	1.2	3000	0.020	7.81×10^7	1.30×10^7	2.97×10^7

Notes. Parameters a_b and k_b are used in the SFR correction to account for finite resolution, assuming *WMAP-5* cosmology. $f_{\rm esc}$ is the assumed escape fraction, $m_{\rm dm}$ is the mass resolution of dark matter particles while $m_{\rm bar}$ is the mass resolution per AMR grid cell. Also shown is the minimum star particle mass. All simulations were performed with 1024^3 dark matter particles.

2. METHODOLOGY

2.1. Simulations

The cosmological simulations analyzed in this work were produced using RAMSES (Teyssier 2002). The cosmological parameters follow the *WMAP-5* constraints (Komatsu et al. 2009), and the initial conditions (ICs) were generated using the MPGrafic package (Prunet et al. 2008). We have generated four sets of Gaussian random fields with different box sizes, based however on the same Poisson shot noise, so that the same structure should form at the same location, although with different timings. The number of cells and dark matter particles was set to 1024^3 and we allow for two more levels of refinement, resulting in the mass and spatial resolution elements quoted in Table 1. The grid was dynamically refined up to the maximum allowed resolution, using a quasi-Lagrangian strategy: when the dark matter or baryons mass in a cell reaches eight times the initial mass resolution, it is split into eight children cells.

Gas dynamics is modeled using a second-order unsplit Godunov scheme (Teyssier 2002; Teyssier et al. 2006; Fromang et al. 2006) based on the HLLC Riemann solver (Toro et al. 1994). We assume a perfect gas Equation of State (EoS) with $\gamma = 5/3$. Gas metallicity is advected as a passive scalar and is self-consistently accounted for in the cooling function. Note that in the present work, no radiation background was considered for the cosmological simulation. As gas cools down and settles into centrifugally supported disks, we need to provide a realistic model for the interstellar medium (ISM). Since the ISM is inherently multiphase and highly turbulent, it is beyond the scope of present-day cosmological simulations to try to simulate it self-consistently. It is customary to rely on subgrid models, providing an effective EoS that captures the basic turbulent and thermal properties of this gas. Models with various degrees of complexity have been proposed in the literature (Yepes et al. 1997; Springel & Hernquist 2003; Schaye & Vecchia 2008). We follow the simple approach based on a temperature floor given by a polytropic EoS for gas,

$$T_{\rm floor} = T_* \left(\frac{n_{\rm H}}{n_*} \right)^{\Gamma - 1}, \quad (1)$$

where $n_* = 0.1$ H/cc is the density threshold that defines the star-forming gas, $T_* = 10^4$ K is a typical temperature mimicking both thermal and turbulent motions in the ISM, and $\Gamma = 5/3$ is the polytropic index controlling the stiffness of the EoS. Gas

is able to heat above this floor, but cannot cool down below it. Note that because of this temperature floor, the Jeans length in our galactic discs is always resolved. We also consider star formation using a similar phenomenological approach. In each cell with gas density larger than n_*, we spawn new star particles at a rate given by

$$\dot{\rho}_* = \epsilon_* \frac{\rho_{\rm gas}}{t_{\rm ff}} \text{ with } t_{\rm ff} = \sqrt{\frac{3\pi}{32G\rho}}, \quad (2)$$

where $t_{\rm ff}$ is the free-fall time of the gaseous component and $\epsilon_* = 0.01$ is the star formation efficiency. The star particle mass depends on the resolution (see Table 1). For each star particle, we assume that 10% of its mass will go supernova after 10 Myr. We consider a supernova energy of 10^{51} erg and one $M_\odot$ of ejected metals per 10 $M_\odot$ average progenitor mass. This supernovae feedback was implemented in the RAMSES code using the "delayed cooling" scheme (Stinson et al. 2006).

To summarize, we used for this simulation suite rather standard galaxy formation recipe, which has proven only recently to be quite successful in reproducing the properties of field spirals (Mayer et al. 2008; Governato et al. 2009, 2010) and dwarf galaxies. The only missing ingredient is the radiation field, which will be considered in a second step using our radiation solver.

2.2. Radiative Transfer

Each snapshot of the simulations is post-processed using the ATON code, described and tested in detail in Aubert & Teyssier (2008), and briefly summarized in this section. The method relies on a momentum description of the radiative transfer equations with an M1 closure relation (González et al. 2007). Radiation is described in terms of the first three moments of the distribution function of photons: the radiative energy density N, the radiative flux $\mathbf{F}$, and the radiative pressure tensor $\mathbf{P}$. These quantities are averaged over a group of frequencies and satisfy the usual conservation relations:

$$\frac{\partial N}{\partial t} + \nabla \mathbf{F} = -\kappa \mathbf{N} + \mathbf{S}, \quad (3)$$

$$\frac{\partial \mathbf{F}}{\partial t} + c^2 \nabla \mathbf{P} = -\kappa \mathbf{F}, \quad (4)$$

where κ stands for the local absorption rate and $S(\mathbf{x}, \mathbf{t})$ is the source field which includes the production sites of photons as well as the recombination radiation. The Eddington tensor $\mathbf{D}$ closes the system through an EoS:

$$\mathbf{P} = \mathbf{D}N, \quad (5)$$

where $\mathbf{D}$ is approximated by the M1 model (Dubroca & Feugeas 1999):

$$\mathbf{D} = \frac{3\chi - 1}{2}\mathbf{I} + \frac{1-\chi}{2}\mathbf{n} \otimes \mathbf{n}. \quad (6)$$

The quantity χ depends only on the reduced radiation flux $f = |\mathbf{F}|/cN$, spans values from 1/3 (pure diffusion regime) to 1 (pure transport regime), and depends only on the local properties of the radiation fields. The exact formula for χ can be found in Aubert & Teyssier (2008). Such a formulation guarantees that the two extreme regimes are properly captured, while all intermediate situations are approximated by a superposition of diffusion and transport. It should also be noted that this scheme differs from the common first-order flux limiter approach by its ability to cast shadows behind absorbents (see Aubert & Teyssier 2008).

The previous radiation conservation laws are solved using an explicit time integration, resulting in a stringent CFL condition on the time step due to the high value of the speed of light:

$$\frac{\Delta x}{c} > \Delta t. \quad (7)$$

However, thanks to GPU acceleration, we can speed up each individual time step so that the resulting scheme can still achieve high performance. The details of the GPU implementation are given in the Appendix. Originally the code is able to evaluate intercell fluxes using both the Haardt–Lax–van Leer and the simpler Lax–Friedrich scheme, but only the latter has been used in the current work. Let us emphasize that this choice of an explicit scheme is mainly driven by its simplicity rather than any kind of numerical/accuracy advantage over implicit methods: parallelization is much easier to implement in the former scheme whereas an implicit solver involves sparse matrix solvers which are difficult to optimize on GPU architectures.

The photochemistry in ATON is currently limited to hydrogen with the associated cooling processes. Again, the energy conservation and the chemistry are solved in an explicit fashion and are sub-cycled during a radiative transfer step using a scheme in the spirit of Anninos et al. (1997). It turns out that most of the time the characteristic time scales involved in cooling and chemistry are longer than radiative time steps, thus limiting the impact of "microphysics" calculations on the overall computation.

All the processes (transport, cooling, and chemistry) and their equations are solved in a single frequency group where $\nu > 13.6$ eV and involve average quantities such as the hydrogen photoionization cross sections $\sigma_e = 2.49 \times 10^{-18}$ cm^2 (energy averaged), $\sigma_n = 2.93 \times 10^{-18}$cm^2 (number averaged), and the typical ionizing photon energy $e = 20.27$ eV, where a 50,000 K blackbody spectrum is assumed.

Typically, one complete radiative transfer simulation requires between 30,000 and 240,000 time steps depending on the resolution which defines the time step and the starting redshift. The 800 Myr of cosmic evolution we would like to cover is described with a time resolution of 3500 yr for radiative transfer. Typically 45 post-processed outputs were produced by the radiative transfer solver for the subsequent analysis, while 125 snapshots were provided by RAMSES to describe the evolution of the gas and the sources. The code has been deployed on GPU architecture using the API CUDA 2.2, (becoming thus CUDATON) developed for devices built by the Nvidia company. The code runs independently from the CPU, without any transfer between the host and the device except during the initial setup and for the outputs on hard drives. The typical acceleration observed compared to single-CPU runs is close to 100. Using an additional message passing interface (MPI) layer, CUDATON is able to run on multi-GPU architecture with communications between the devices, which requires additional transfer between hosts and GPUs. The additional cost is close to 10% of the total computing time since data have to be transferred through PCI-Express ports. All the calculations here were performed on 128 Tesla C1060 devices on the Titane supercomputer of the CCRT computing center. Typically a single radiative post-processing run on a 1024^3 grid is performed in 2.5 hr but can be as short as 1 hr for coarse simulations with simple physics and as long as 18 hr for our most realistic calculations. During the course of this project, a couple hundred of calculations over 6 months

were performed to improve the code and to test our various recipes.

2.3. Source Modeling for the Radiative Post-processing

The sources of photons, namely, young stars, are produced by the cosmological simulations that return for each stellar particle its position, velocity, age, mass, and metallicity. From there, the source modeling is inspired by the procedure described in Baek et al. (2009). Stellar particles are assumed to satisfy a Salpeter initial mass function resulting in a global spectra well approximated by a 50,000 K blackbody. Individual lifetimes of stellar particles as ionizing sources are drawn randomly between 5 and 20 Myr. Overall, for each source, the production of ionizing photons lies between 24,000 and 98,000 per stellar baryon over its lifetime. Because the sources appear at discrete times, due to the discontinuous production of snapshots, the sources' contribution is smoothed out over all the duration between two successive snapshots using the following strategy. When modeling the radiative transfer from time t_p to t_{p+1}, we consider only star particles contained in snapshot $p+1$. Knowing their age a_{p+1}, we calculate their age $a_p = a_{p+1} - (t_{p+1} - t_p)$ at time step p, which can be negative if the star appeared at a time a^* between the two snapshots. Then:

1. If a_p is greater than the source's lifetime L, it is discarded.
2. If $a_p < 0$, the source has been created between the two snapshots. However, it will contribute to the photon emission from t_p to t_{p+1} with a "diluted photon" emission rate given by $(a_{p+1} - a^*)/(t_{p+1} - t_p)\dot{N}$.
3. If $0 < a_p < L$ the source will have an emission rate given by $\min(1, L - a_p)/(t_{p+1} - t_p)\dot{N}$. If the source ends its ionizing phase between the snapshots, it will nevertheless contribute continuously from t_p to t_{p+1} with a "diluted" emission rate.

From there, sources are projected on the three-dimensional grid using the Nearest Grid Point assignment scheme. Emission is modeled as a field where each cell acts as a single photon source.

These intrinsic luminosities are modulated by two additional factors to give the effective source luminosity:

$$\dot{N}_{\rm eff} = \dot{N} \times f_{\rm esc} \times B. \quad (8)$$

The first factor is the escape fraction $f_{\rm esc}$ which models the actual fraction of radiative energy which manages to escape the stellar environment. Typical values can be as high as 20% and is essentially a free parameter which allows to tune the reionization redshift. The second factor, B, is called here the *boost factor*. It is a correction term that compensates from the unresolved star-forming halos in the simulation, a major resolution effect on the simulated star formation history (SFH). As shown in, e.g., Rasera & Teyssier (2006), the mass resolution has a significant impact on the SFH if large simulation volumes are considered, when the minimum resolved halo mass (optimistically set to 100 dark matter particles) is larger than the minimum mass for star-forming halos, based on atomic cooling arguments (Gnedin 2000; Rasera & Teyssier 2006; Hoeft et al. 2006). This minimum mass (also referred to as the Filtering mass) starts around $10^7\ M_\odot$ before reionization and then rises steadily as $(1 + z)^{3/2}$ from redshift 6 to the final epoch (Rasera & Teyssier 2006; Hoeft et al. 2006). Resolving this minimum mass before reionization will require a dark matter particle mass below $10^5\ M_\odot$, a rather strong requirement for cosmological simulation. Only our smallest box size ($12.5\,{\rm Mpc}\,h^{-1}$ with 1024^3 particles) barely satisfies this criterion.

As an illustration, the top panel of Figure 1 shows the evolution of the integrated number of photons with time in four simulations at different resolutions, which depends directly on the simulated star formation rate (SFR). Clearly the difference in resolution has an impact on the apparition of the first sources: low-resolution simulations require a longer time to reach the epoch of the formation of the first stars: $z \sim 11$ for the $100\,{\rm Mpc}\,h^{-1}$ simulation versus $z \sim 18$ for the $12.5\,{\rm Mpc}\,h^{-1}$ simulation. Furthermore, this late start is not compensated for by a higher SFR and at $z = 6$, the number of emitted photons decreases as lower spatial resolution is considered.

We used the analytical model of Rasera & Teyssier (2006) to compute the expected converged SFH. We can compensate for the unresolved star-forming halos by boosting each resolved UV emitting source by a "boost factor," derived to put the actual simulated SFRs (and hence the number of emitted photons) in accordance with the converged one. We have used for the boost factor the following simple functional form:

$$B(t) = \min(1, a_b \exp(k_b/t)) = \frac{{\rm SFR}_{\rm converged}(t)}{{\rm SFR}_{\rm actual}(t)}, \quad (9)$$

where t is the age of the universe. The parameters a_b and k_b are fitted in the measured SFH in each simulation. They hence depend on the resolution and are given in Table 1 for $1024^3 + 2$ levels of refinement simulations with the *WMAP-5* cosmology. The resulting integrated photon numbers are shown in the bottom panel of Figure 1 and exhibit a good level of convergence at redshifts $z < 9$. Let us emphasize that the two parameters, $f_{\rm esc}$ and B, are different by nature: B is not a free parameter and follows from the proper analysis of resolution effect and is in some sense a pure numerical correction. Meanwhile, $f_{\rm esc}$ remains as a physical parameter which models, e.g., the subgrid physics and in the end, serves mostly to set the redshift of reionization. Of course, this simple prescription does not fix the late apparition of stars at low resolution since it only corrects existing stars without creating new sites of stellar formation. In our investigations, it appears that low-resolution simulations (with the largest correction) exhibit similar behaviors than highly resolved ones, but we admittedly focus on average quantities and global distributions. Fine geometrical details, on the other hand, are likely to be poorly captured by this boost factor approach because of the lack of small emitters in unresolved site of stellar formation.

3. RESULTS

Several aspects were investigated during this work, starting from basic numerical experiments focusing mostly on the resolution effects to slightly more complex modelization where we attempt to fit observational data. First, we describe our fiducial results obtained from the post-processing of adaptive mesh refinement (AMR) simulations. From there, we discuss the impact of subcell clustering to the modelization with a focus on the quantity of absorbents at $z \sim 6$.

3.1. Fiducial Experiments

3.1.1. Global Properties

The fiducial experiments consist of four simulations described in Table 1, with comoving box size ranging from $100\,{\rm Mpc}\,h^{-1}$ to $12.5\,{\rm Mpc}\,h^{-1}$. The dynamical simulations were performed

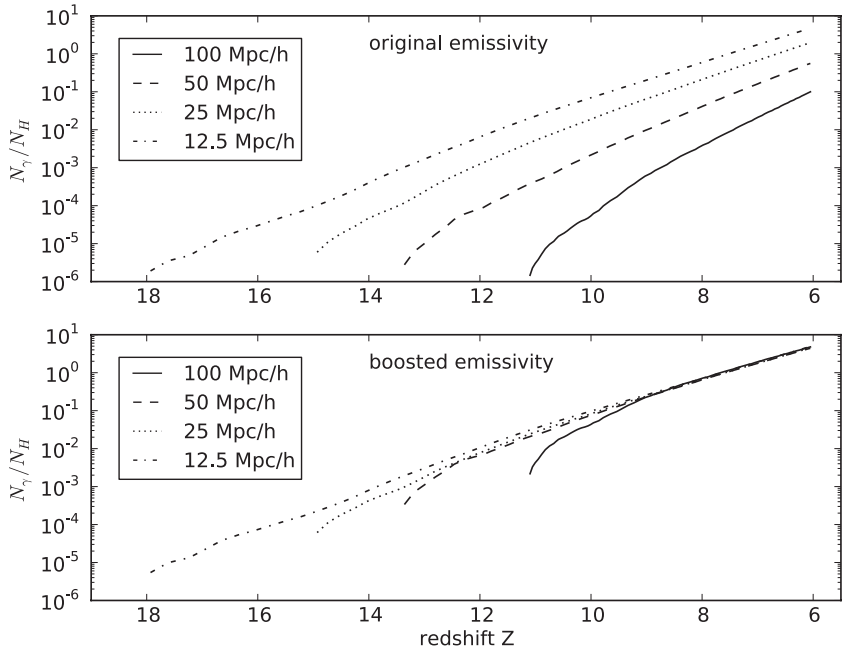

Figure 1. Integrated number of emitted photons in units of the number of hydrogen atoms as a function of redshift in four (1024^3+2 levels of refinement) hydrodynamical simulations with different coarse resolutions. The top panel shows the original emissivity due to the simulated stellar population, while the bottom panel shows the boosted emissivity which compensates for the impact of resolution on the simulated SFR. All plots are performed with $f_{\rm esc} = 0.03$.

on a 1024^3 coarse grid + 2 level of refinement at $z \sim 6$ while the radiative transfer post-processing was computed on a 1024^3 regular grid. Highly resolved simulations are expected to better resolve the small-scale photon sinks but lack the strong and rare sources that populate large scale volumes. Conversely, large simulations have a better representativity of rare and strong events but lack the resolution on absorbents. These features are somehow reflected in the values of the escape fraction shown in Table 1: for a redshift of reionization chosen to be $z_{\rm ion} = 6.5$, $f_{\rm esc}$ decreases with the box size from 0.055 to 0.02. Highly resolved simulations have sources embedded in highly clustered gas, implying a more efficient recombination, and these sources cannot be as strong as the ones found in large volumes. Overall, such simulations require a larger amount of photons to reionize.

Maps of the distribution of neutral gas are shown in Figure 2 at half reionization. Let us recall that these four simulations were performed with ICs that shared the same set of phases leading to similarities in the global spatial distributions. These maps exhibit the expected global behaviors: high resolution simulations present complex ionization fronts, which result from the highly inhomogeneous structure of the absorbing regions. Meanwhile, low-resolution simulations fail to resolve small-scale structures leading to smoother fronts. Looking at the details of zoomed maps (in Figure 3), highly resolved simulations present dense neutral clumps within ionized regions whereas these absorbents are absent from large under-resolved boxes. The failure of large simulation boxes to resolve these small scales will prove to be crucial in our ability to reproduce the data at $z \sim 6$.

The typical temperature of the intergalactic medium after the post-processing step is $T \sim (1\text{–}1.5) \times 10^4$ K which is the value expected. However, this temperature is achieved only after our post-processing and the dynamical simulation performed by RAMSES does not include any photo-heating. It has limited consequence prior to reionization, since by definition the UV intensity is low but can have an impact at $z \sim 6$. At these times, the probability density function (pdf) of the density is not exactly consistent with the one we expect when the gas is heated by, e.g., a uniform background. For comparison, we included the pdf of the gas density of our four boxes and compared it with the model suggested by Miralda-Escudé et al. (2000), taken at $z = 6$ (see Figure 4). At high densities, the behavior is correct and convergence is achieved for the smallest box. At low density, however, we detect voids that are not present in the Miralda-Escudé et al. (2000) model and they are due to the lack of photoionization effects on the gas dynamics. We tend to think that the impact on our result is limited since the distributions are in agreement for a contrast $\Delta > 0.1$ which include the majority of the gas and the most probable value. Still, a definite answer can only be given by fully coupled simulations.

3.1.2. Neutral Fraction

The evolution of the volume-weighted ionized fraction x_v and neutral fraction $1 - x_v$ is shown in Figure 5. Escape fractions were chosen to achieve reionization at $z \sim 6.5$ and it can be seen that all four calculations present similar behavior for $z < 9$. Distribution of values is shown as colored contours in Figure 6 and it can be seen that x_v is representative of the distribution of ionized fraction in the boxes as it tracks accurately the most probable value. For earlier times, notable differences arise from the impact of resolution on star formation and the production of photons. Highly resolved simulations have ionization history that expand up to $z = 18$ where their first stars form. Conversely the largest simulation forms stars only at $z = 11$. Because of the introduction of a time-dependent boost of their source luminosity, these large simulations quickly catch up the highly resolved one, resulting in a strong slope for x_v. The "catching up" effect is clearly noticeable in the 100 Mpc h^{-1} calculation but already much limited for the 50 Mpc h^{-1} simulation, almost unnoticeable for the 25 Mpc h^{-1} box and for $z > 9$ the calculations have all converged. This

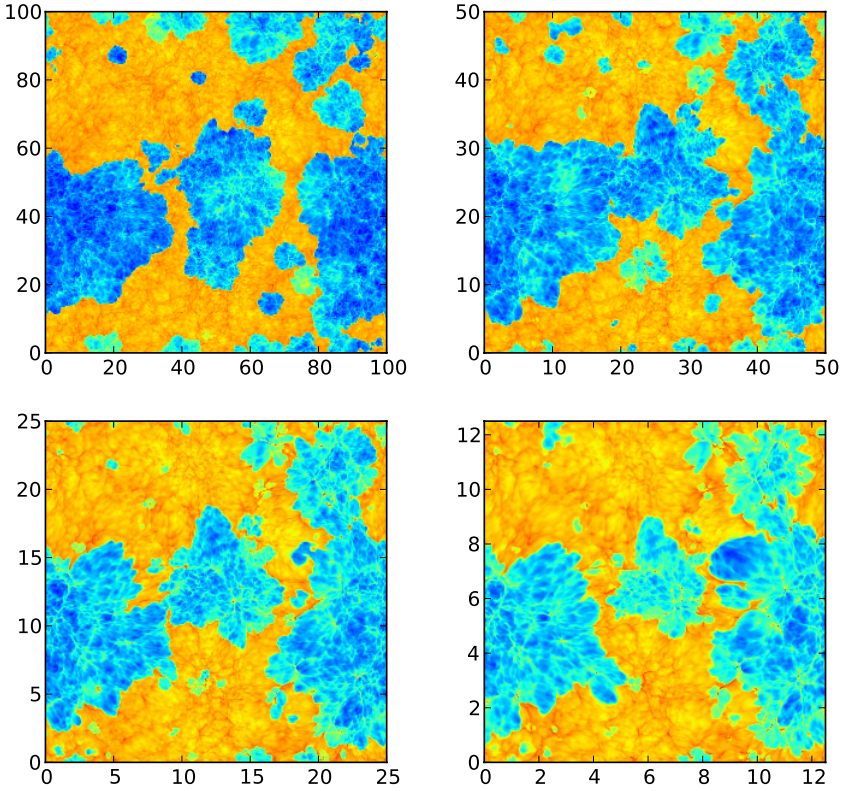

Figure 2. Neutral hydrogen density maps at $z \sim 7.3$ and $x \sim 0.5$ (volume-weighted) for boxes of comoving lengths 100, 50, 25, and 12.5 Mpc h^{-1}. All maps have a resolution of 1024^2 and a thickness of 5 Mpc h^{-1}. The color scale is logarithmic with blue and red regions standing, respectively, for low and high densities of neutral hydrogen. Coordinates are expressed in comoving Mpc h^{-1}.

(A color version of this figure is available in the online journal.)

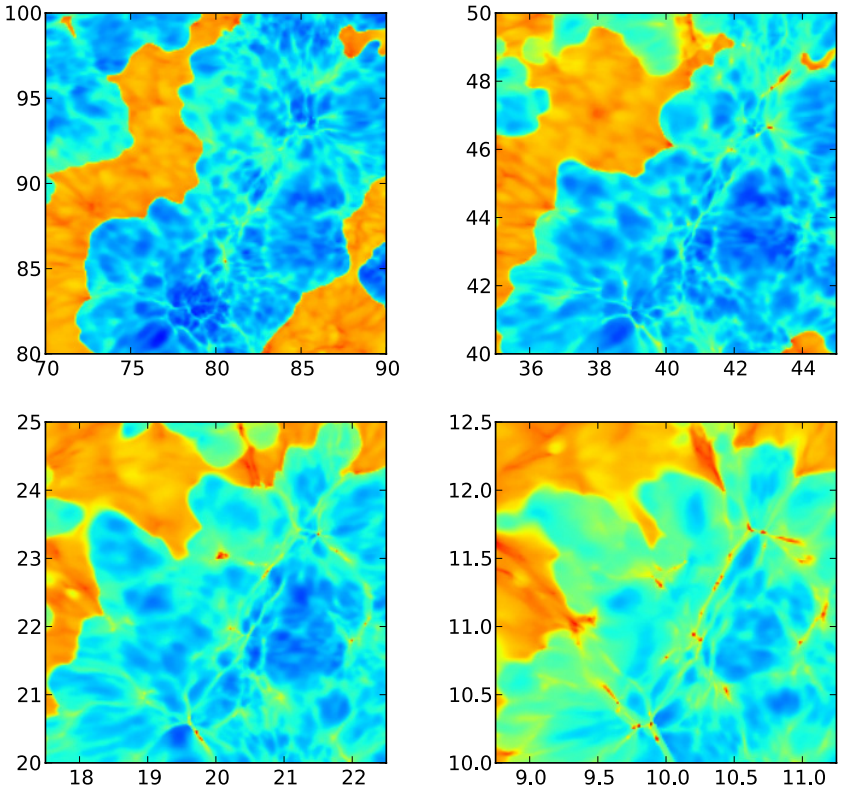

Figure 3. Same as Figure 2 but zooming on a photoionized region. Coordinates are expressed in comoving Mpc h^{-1}.

(A color version of this figure is available in the online journal.)

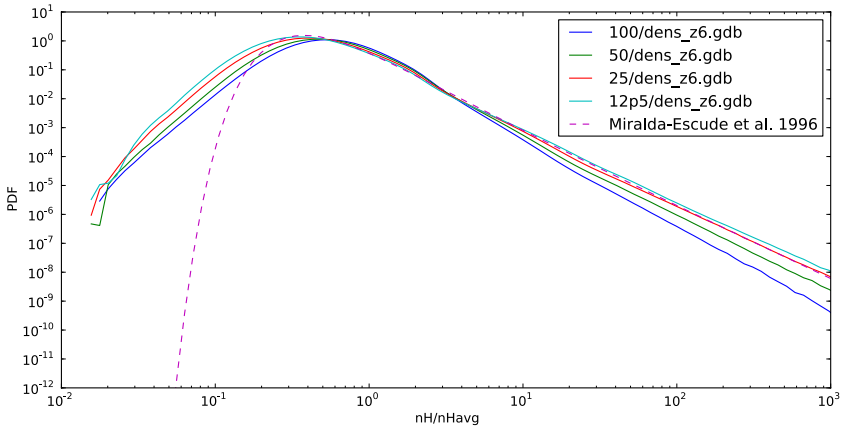

Figure 4. Probability density function of the baryon density at $z = 6$ in the different boxes compared to the model given by Miralda-Escudé et al. (2000).
(A color version of this figure is available in the online journal.)

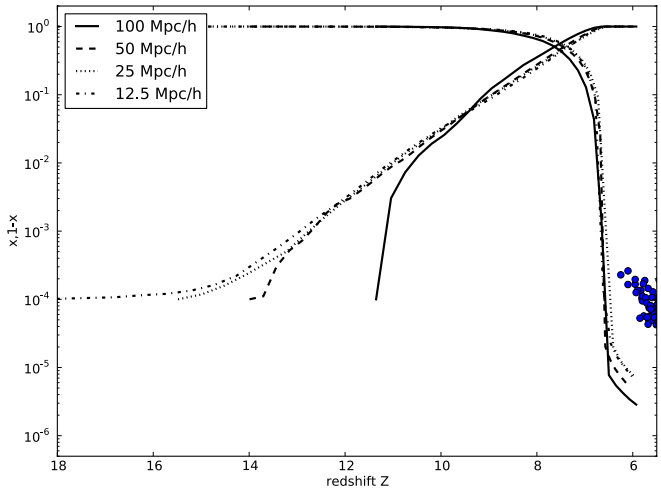

Figure 5. Neutral and ionized volume-averaged fraction as a function of redshift measured in the 100, 50, 25, and 12.5 Mpc h^{-1} boxes. Dots stand for observational constraints given by Fan et al. (2006).
(A color version of this figure is available in the online journal.)

should not come as a surprise since the boost factor approach was designed precisely for that reason. Nevertheless, the different clustering of gas, of sources as well as their number could have resulted in a difference in the ionization history of the different calculations even though they share the same global amount of photons emitted. Since we do not observe such a discrepancy, it suggests that small photon sources missing from the large boxes are located roughly at the location of the large photon sources present in these boxes: by boosting their luminosity, we compensate at the subgrid level for the lack of stellar particles at the correct location.

Let us also point out that the neutral fraction calculated at $z = 6$ spans from 3×10^{-6} for the 100 Mpc h^{-1} simulation to 10^{-5} for the 12.5 Mpc h^{-1}. Such levels of neutral fraction are inconsistent with constraints provided by Fan et al. (2006) from GP troughs in quasars spectra which imply a typical level of 10^{-4} at $z = 6$. Even though this estimation relies on assumption on the distribution of gas and on a homogeneous field of radiation, this level of neutral gas has been reproduced by, e.g., Trac & Cen (2007), Kohler et al. (2007), Shin et al. (2008) and on highly resolved simulations by Gnedin & Fan (2006). On the other hand, Finlator et al. (2009b) present the same level of discrepancy at admittedly much lower resolution. We investigate this point further on in subsequent sections, but at the current stage, all our simulations fail to reproduce the observed neutral fraction without additional modelization.

3.1.3. Radiation Field-UV Intensity

The moment-based description of radiative transfer allows us to track the radiation intensity in each cell, usually described in terms of J_{21} or in terms of photoionization rate Γ_{12}. The evolution of the volume-averaged intensity is shown in Figure 7 as well as the constraint provided by Bolton & Haehnelt (2007). For our four boxes, the evolution of the radiation intensity exhibits a "cobra-like" shape with a sharp increase until the reionization epoch, during which the increase is the steepest, followed immediately after by a flattening of the slope at the end of reionization. The amount of radiation is larger by a factor of three at $z = 6$ when comparing the 100 Mpc h^{-1} and the 12.5 Mpc h^{-1} box, while the 25 Mpc h^{-1} and 12.5 Mpc h^{-1} calculations seem to have converged. This trend is consistent with the differences in the residual neutral fraction calculated at $z = 6$ (see Figure 5), where the highly resolved calculations present a larger amount of neutral gas than the poorly resolved ones. A stronger radiation field implies a larger photoionization rate and therefore a smaller amount of neutrals.

Unfortunately, these calculations all agree on one point: they are inconsistent with observational constraints, such as the one provided by Bolton & Haehnelt (2007) by a factor of 20–50. Again, such a discrepancy has recently been found at lower resolution by Finlator et al. (2009b) using an alternative implementation of moment-based radiative transfer. This excess of radiation goes along with the lack of neutral gas at the end of reionization in our computations. Furthermore, an inspection of Figure 8 reveals that the average intensity is representative of the most probable value in the boxes, discarding any possibility of a biased value.

3.1.4. Optical Depth

The Thomson scattering of CMB photons by the electrons released during the reionization is quantified through the Thomson optical depth given by

$$\tau = c\sigma_t \int_{z_{\rm rec}}^{0} n_e(z) \frac{dt}{dz} dz, \qquad (10)$$

where σ_t is the corresponding cross section and $n_e = x n_{\rm H}$ is the density of electrons released by ionized hydrogen atoms.

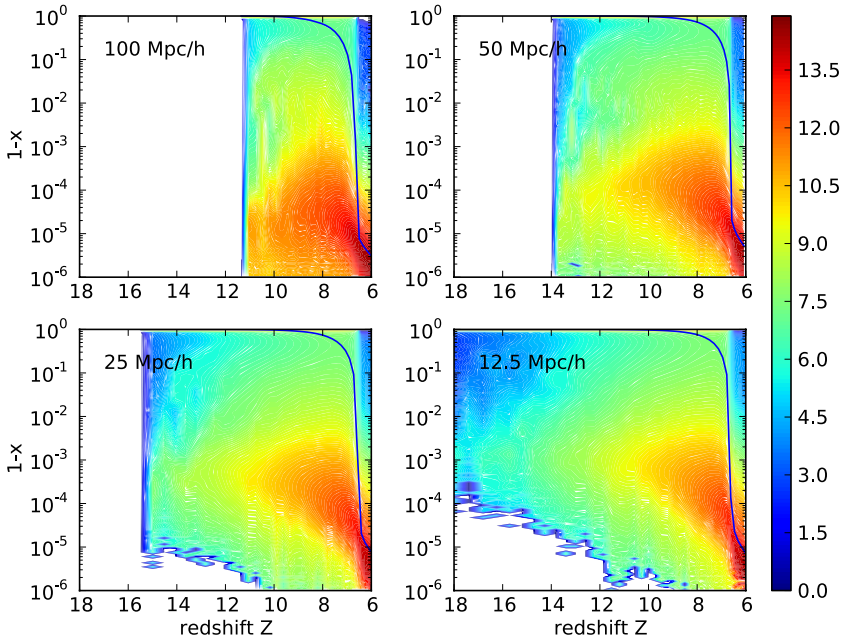

Figure 6. Evolution of the hydrogen neutral fraction $f_{\rm HI}$ in the 100, 50, 25, and 12.5 Mpc h^{-1} boxes (from top to bottom). The blue lines stand for the evolution of the volume-weighted average value and the color levels show the evolution of the $f_{\rm HI}$ distribution with redshift. Values with a high probability density are shown in red and values with a low probability density are shown in blue, the scale being logarithmic.

(A color version of this figure is available in the online journal.)

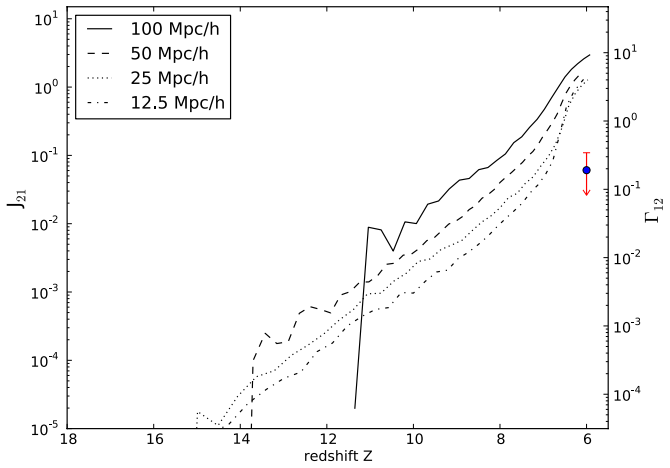

Figure 7. Evolution of the mean intensity of ionizing radiation in the 100, 50, 25, and 12.5 Mpc h^{-1} boxes. The upper limit at $z \sim 6$ stands for the constraint given by Bolton & Haehnelt (2007).

(A color version of this figure is available in the online journal.)

Our calculations of the optical depth are shown in Figure 9 for the four simulations. Also presented is the constraints range obtained from the five years release of CMB measurements made by the *WMAP* collaboration (Komatsu et al. 2009) at the 1σ level.

The four fiducial experiments were performed at different resolution and therefore exhibit different ionization histories but have converged in terms of optical depth. This agreement results from the fact that the bulk of electrons production lies within the convergence redshift range ($z < 11$). The four of them present an optical depth $\tau = 0.051$ which lies at 2σ from the CMB value. The inclusion of helium electrons would slightly increase the amount of electrons (typically 10% see, e.g., Finlator et al. 2009b) but not at levels that would make it consistent with the expected *WMAP* value. This discrepancy has already been noted by Gnedin & Fan (2006), Trac & Cen (2007), or Finlator et al. (2009b) for simulations with similar ionization redshifts. Also shown in Figure 9 is the optical depth measured in the largest box (100 Mpc h^{-1}), but with different escape fractions. These calculations were performed at the same level of resolution as the fiducial experiments and under the same protocol. Clearly, the resulting τ gets closer to the CMB value as a consequence of the larger density of electrons due to ionizing photons at earlier times. It indicates that a larger escape fraction could be the solution toward an agreement, however it comes at the cost of a larger redshift of reionization as shown in Figure 10 which would place the $z = 6$ neutral fraction even further to the observed constraints than the fiducial ones. It is therefore likely that a plausible path toward an agreement between the observed and the calculated τ lies in a varying escape fraction which would ensure an extended ionization history to increase τ while keeping the reionization redshift reasonably low. For instance, investigations by Wise & Cen (2009) suggest that higher escape fractions could exist in small galaxies. Furthermore Iliev et al. (2007) argue that early populations in small halos at $z \sim 22$ are important contributors to the optical depth, and these objects are missing in our calculations. Other routes toward agreement may lie in additional physics such as the inclusion of specific population III sources (Trac & Cen 2007). Finally, it should be noted that Shull & Venkatesan (2008) suggest that inaccuracies in the reionization history and degeneracies in cosmological parameters lead to a larger range of possible values of τ, 0.06–0.11 at 1σ: the discrepancy between our calculation and the CMB constraint could be resolved in between at $\tau \sim 0.07$.

3.1.5. Density Dependence of UV Intensity and Neutral Fraction

In order to investigate our results on the *average* neutral fraction and UV intensity, we have computed the distribution $(1 - x)(n_{\rm H})$ and $J_{21}(n_{\rm H})$ just after the reionization at $z = 6.25$. The results are shown in Figures 11 and 12. From Figure 11, we can see that the radiation field is not strictly homogeneous:

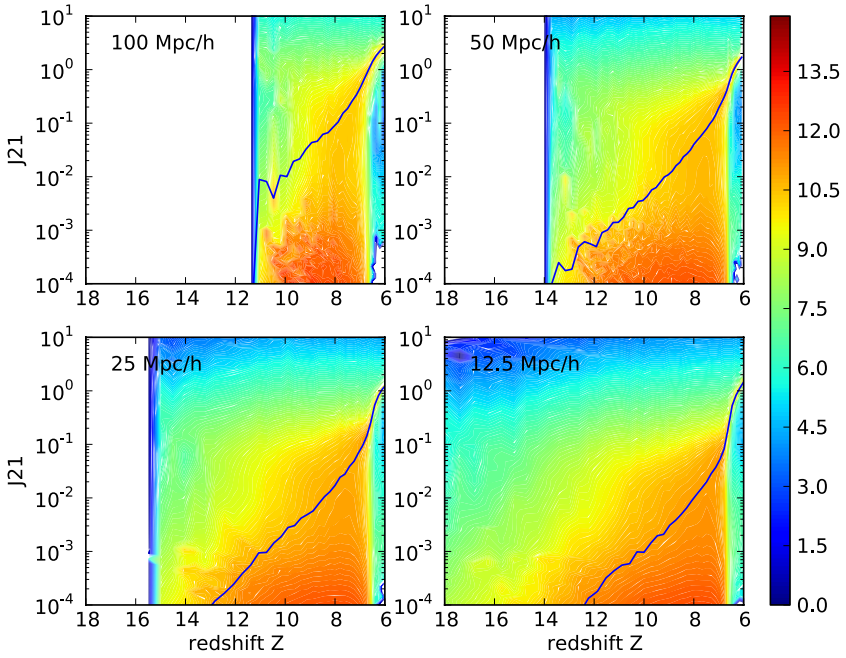

Figure 8. Evolution of the mean intensity J_{21} in the 100, 50, 25, and 12.5 Mpc h^{-1} boxes (from top to bottom). The blue lines stand for the evolution of the volume-weighted average value and the color levels show the evolution of the J_{21} distribution with redshift. Values with a high probability density are shown in red and values with a low probability density are shown in blue, the scale being logarithmic.

(A color version of this figure is available in the online journal.)

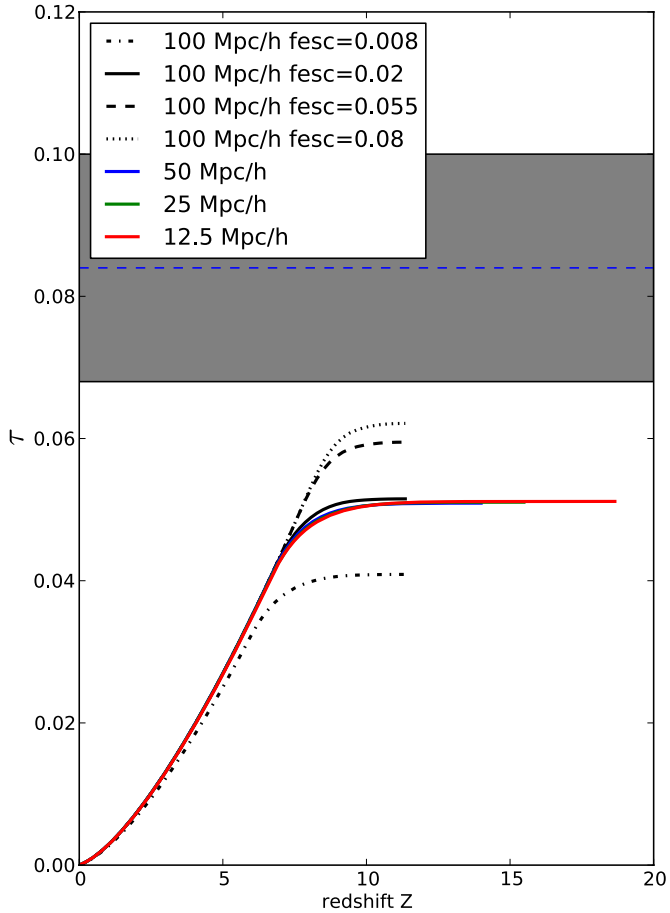

Figure 9. Thomson optical depth computed from the average mass-weighted electron density. The gray area stands for the *WMAP-5* measurements allowed range at the 1σ level (see Komatsu et al. 2009). The 50, 25, and 12.5 Mpc h^{-1} curves are almost superimposed.

(A color version of this figure is available in the online journal.)

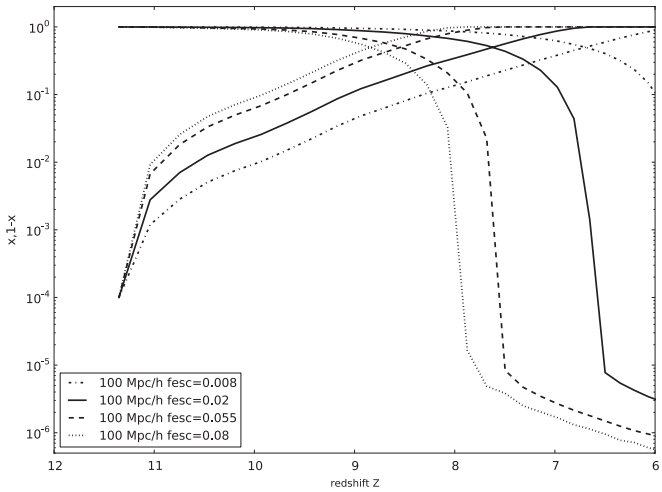

Figure 10. Evolution of the volume-weighted neutral and ionized hydrogen fraction in comoving 100 Mpc h^{-1}–1024^3 boxes with different escape fractions.

although it is quasi-constant for densities $n_{\rm H} < 0.001$ cm^{-3} ($\Delta < 10$), we see a significant decrease of the flux in the densest regions. In particular, an accumulation of obscured regions is apparent at densities close to 5×10^{-2} cm^{-3} ($\Delta \sim 250$) with a radiation field 1000 times weaker than the volume average. This feature is more prominent in highly resolved simulations (12.5 and 25 Mpc h^{-1}) and corresponds to a better treatment of dense, small absorbents, which we fail to model properly in large boxes. The strong cutoff of radiation in high-density regions is a manifestation of self-shielding, where dense clumps are protected from the ionizing background by their own high densities.

We model this high-density behavior using two different models with different levels of exponential cutoff:

$$J_1(n_{\rm H}) = J_0 \exp(-n_{\rm H}/n_1^*) \quad (11)$$

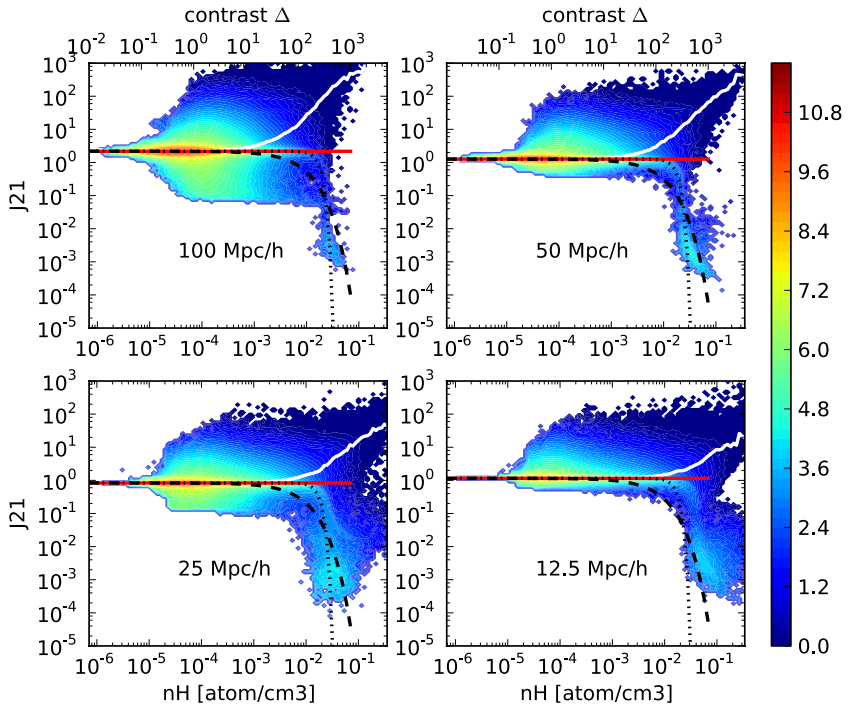

Figure 11. Density contrast vs. ionizing intensity (as J_{21}) relations measured in the 100, 50, 25, and 12.5 Mpc h^{-1} boxes at $z \sim 6.25$, which corresponds to a fully ionized simulation. Red (resp. blue) regions stand for high (resp. low) probability densities in the distributions, the scale being logarithmic. The white line stands for the average ionizing intensity per density bin $\langle J_{21}\rangle(nH)$. The dashed red line stands for the volume-averaged $J_{21} = J_0$ value, the dashed black line stands for the $J_{21} = J_0 \exp(-n_{\rm H}/n_1^*)$ model, and the dotted line stands for the $J_{21} = J_0 \exp(-(n_{\rm H}/n_4^*)^4)$ model.

(A color version of this figure is available in the online journal.)

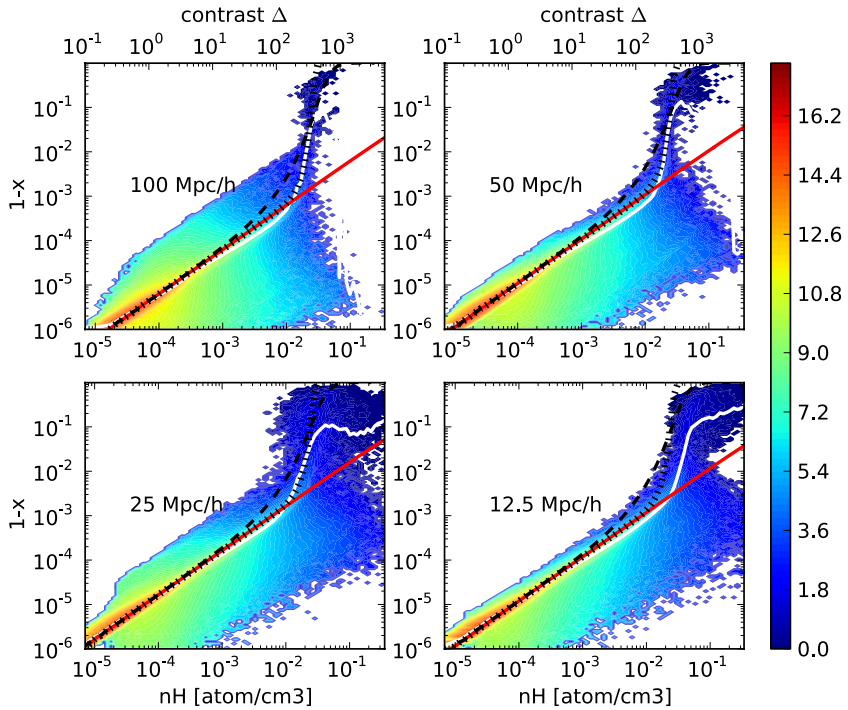

Figure 12. Density contrast vs. neutral fraction relations measured in the 100, 50, 25, and 12.5 Mpc h^{-1} boxes at $z = 6.25$, which corresponds to a fully ionized simulation. Red (resp. blue) regions stand for high (resp. low) probability densities in the distributions, the scale being logarithmic. The white line stands for the average neutral fraction per density bin $\langle 1 - x\rangle(n_{\rm H})$. The red line stands for the expected neutral fraction assuming equilibrium and an ionizing intensity equal to J_0. The dashed and dotted lines stand, respectively, for the expected neutral fraction assuming equilibrium for the $J_{21} = J_0 \exp(-n_{\rm H}/n_1^*)$ and $J_{21} = J_0 \exp(-(n_{\rm H}/n_4^*)^4)$ models.

(A color version of this figure is available in the online journal.)

and

$$J_4(n_{\rm H}) = J_0 \exp((-n_{\rm H}/n_4^*)^4), \quad (12)$$

where J_0 stands for the average intensity field at low density and $n_{1,4}^*$ is the characteristic density at which the exponential cutoff operates. For the sake of simplicity, we arbitrarily assigned the same values for the characteristic self-shielding densities for the four simulations, namely, $n_1^* = 0.007$ cm^{-3} ($\Delta \sim 100$) and $n_4^* = 0.018$ cm^{-3} ($\Delta \sim 250$), which accurately reproduce the global $J_{21}(n_{\rm H})$ behaviors in the three largest boxes and are slightly off for the 12.5 Mpc h^{-1} calculation. On the other hand, the plateau value is computed separately for the four box sizes. Clearly, the J_4 model is a better representation of the density dependence but the other model J_1 has also been considered for the sake of comparison. It should be noted that the volume-average value $\langle J_{21}\rangle(n_{\rm H})$ per density bin is also shown in Figure 11 as the white solid line. Surprisingly, it increases to levels up to 100 times the volume average value, instead of following the exponential cutoff. While the latter is a good model for the most probable radiation flux, it does not predict the correct volume-averaged value. The discrepancy starts at densities close to 10^{-3} cm^{-3} and is likely to be due to strong sources of radiation which are found inside galaxies. This illustrates the fact that the radiative intensity may be subject to strong biasing effects, especially at high densities, and its average value must be therefore taken with caution.

These models for the UV radiation field can be used to compute the expected neutral fraction, assuming photoionization equilibrium. The result of such a procedure is shown in Figure 12 for the four fiducial simulations with the equilibrium $(1-x)(n_{\rm H})$ curves computed for a uniform radiation field equal to J_0 and for the small/strong cutoff models J_1 and J_4. Also shown is the average neutral fraction per density bin. Clearly the average neutral fraction follows the equilibrium trend at low densities ($n_{\rm H} < 0.001$ cm^{-3}, $\Delta < 15$) which is reproduced well by the three models. For higher densities ($n_{\rm H} > 0.01$ cm^{-3}, $\Delta > 150$), the average neutral fraction rises sharply and its distribution presents a significant tail toward neutral gas, even though the spread remains quite important: for instance at $n_{\rm H} \sim 2\times10^{-2}$ neutral fractions from 10^{-6} to 1 can be found with almost equal probabilities. This tail cannot be reproduced by the uniform UV field model J_0 which is not surprising since J_0 is not representative of the radiation field in which self-shielded high-density regions lie. The small cutoff model J_1 does a better job at reproducing the tail but underestimates the strength of rise toward larger neutral fractions. The strong cutoff model J_4 is in better agreement which is also expected since it models more accurately the typical trend of the UV field as a function of density (see Figure 11). This overall agreement between the computed radiation field and neutral fractions indicates that the global neutral fraction can be recovered assuming photoionization equilibrium, as long as the correct model for self-shielded UV radiation is used[4].

3.2. Subgrid Clumping Model

Our fiducial numerical experiments, albeit highly resolved in terms of radiative transfer, lack some resolution for the underlying gas distribution. From Table 1 the largest simulation fails to resolve star-forming halos at $z > 11$ and $\sim 10^7\,M_\odot$ mini-halos which are expected to act as a sink of photons during the reionization process. On the other end of our sample of simulation, the smallest boxes reasonably resolve these scales but are too small to, e.g., provide a correct description of the cosmological H II regions which are expected to be as large as tens of Mpc (see, e.g., Barkana & Loeb 2001; Furlanetto et al. 2004; McQuinn et al. 2007). From now on, we focus on the largest simulation (100 Mpc h^{-1} with 1024^3 dark matter particles) but with an additional subgrid model, in order to combine large-scale statistics with a corrected small-scale physical model. From now on, we only compare our calculations to the constraints at $z = 6$ from quasars spectra and put Thomson optical depth measurements aside. Since this quantity is more sensitive to the global reionization history, the fact that our SFH starts at $z = 11$ in the largest box cannot be compensated for by any other means than just increasing the mass resolution or drastically changing the star formation recipe. Exploring these possibilities is postponed to future work.

3.2.1. Clumping Factors

Considering the hydrogen chemical balance equation, one gets

$$\frac{dn_{\rm H_I}}{dt} = \alpha n_e n_{\rm H_{II}} - \Gamma n_{\rm H_I}, \quad (13)$$

which is modified in the following manner if one considers the ionization fraction x:

$$-n_{\rm H}\frac{dx}{dt} = \alpha n_{\rm H}^2 x^2 - (1-x) n_{\rm H}\Gamma, \quad (14)$$

where $n_{\rm H}$ stands for the hydrogen number density (neutral+ionized), α and Γ are, respectively, the recombination and photoionization coefficients, and x is the usual ionized fraction. As we deal with fields defined on a grid, we only have access to quantities averaged within the cells such as $\langle n_{\rm H}\rangle$ which lack some information on the subgrid variations. Defining a recombination clumping factor as $C_R = \langle n_{\rm H}^2 x^2\rangle / \langle n_{\rm H} x\rangle^2$ and a photon-atomic density cross clumping factor $C_I = \langle n_\gamma (1-x) n_h\rangle / \langle n_\gamma\rangle\langle (1-x) n_{\rm H}\rangle$, the chemistry equation can be rewritten as

$$\frac{dx}{dt} = -(\alpha)\langle n_{\rm H}\rangle^2 x^2 C_R - (1-x) c\sigma \langle n_{\rm H}\rangle\langle n_\gamma\rangle C_I. \quad (15)$$

The choice of definition for a clumping factor is not unique; for instance, Kohler et al. (2007) define clumping factors where the averages are taken over the whole terms such as $\langle \alpha(T) n_{\rm H}^2 x^2\rangle$, which depends on density, ionization fraction, and also temperature. Our choice of clumping factors basically assumes that temperature is distributed uniformly within the computational cells.

We compute the clumping factors for the 100 Mpc h^{-1} simulation using the 12.5 Mpc h^{-1} simulation. For $6 < z < 18$, the total Jeans Mass in the case of an adiabatically cooling gas decreases from $1.5 \times 10^4\,M_\odot$ to $4 \times 10^3\,M_\odot$ while the baryonic filtering mass evolves from $10^5\,M_\odot$ to $5 \times 10^4\,M_\odot$ during the same interval (see, e.g., Gnedin & Hui 1998; Barkana & Loeb 2001; McQuinn et al. 2007). From Table 1, one can see that our mostly resolved simulation almost achieves this level of resolution. Our model should therefore provide a reasonable description of the density distribution at small scale.

All the $8 \times 8 \times 8$ cubes in the 12.5 Mpc h^{-1} simulation are considered in the present analysis. Clumping factors are computed by averaging the relevant quantities on these 512 cells. This 8^3 cell volume in the 12.5 Mpc h^{-1} corresponds to the volume of one cell in the 100 Mpc h^{-1}. The distributions of

[4] Incidentally, it also shows that some consistency is achieved within our code.

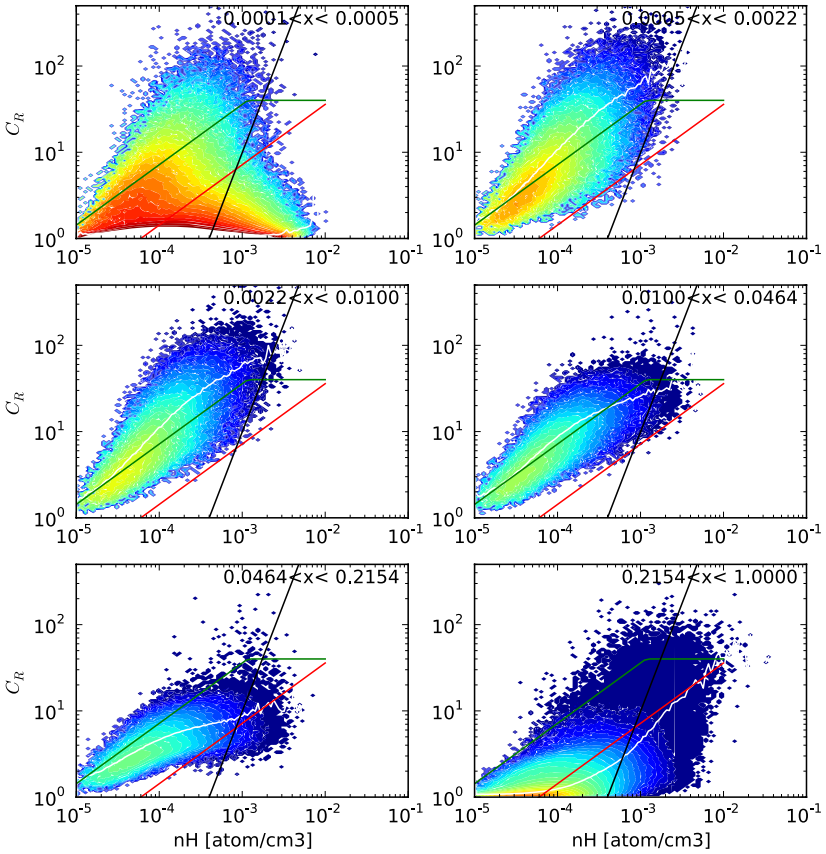

Figure 13. Clumping factors C_R as a function of the density computed from $8 \times 8 \times 8$ cells in the 12.5 Mpc h^{-1} simulation and used in the 100 Mpc h^{-1} experiment. The six panels show the C_R distributions at different ionization levels: the solid line shows the $\langle C_R \rangle (n_{\rm H})$ trend while the green dotted and the red dashed line stand for $C_R \sim n_{\rm H}^{0.7}$ models with, respectively, a high and a low normalization. The black dashed line stands for models with $C_R \sim n_{\rm H}^{2.5}$ at high densities.

(A color version of this figure is available in the online journal.)

C_R and C_I as a function of the density $n_{\rm H}$ are shown in Figures 13 and 14 for six bins of ionization levels with a roughly equivalent number of data points. This distributions were performed by averaging six snapshots between $13.1 < z < 5.9$, yielding clumping factors which do not depend on redshift but only on the physical properties of the cell. However, we do not imply that no redshift dependence exists: for example, our measurements show that the clumping increases at a given density value with time as a result of clustering of subgrid structures (not shown here). Still, we checked that the models described below still encompass the redshift by redshift results. This choice aims at simplicity, is not a physical statement, and can suffer from two caveats, one statistical, the other more physical. First, the distribution can be skewed by a snapshot which dominates the other. Second, by summing the contributions at all redshift we basically ignore a possible redshift dependence of the clumping and somehow ignore the ionization history of a cell.

Clearly, Figures 13 and 14 present an important spread of values around the mean trend (shown in white). It should not come as a surprise since we somehow projected the clumping factors on the $x, n_{\rm H}$ space, putting aside, e.g., the temperature or the local ionization field. Still some trends can be fitted to a good level of approximation. Considering the distributions of C_R first the distribution is reasonably fitted by $C_R \sim n_{\rm H}^{0.7}$ trends, especially considering ionized fractions greater than 0.005. The fit is poorer for smaller x but such a level of neutral gas does not contribute much to recombination. The same normalization can be applied for $0.005 < x < 0.2$ (see the green dotted curves) but a smaller normalization (red dashed line) seems to be necessary to fit the mean trend for the last class of ionization. Such results are consistent with the "B" clumping factors found by Kohler et al. (2007) with their mass resolution being lower by an order of magnitude and a larger averaging coarse grid. The agreement first suggests that some convergence has been achieved for clustering measurements at this scale. It should also be noted that a clustering law close to proportionality does not strongly depend on the size of the coarse cell used to compute its behavior. Finally, Figure 13 exhibits an important dispersion, therefore the $n_{\rm H}^{0.7}$ trend is consistent with the distribution but not in the strongest fashion. For instance, we also recover their "A" clumping factors which are biased toward high densities and which follow a $C_R \sim n_{\rm H}^{2.5}$ power law (shown as dashed dark lines): a subsample of cells follows this trend for high densities like the outliers in the $0.04 < x < 0.2$ panel or the high-density rise of the distribution in the $x > 0.2$ panel. By looking at the distribution for single redshifts (not shown here), C_R coefficients with the same "A" trend can be found for all the ionized fractions $x > 0.0005$ at the highest densities. But their overall weight is such that volume averaged trends tend to follow a "B" law with a gentler slope (shown in white), 0.7 instead of 2.5.

From there it is clear that a single normalization of the $C_R \sim n_{\rm H}^{0.7}$ law cannot be representative of all the ionization fraction classes. Therefore, we choose to perform two sets of runs, one with a "high normalization" (shown in green in Figure 13) and one with a "low normalization" (shown in red). The former is adequate for $x < 0.2$ but overestimates clumping in ionized regions while the latter underestimates clumping

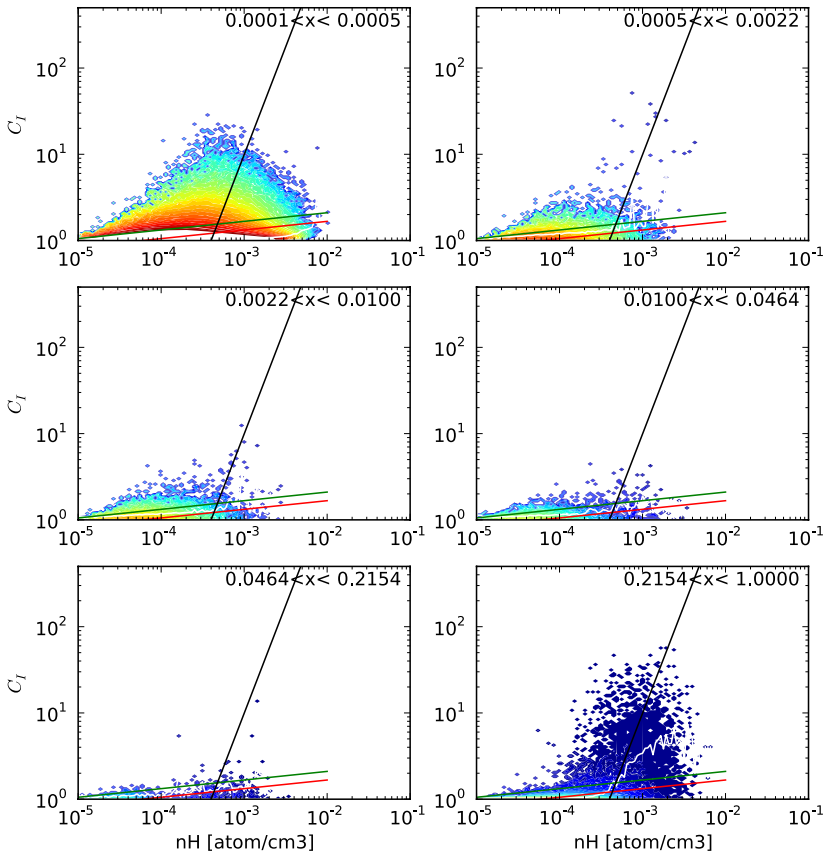

Figure 14. Clumping factors C_I as a function of the density computed from $8 \times 8 \times 8$ cells in the 12.5 Mpc h^{-1} simulation and used in the 100 Mpc h^{-1} experiment. The six panels show the C_I distributions at different ionization levels: the solid line shows the $\langle C_I \rangle (n_{\rm H})$ trend while the green dotted and the red dashed line stand for $C_I \sim n_{\rm H}^{0.1}$ models with, respectively, a high and a low normalization. The black dashed line stands for models with $C_I \sim n_{\rm H}^{2.5}$ at high densities.

(A color version of this figure is available in the online journal.)

in regions with high neutral fraction but is a better fit of the clumping in ionized cells. Again, this is consistent with the result found by Kohler et al. (2007) who studied the neutral fraction dependence of the clumping factors. Furthermore and as shown hereafter, no strong differences can be noted between these two calculations suggesting that a more detailed C_R with an x dependence (which should lie in between) would lead to similar results.

Considering next the photoionization clumping C_I, which distributions are shown in Figure 14, it clearly appears that the clumping is less pronounced than for recombination. The mean trend can be fitted by $C_R \sim n_{\rm H}^{0.2}$ laws (shown as red and green lines in the panels), which we used in the subsequent experiments. It should be noted that we recover again a $C_I \sim n_{\rm H}^{2.5}$ for high-density outliers but they are not representative of the overall distributions of C_I values. In the end, it appears that the photoionization clumping factors are much smaller than the recombination ones and incidentally during this work we performed calculations with $C_I = 1$ (and $C_R \sim n_{\rm H}^{2.5}$). It is equivalent to assuming a homogeneous UV background: from Figure 11 it is quite clear that J_{21} does not depend strongly on the density for a large range of values and the approximation made by choosing $C_I = 1$ should be reasonably close to the actual clustering.

3.2.2. Results

The simple clumping models described in the previous section were added to the basic version of the chemistry/cooling module, and radiative transfer has been performed again for the 100 Mpc h^{-1} box, with the same boost factor as the one given in Table 1. The simple visual inspection of the fields is quite informative on the impact of our subgrid clumping model on transfer calculations: Figure 15 presents the same slice within the box, at the same instant during the pre-overlap phase but for calculations with or without a subgrid model. First, the overall geometry can be recognized in both calculations; however, the experiment with a subgrid clumping model presents less extended ionized regions indicating that the overall chronometry has been modified: with clumping the radiation field is less efficient in ionizing the gas and requires therefore a longer time to achieve a certain level of ionization. Moreover, the experiment without clumping presents ionization front(s) which appear smoother than they are in the subgrid clumping model, reflecting again the increased difficulty for radiation in passing through higher density regions. Finally, if one looks closer at photoionized regions, much more pockets of neutral gas are seen in the clumping model, as a consequence of the larger recombination rate.

To assess these aspects more quantitatively, we present in Figure 16 the same distributions as in Section 3.1.5, namely, $x(n_{\rm H})$ and $J_{21}(n_{\rm H})$ but within our clumping factor model. These distributions are given at a post-overlap redshift ($z = 5.92$). Like previously, the volume-averaged radiation field follows closely the flux in low-density regions ($n_{\rm H} < 5 \times 10^{-4}$ cm^{-3}), where its intensity is quasi-constant. For larger densities, a strong exponential cutoff is observed, with a radiation field of three orders of magnitude smaller than the average value. Clearly, high-density regions live in a radiation field different

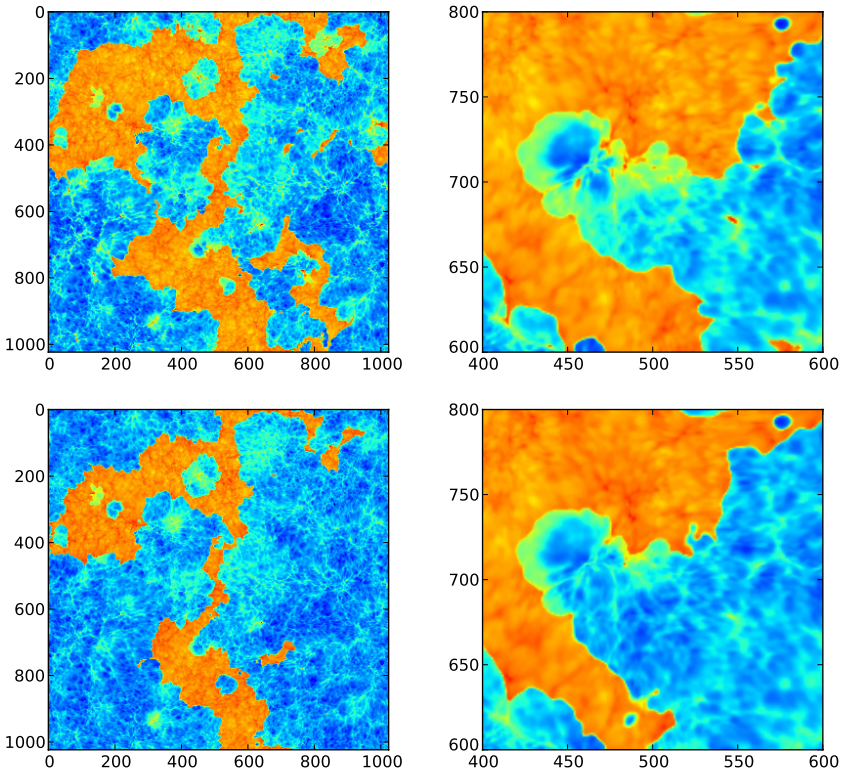

Figure 15. Neutral fraction maps (red zones are neutral, blue ones ionized) in a slice of thickness equal to 9.7 kpc h^{-1} comoving at $z = 6.97$. Distances are given in pixels and each map side covers 100 Mpc h^{-1} comoving (left column) and 19.53 Mpc h^{-1} comoving (right column). The top row pictures stand for the calculations with subgrid clumping and the bottom row ones for the same calculation without subgrid clumping.

(A color version of this figure is available in the online journal.)

than the rest of the simulated volume. We use the same type of models J_0, J_1, and J_4 as previously with $n_1^* = 0.006$ cm^{-3} for both clumping models and $n_4^* = 0.016$–0.025 cm^{-3} for resp. the high- and low-normalization models. We recompute the equilibrium ionized fraction and compare it to the distribution actually found in the numerical experiment. The calculations are performed assuming the same clumping models as the one used during the simulation and shown in Figure 17. When compared to the calculation without subgrid clumping, it can be noted that the fraction of neutral is more important and that gas tends to be more neutral at a given density. The "cobra rise" of the average J_{21} as a function of z is steeper with clumping and saturates at lower density than it used to. At density close to 5×10^{-2} cm^{-3}, the distribution of neutral fraction is clearly bimodal: a first peak stands for high-density regions which are ionized $1 - x \sim 0.001$ while a second population has $1 - x \sim 1$. It indicates that some gaseous regions are sufficiently embedded and/or recombine fast enough to be "spared" by the radiation field. Again the J_0 model is completely off the mean $1 - x(n_{\rm H})$ trend for high-density regions $n_{\rm H} > 0.001$ cm^{-3}, even though this level of radiation is effectively the one measured when averaging over the whole volume. Meanwhile the J_1 does a better job and J_4 appears to be a good match which is not surprising since they are better fits to the local density dependence of the UV background. Because of self-shielding, it is clear that the ionized state of high-density regions cannot be deduced by the simple generalization of the average UV field found in the simulated volume.

3.2.3. Comparison to Observational Constraints

We first consider the evolution of the volume- and mass-averaged neutral fraction, shown in Figure 18. Compared to the experiments without subgrid clumping, the fractions are typically one order of magnitude larger when subgrid structures are modeled. At $z = 5.9$, the volume averaged neutral fractions are equal to $x \sim 5 \times 10^{-5}$ and the mass averaged to $x \sim 3 \times 10^{-3}$. The differences between the low- and high-normalization models for the clumping are small with higher neutral fraction for the high-normalization model, which is expected since it overestimates the recombination rate in the most neutral regions. These levels of neutral gas are consistent with observational constraints provided by Fan et al. (2006)[5] and for the two types of averaging and indicate that high resolution or subgrid clumping is required to match the data. For instance, the same agreement has already been obtained by Trac & Cen (2007) using directly the particles as a source of clumping in a high-resolution pure dark matter simulation. Other examples are Kohler et al. (2007) using clumping factors and Gnedin & Fan (2006) using small box calculations for the volume-weighted neutral fraction.

The agreement found in the current work is encouraging but should nevertheless be considered with some care. First, the clumping models are simple and lack a detailed dependence on, e.g., the temperature or UV field. But even with a more accurate description of the subcell physics, clumping factors will remain as a trick to cope with inadequate resolution and the forthcoming effort should concentrate on improving the resolution using larger simulations. For instance, the low clumping model exhibits a small increase in the mass-weighted neutral fraction which is clearly inexistent in the observational

[5] The observational constraints shown here were recomputed from the transmission tables of Fan et al. (2006) using the same cosmology as the one used in our calculations. It results in a relative variation of 20%.

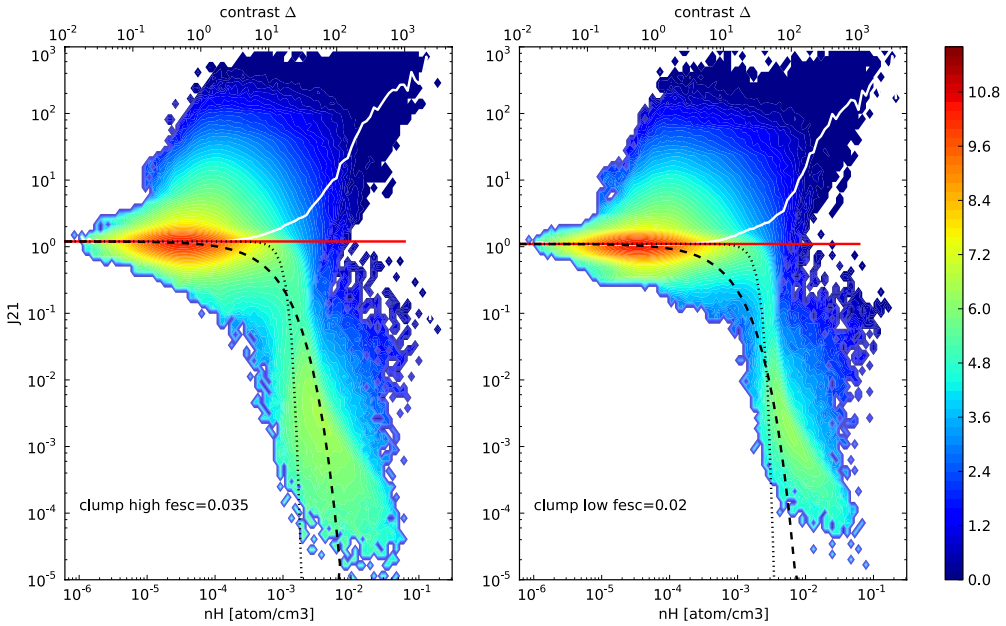

Figure 16. Density dependence of the ionizing intensity in the 100 Mpc h^{-1} box with subgrid clumping. The red dashed line shows our constant ionizing background model J_0, the dashed black line shows our $J_{21} = J_0 \exp(-n_{\rm H}/n_1^*)$ model, and the dotted black line shows our $J_{21} = J_0 \exp(-(n_{\rm H}/n_4^*)^4)$ model. The white lines show the average intensity per density bin. The top row results were obtained assuming the clumping law with a high normalization and the bottom row ones with a low normalization.

(A color version of this figure is available in the online journal.)

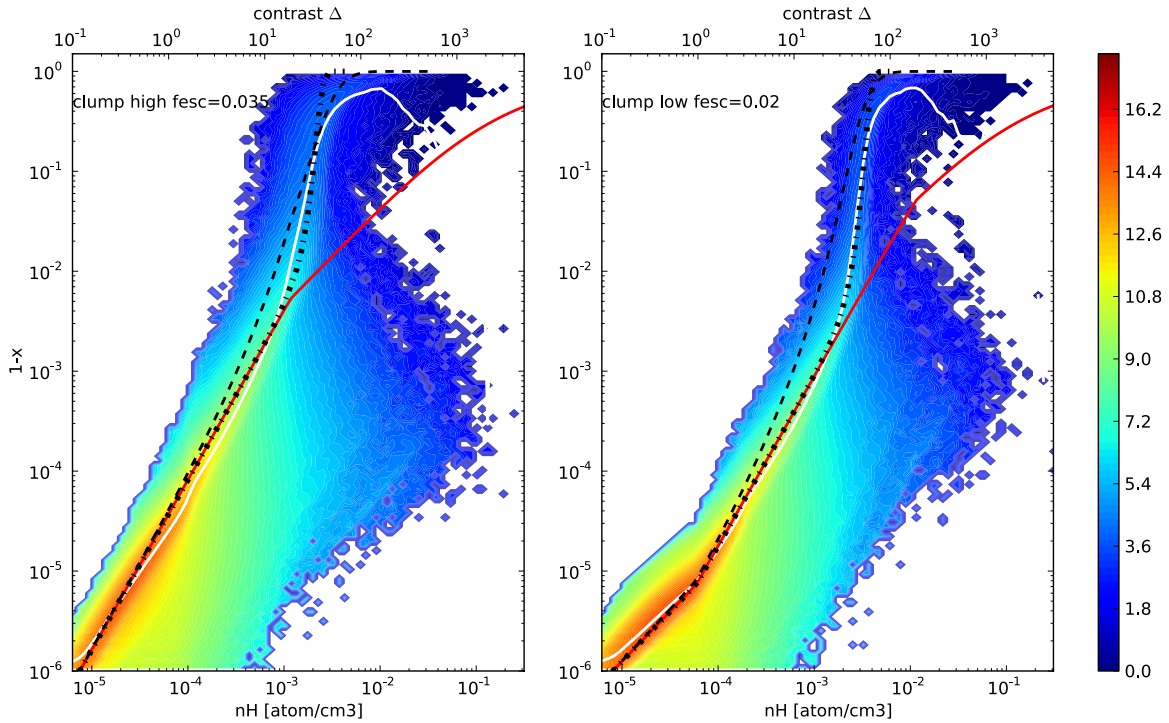

Figure 17. Density dependence of the neutral fraction in the 100 Mpc h^{-1} box with subgrid clumping. The lines show the neutral fraction for our three ionizing background models, assuming photoionization equilibrium and a clumping factor model. The white lines show the average neutral fraction per density bin. The top row results were obtained assuming the clumping law with a high normalization and the bottom row ones with a low normalization.

(A color version of this figure is available in the online journal.)

data and also in other calculations. This model is peculiar since only the densest regions ($\Delta > 10$) experience any clumping, whereas the bulk of the gas is considered as unclumped. With such a model it is difficult to guarantee a good agreement of both mass and volume weighted neutral fraction. This slight increase demonstrates the limits of this simple model. Furthermore, we still lack the coupling with the hydrodynamics which is likely to affect small-scale features of reionization and such physics cannot be assessed using only a subgrid clumping model. In the end, we estimate that given the simplistic aspect of our clumping model, the current agreement should be seen as a sign that the overall direction is correct, but it is not clear that improving the

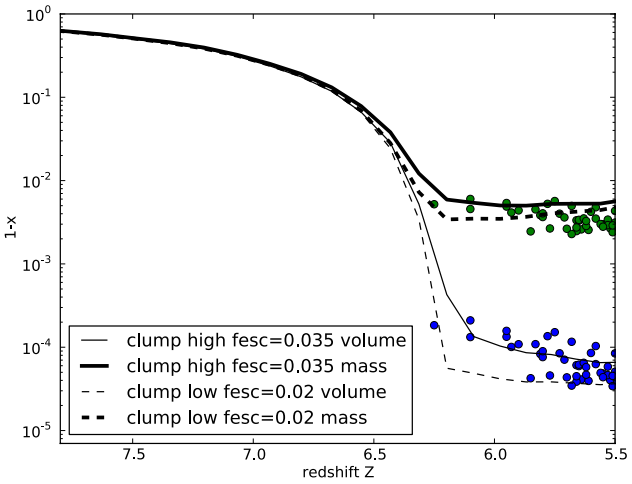

Figure 18. Evolution of the mass- and volume-averaged neutral fraction in the 100 Mpc h^{-1} box with a clumping factor assuming a high/low normalization (thin/thick lines). The values at $z = 6$ are consistent with measurements made by Fan et al. (2006) for both kinds of average methods (dots).

(A color version of this figure is available in the online journal.)

model is worth the effort, compared to increasing the resolution of the simulation. Second, Trac & Cen (2007) already noted that the constraints provided by Fan et al. (2006) are model dependent. In the most pessimistic case, the present agreement can be fortuitous, even though an agreement on both the volume and mass-averaged neutral fraction is unlikely to happen by accident.

It should however be noted that the average neutral fractions do not correspond to any feature in the probability density distributions. In Figure 19, we overlay the evolution of the averaged neutral fractions on the evolving distributions of probabilities. Red regions stand for high probabilities while blue ones stand for lower probabilities: after the overlap, the distributions are clearly dominated by strongly ionized regions ($x \sim 10^{-5}$) which correspond mostly to low gas densities. The mass-weighted distribution enhances the contribution of dense cells and which in turn push the average neutral fraction toward higher values ($x \sim 10^{-2}$). However, the values of the average neutral fractions are never coincident with the maximum of the distribution. On the contrary, these averages lie in the transition region between low neutral fraction and high neutral fraction regions. In other words, the average values lie in between two characteristic regions of the gas distribution, but are representative of neither of them and consequently are not a good proxy of the physical states that coexist inside the simulation.

In Figure 20, we present the evolution of the distribution of the UV background in the 100 Mpc h^{-1} boxes with subgrid clumping with the mean value superimposed and the constraint provided by Bolton & Haehnelt (2007) after the overlap. Compared to the same calculation without clumping, some improvement can be seen and the intensity of the UV background has been reduced by a factor of two or three (depending on the kind of normalization applied to the recombination clumping). It can be noted that the same ratio was already observed between the 100 Mpc h^{-1} and 12.5 Mpc h^{-1} (which served to calibrate our clumping model) boxes in our fiducial calculations. Still, the discrepancy in the observational constraints remains quite large, almost 1 order of magnitude. Furthermore, the inspection of the distribution indicates that the mean value of the UV background effectively tracks the maximum of the J_{21} distribution. Therefore, no bias or multimodal distribution can be invoked as a valid reason for the discrepancy.

This failure can be explained by the fact that very different regions are at the origin of the average value for the neutral fraction and for the UV background values. After the overlap, the neutral fraction is intrinsically low in low-density regions ($1 - x \sim 10^{-5} - 10^{-6}$) and few regions with high neutral fraction push the average to higher values: at face value a single fully neutral cell weighs as much as 10^6 cells with a 10^{-6} neutral fraction. Furthermore, if mass-weighting is considered, the impact of such cells is even higher since they are usually more neutral. As a consequence, the average neutral fraction is pushed toward values higher than the peak of the distribution and dense cells (even less numerous) have an important impact. Considering now the average UV background, dense cells lie in regions where the radiation intensity is typically 1000 times smaller than the typical value computed in low-density regions because of self-shielding, which implies a smaller impact on the average J_{21}. Consequently, the mean neutral fraction is not related to the mean UV background: the former is influenced by dense clumps while the latter is mainly set by voids. The fact that dense regions do not lie in the typical UV background explains our ability to reproduce the neutral fraction and our failure to satisfy the constraints on the ionizing radiation field.

One may therefore ask how do we balance a discrepancy in the photoionization rate and an agreement in neutral fraction? First let us recall that the actual quantity measured in quasar spectra is the transmission $\mathcal{T}$, i.e., the ratio of the observed flux to the unobscured one over a given range of redshifts (see, e.g., Fan et al. 2002, 2006). At the redshift considered here, the equivalent comoving distances are of the order of 60 Mpc h^{-1} and are therefore comparable to the experiments presented in the current work. Hence, observations give a constraint on

$$\langle \mathcal{T} \rangle \sim \int p(\Delta) e^{-\alpha Q(z)\Delta^2/\Gamma} d\Delta, \tag{16}$$

where Δ stands for the density contrast, α is the recombination rate (mostly homogeneous), $Q(z)$ depends on the physical parameters of the Lyα transmission and cosmology, and finally $p(\Delta)$ is the pdf of the density. A typical example of such a pdf is given by Miralda-Escudé et al. (2000) which is used in the models of Fan et al. (2002), Fan et al. (2006), or Bolton & Haehnelt (2007). In Equation (16), the photoionization rate can be deduced from $\langle T \rangle$ and assuming photoionization equilibrium the neutral fraction can also be deduced. We have seen that the latter assumption is mostly verified. The relation also assumes that the universe is mostly ionized and that the UV background is homogeneous. From the expression in Equation (16), it can be easily seen that the exponential cutoff implies that the actual density distribution at high density has no influence on the observed quantity, therefore their departure from homogeneity and low neutral fraction (measured in simulations) does not impact on the transmissions. Conversely, it implies that the quantities which are mostly constrained are inferred from low-density regions (for a detailed discussion see, e.g., Oh & Furlanetto 2005). In other words, the photoionization rate is more "reliable" or more directly constrained than the neutral fraction and should be reproduced first by simulations: in principle the agreement on the neutral fraction should follow. In the current work, we show however that reproducing first the neutral fraction does not automatically imply an adequate photoionization rate.

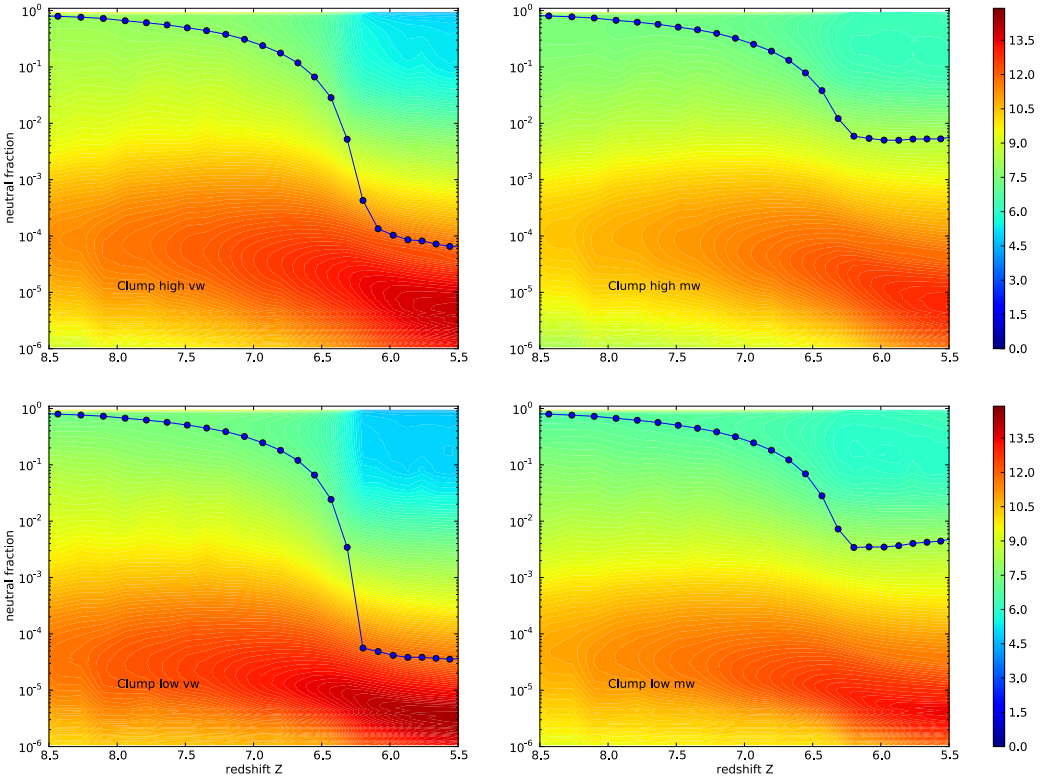

Figure 19. Evolution of the neutral fraction distribution with redshift along with the evolution of the average value using mass weighting (bottom row) and volume weighting (top row) for the 100 Mpc h^{-1} box with clumping. The left column stands for experiments with high normalization clumping and the right one for the low normalization clumping model. Contours show the density probabilities of neutral fraction with high probability densities in red and low ones in blue, the scale being logarithmic.

(A color version of this figure is available in the online journal.)

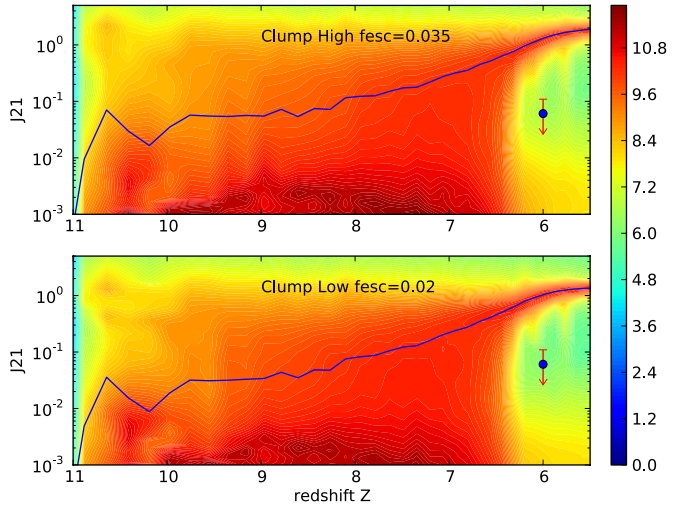

Figure 20. Evolution of the average ionizing radiation (blue line) in the 100 Mpc h^{-1} box assuming a subgrid clumping factor model with a low/high normalization (top/bottom panel). The colored isocontours stand for the distribution of J_{21} values at each redshift with high probability densities in red and low ones in blue, the scale being logarithmic. The marker at $z \sim 6$ shows the constraint provided by Bolton & Haehnelt (2007).

(A color version of this figure is available in the online journal.)

To further emphasize this point, we computed the evolution of the effective Gunn–Peterson optical depth $\tau_{\rm eff} = -\log\langle e^{-\tau}\rangle$ and compared it to the values measured by Fan et al. (2006). Again, this quantity is directly observed and is therefore a more stringent test of our calculations. The results of our calculations for the high clumping model in the 100 Mpc h^{-1} box are shown in Figure 21. Clearly we underestimate the GP optical depth by a factor of two, which confirms the above reasoning: by overestimating the transmission, we end up with an overestimate of the photoionization rate even though the level of neutral is reproduced. We proceeded by taking the average on one-dimensional skewer with a length corresponding to $\Delta z = 0.15$, which is different than computing $\langle\tau\rangle$, which is much larger because of a large value bias. The distribution of τ is also shown as a color map and interestingly the observed optical depth roughly corresponds to the most probable value of τ. Still, the comparison should be made on $\tau_{\rm eff}$ since the transmission (or $\tau_{\rm eff}$) is the observed quantity: the disagreement on the intensity of the UV background cannot be put at the same level as the agreement on x and remains to be resolved.

3.2.4. Improving the Model

Which path should be taken toward a complete agreement between observational constraints and our calculations? The most obvious free parameter we have access to is the escape fraction. We present in Figure 22 the evolution of the averaged UV background and neutral fraction for various escape fractions and using the same 1024^3 particles 100 Mpc h^{-1} simulation with the high normalization clumping factor model. Plotted along are the constraints provided by Fan et al. (2006) and Bolton & Haehnelt (2007). As expected, lowering the escape fraction makes the simulated UV background more consistent with observations. However, also as expected, the redshift of reionization decreases and for the lowest value of $f_{\rm esc} = 2.5\%$ presented here, overlap is not complete and the average neutral

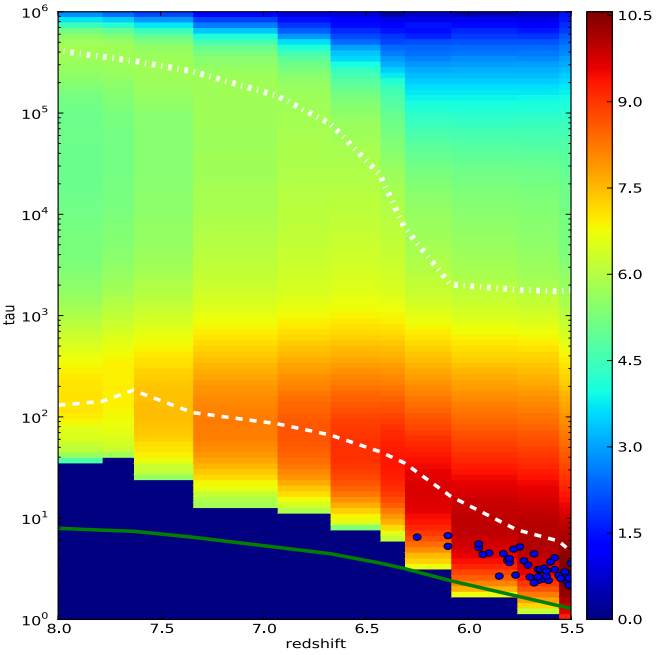

Figure 21. Evolution of the effective GP optical depth $\tau_{\rm eff}$ in the 100 Mpc h^{-1} box with high normalization clumping. Dots: measures of the effective optical depth $\tau_{\rm eff} = -\log T$ made by Fan et al. (2006). Green line: the average effective optical depth measure from our 100 Mpc h^{-1} simulation with subgrid clumping. Color map: probability distribution of $\tau \neq \tau_{\rm eff}$ measured in the same simulation, the scale being logarithmic. White dash-dotted line: the redshift evolution of $\langle\tau\rangle$. White dashed line: the redshift evolution of the maximum of the pdf of τ.

(A color version of this figure is available in the online journal.)

fraction is only at 5% at $z = 6$. Such a scenario is problematic because it implies that the neutral fraction must decrease very sharply: observational data exhibit some transmission for quasars at redshifts $z \sim 5.9$, i.e., at levels of neutral fraction close to $1 - x \sim 0.0001$ and it would imply a sudden decrease from $x \sim 0.1$ to 10^{-4} in a small redshift interval of $\Delta z \sim 0.1$. Furthermore such a trend would also go against a better agreement with the optical depth measured from the CMB data which favor a higher escape fraction. One option would be to use an evolving escape fraction, from ∼10% at $z \sim 10$ down to $f_{\rm esc} \sim 1\%$ at $z \sim 6$. Preliminary experiments (not shown here) indicate that, albeit helpful, this option does not easily provide a solution to the discrepancy. Furthermore, even if a good match is obtained, it would only consist in a proof of concept and one would have to relate this evolution to a physical process (like, e.g., star formation). Other routes can be used to reduce this discrepancy. The lack of multi-frequency transfer implies among other things that no preheating occur behind ionization fronts. In particular, it would reduce both the recombination rate of the gas and the required number of photons per baryons to complete reionization. Finally, the proper coupling of radiative transfer with hydrodynamics may prove to be crucial: low-density regions or mini-halos are likely to react to any kind of heating due to radiative transfer and the source production (namely, star formation) may be affected (Iliev et al. 2005, 2007), even though self-shielding, which has been shown to be quite effective in our calculations, could go in the opposite direction. These points will be investigated in future work.

4. SUMMARY AND PROSPECTS

We have presented a set of radiative cosmological simulations in order to model the reionization epoch from $z \sim 18$ down to $z \sim 6$. The gas and dark matter dynamics, as well as the associated star formation have been performed with the RAMSES code, while radiative transfer has been computed by means

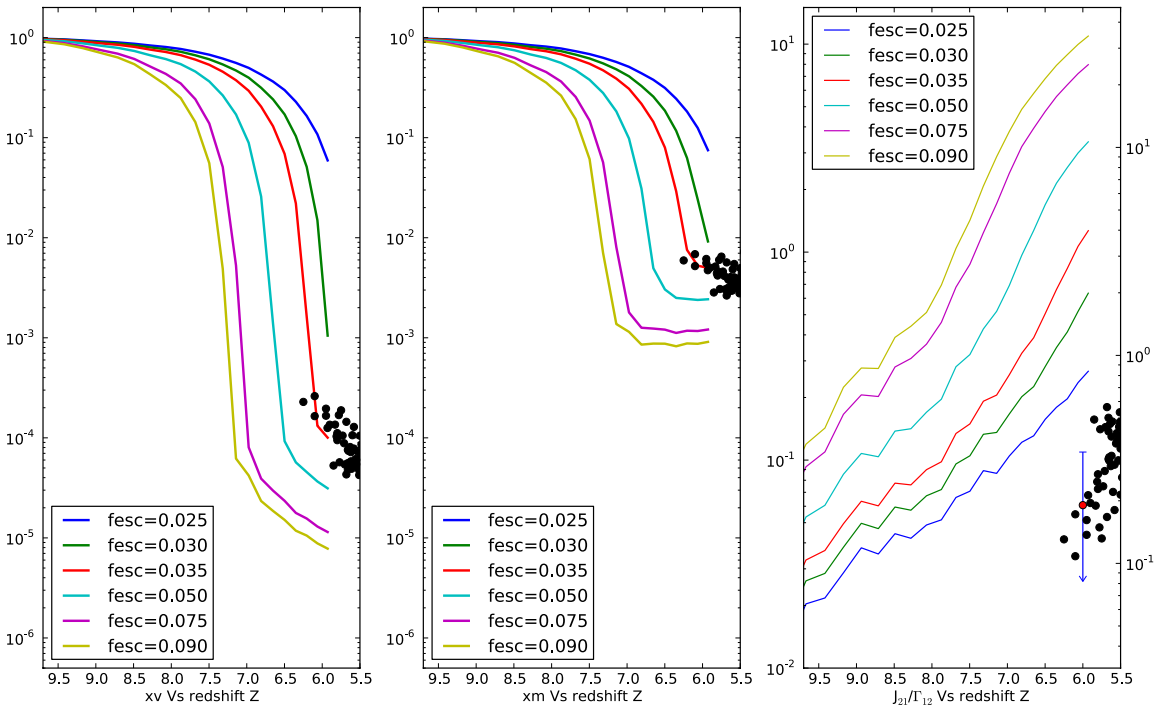

Figure 22. Evolution of the mass, volume-averaged neutral fraction, and ionizing rate in the 100 Mpc h^{-1} box with a clumping factor assuming a high normalization for various escape fractions. The values at $z = 6$ are consistent with measurements made by Fan et al. (2006; dots). The red arrow at $z \sim 6$ shows the constraint provided by Bolton & Haehnelt (2007).

(A color version of this figure is available in the online journal.)

of a moment-based formalism using the M1 closure relation, implemented in the ATON code. The latter has been ported on a multi-GPU architecture using CUDA, providing an acceleration close to 100x, which allows us to tackle radiative transfer problems at high resolution (a 1024^3 base grid and 2 levels of refinement for the hydrodynamics and a 1024^3 Cartesian grid for the radiative transfer).

A good level of convergence on average quantities (neutral fraction, UV background, and Thomson optical depth) has been observed between different simulations of increasing mass and spatial resolution, as long as the effect of finite mass resolution on the simulated SFH is properly taken into account. We have also shown that the density dependence of the neutral fraction is close to the one predicted by photoionization equilibrium, as long as the effect of self-shielding is considered when defining the properties of the UV field. It also appears that without any other ingredients, our simulation fails in reproducing the $z \sim 6$ constraints on the neutral fraction of hydrogen and the intensity of the UV background, in a similar manner to Finlator et al. (2009b).

By combining our best resolved simulation ($12.5\,\mathrm{Mpc}\,h^{-1}$ and 1024^3 particles) with our largest simulated volume ($100\,\mathrm{Mpc}\,h^{-1}$ with 1024^3 particles), we have introduced a sub-grid clumping model in our chemistry solver, consistent with the one derived by, e.g., Kohler et al. (2007). We have shown that, although this clumping factor model is quite simplistic, it has allowed us to reproduce the level of neutral gas deduced from the spectra of high-redshift quasars, as did previously Gnedin & Fan (2006) or Trac & Cen (2007) among others. However, our estimation of the average photoionization rate is still at least a factor of two above the observational constraints. This "semi-success" can be explained by the fact that the average radiation intensity and the average neutral fraction depend on different regions of the gas distribution and one cannot simply deduce one from the other using photoionization equilibrium: in other words, if one constraint is satisfied, the other cannot be. However, we have argued that the photoionization rate is probably a more robust observational constraint than the neutral fraction. This suggests that some effort should still be done in our modelization to reproduce the level of the UV background at $z \sim 6$.

Among several prospects, one obviously thinks of increasing the resolution of the calculations. With GPU acceleration 2048^3 hydro + radiative transfer calculations are within reach. However, it clearly appears that coupled hydrodynamics and radiative transfer simulations are necessary at this stage (such as, e.g., Trac et al. 2008 at a comparable resolution), since an increase in resolution will inevitably raise the question of the impact of radiation on mini-halos or on the SFH. Also additional physics should be implemented, such as multi-group radiative transfer, where the importance of preheating by X-rays could therefore be fully assessed (see, e.g., Furlanetto 2006; Shull & Venkatesan 2008), but also Population III stars (Trac & Cen 2007) and varying star formation efficiencies and escape fractions (Gnedin et al. 2008; Wise & Cen 2009). Overall, on a final positive note, our current results indicate a satisfying trend of cosmological calculations toward satisfying observational constraints.

D.A. is supported by the ANR grant LIDAU and a *Conseil Scientifique* Grant from the University of Strasbourg. This work was granted access to the HPC resources of CCRT under the "Grand Challenge Applications" allocation for 2009.

APPENDIX

ON THE GPU IMPLEMENTATION OF ATON

This section comments in further detail on the implementation of ATON on GPU. The whole development has been performed using the version 2.2 of the CUDA extension to the C language, developed by Nvidia for its graphics devices. However, this section should be seen as a commentary of the implementation process rather than a full description of the programming details: the field of GPU programming is currently in full expansion, several standards/programming languages are competing with each other and many specific programming details are likely to be quickly outdated. For these reasons, we choose to stick to fairly general techniques, comment on the suitability of the calculations to multithreaded calculations, and provide the general tricks of our development to optimize the performances.

Let us first recall that GPU computing relies on two separate hierarchies: a hierarchy of memories and a hierarchy of tasks. Regarding the first aspect and because GPUs are devices physically separated from the host, they possess their own memory known as the video memory or "global memory" in the CUDA nomenclature. The transfer rate between the host and the GPU is therefore strongly limited by a bus and the best performances can be achieved if all the calculations are performed on the device, i.e., without transfer between the host and the GPU. An ideal situation would be to transfer the ICs on the device and let it process the calculation on its own with the host acting merely as a driver of the calculation. This constraint is satisfied by the GPU implementation of ATON: the host sends signals to the GPU in a regular fashion to advance the simulation within a time step and from one time step to another but it never actually computes anything on the data. More precisely, the ICs are sent on the GPU, then the host asks the GPU to compute the radiation transport, then the chemistry and the cooling. Then the same signals are sent again to the GPUs to perform the next time step until the simulation is completed. If required, the data are transferred back on the host to write the snapshot on the disk but such a situation occurs only once in a while (typically once every 5000 time steps in our case). In such a procedure, only the host is aware of the fact that a time evolving calculation is performed but only the GPU does actual calculations. Furthermore, the global memory can be as big as 4GB on current devices and is used to store the data (like, e.g., a large grid of values). This memory space is usually sufficient but slow to access. On-chip memory also exists, with fast access, but is usually small (of the order of 16 kB) and more importantly, still requires access to the slow video memory to be filled. Therefore if possible, any memory access should lead to a significant "number crunching" in order to make these memory transfers worthwhile. ATON fulfills this requirement quite easily since every time a cell is accessed (containing an energy density, flux, temperature, ionization fraction, and baryon density), the cell is fully updated and requires a transport calculation and the resolution of the ionization and energy balance equations which are quite demanding in terms of arithmetic intensity.

Regarding the hierarchy of tasks, GPUs are efficient in performing tasks (or execution *threads*) in parallel which are as follows.

1. *Independent*. If a given thread has to wait for the completion of another one to perform its calculation (e.g., in a naturally sequential algorithm like a reduction Sengupta et al. (2007)) or if two threads try to update simultaneously the same value (in, e.g., histogramming calculations, Aubert et al. 2009),

specific algorithmic techniques must be employed to keep the parallelism efficient. On the other hand, if calculations do not interfere with each other then porting these tasks on GPU architecture is usually quite easy.

2. *Predictable*. Tasks can be unpredictable in their operations (if-else branches) or in their memory accesses. The former lead to divergence between threads where their execution tracks are executed sequentially by the GPUs thus reducing the efficiency of parallelism. If divergences are limited to exceptions (i.e., have a small chance to happen) and are hidden in intensive calculations their impact remain small. The latter lead to non-coalescent and non-aligned memory access which greatly impacts on the performances.
3. *Compact*. A task is compact if it uploads data to a compact region in the memory. Again, compact calculation leads to coalescent memory access which greatly improves the acceleration of the calculation.

It turns out that ATON possesses these three qualities. To demonstrate it, let us first recall that the radiation transport equations can be written in a generic manner as

$$\frac{du}{dt} + \frac{dF(u)}{dx} = S, \tag{A1}$$

where u is a set of conserved variables (energy density and flux in the case of radiative transfer), $F(u)$ is the associated flux, and S is a generic source term. We considered a one-dimensional transport for simplicity. It translates into

$$\frac{\tilde{u}_i^{p+1} - u_i^p}{\Delta t} + \frac{F_{i+1/2}^p(u) - F_{i-1/2}^p(u)}{\Delta x} = S_i^p, \tag{A2}$$

when one considers an explicit finite difference (FD) scheme in order to update u at position i at time $p+1$. Usually the intercell flux can be exactly solved or approximated using the values in the neighboring cells through an operator g and, for instance,

$$F_{i+1/2} = g(i, i+1/2). \tag{A3}$$

Moreover the chemistry/temperature updates plus the effect of absorption can in our implementation be formally written as

$$(u_i^{p+1}, x_i^{p+1}, T_i^{p+1}) = \Phi\left(\tilde{u}_i^{p+1}, x_i^p, T_i^p\right), \tag{A4}$$

where T and x stand for the temperature and ionized fractions while $\tilde{u}_i^{p+1}$ stand for the conserved variables updated after the transport. As one deals with grid-based structures it is natural to assign a thread to the update of a specific cell. From there it can be seen that ATON is suited well for GPU parallelism according to the three qualities listed before.

1. *Independence.* All these calculations are explicit: the only intermediate results needed are the transport-updated $\tilde{u}_i^{p+1}$, and it is a local value. All the other inputs are initial state values which do not require communications during the calculation per se. As a consequence, all the cell updates (and therefore the threads) are independent and the overall procedure is free of threads collisions or sequential calculations where one thread has to wait for the completion of one or several other tasks.
2. *Predictability.* Here the calculations are at least "memory-predictable." Updating a given cell requires data in a region which is known by advance, i.e., the updated cell plus its six neighbors in three dimensions. Operation branching occurs in the cooling and chemistry calculations and has some impact on the performance (see the subsequent analysis).
3. *Compacity.* Again the calculation requires a 7 cells stencil for a single thread which is quite compact and allows to enforce the coalescent memory access, as shown hereafter.

Finally, these devices can easily be used at full power if two properties are satisfied during the calculation: the data in global memory should be accessed in a *coalescent* and *aligned* fashion. Figure 23 allows us to explain the coalescence in detail assuming a two-dimensional calculation. The data are accessed in a coalescent fashion if a series of threads reads data which are organized in a sequential manner in the memory. In Figure 23, a two-dimensional field is physically stored in memory as a one-dimensional sequence listed by numbers 1–25. If a sequence of threads (shown colored) is set in such a way that threads access the data "vertically" (left scheme), they physically access data which are separated by jumps of five units: such a strategy is non-coalescent. Conversely if the sequence of threads is organized "horizontally" (bold border in the right scheme), they access data which are physically next to each other, i.e., in a coalescent fashion. All the threads in ATON are arranged following this strategy in order to enforce coalescence. For example, the chemistry/cooling step is performed with threads along the physically coalescent direction. The radiation transport step is slightly more difficult to set up as it involves an FD along all the directions:

$$\frac{\tilde{u}_{i,j}^{p+1} - u_{i,j}^p}{\Delta t} + \frac{F_{i+1/2,j}^p(u) - F_{i-1/2,j}^p(u)}{\Delta x} + \frac{F_{i,j+1/2}^p(u) - F_{i,j-1/2}^p(u)}{\Delta y} = S_{i,j}^p. \tag{A5}$$

For the FD performed along the coalescent direction, the coalescence is naturally satisfied. In order to avoid multiple access to the same data by neighboring threads, the coalescent values are uploaded in shared (on chip) memory once and calculations are performed from this shared memory. For the FD performed along the non-coalesced direction (vertical in Figure 23), a naive strategy would have been to upload the vertical values in shared memory (left column), i.e., along the direction of the FD. However this would imply non coalescent access. The correct way to deal with this FD is shown on the right panel of Figure 23. First the threads should be organized along the coalescent direction (shown colored). Then all threads upload the data "above" the region to update (shown with a bold line) in shared memory along the coalescent direction. The same is done for the data "below" the region to update. Finally the FD can be performed. From our experience, switching from non-coalescent to fully coalescent strategies can improve the performance of the GPU calculation by factor of 10–100. It should be said that such a "trick" is not specific to GPU-based calculation but given the high parallelization of the devices such an optimization has a more dramatic impact on their performances compared to usual scalar processors.

Aligned access is more specifically related to the hardware used. Typically, the data should be accessed in sets of 64, 96, or 128 words which is usually satisfied using thread configurations which rely on powers of 2. An additional constraint is that the range of memories accessed by these sets should be aligned with "preferential" memory addresses, usually multiples of 16. When dealing with arrays with dimensions equal to powers of 2, any access of sets of 64, 96, or 128 words will be automatically aligned. Non-aligned access will result in multiple memory queries on aligned addresses in place of a single one. It turns out that such a situation is quite common as boundary conditions

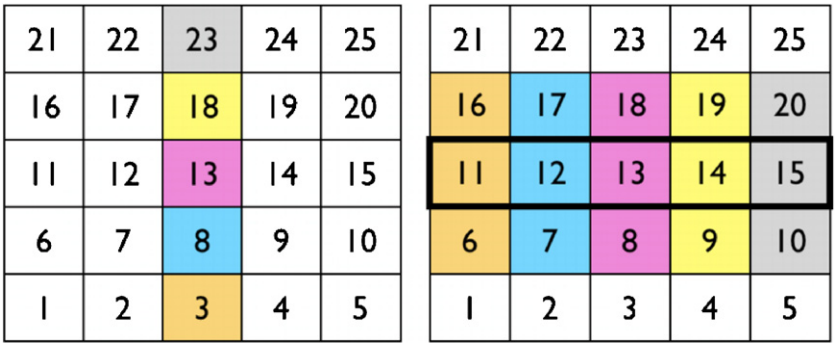

Figure 23. Comparison of two finite difference (FD) strategies on a two-dimensional field in memory. The sequence indicates the actual organization of cells in memory, the coalescent direction. We consider the case where the FD should be performed along the non-coalescent direction. Each color represents the location to be computed by a thread. Left: "vertical strategy" where the threads are arranged along the FD direction. Right: "horizontal strategy" where the threads are arranged perpendicular to the FD direction. The "horizontal strategy" maximizes the performance of the GPUs due to coalescent memory accesses. See the main text for details.
(A color version of this figure is available in the online journal.)

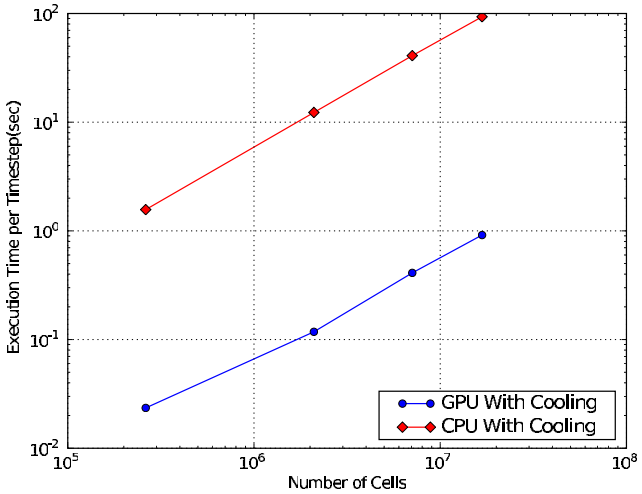

Figure 24. Average time steps duration for the cosmological test of the comparison project for CPU and GPU at different resolutions. The GPU and the CPU are roughly equivalent in terms of generation.
(A color version of this figure is available in the online journal.)

usually add a layer of data around the actual computational volume making, e.g., a 128 × 128 × 128 cube a (128 + 2) × (128 + 2) × (128 + 2) cube, which breaks the alignment. We circumvent this by making the boundary layer larger than required by the code since we are not limited by memory (e.g., 128 × 128 × 128 cube becomes a (128+32) × (128+32) × (128+32) cube). Typically an additional factor of 2–3 of acceleration can be achieved by enforcing alignment.

As an illustration of the computing abilities of GPUs, we show in Figure 24 the average duration of a time step of a radiative transfer post-processing performed on the cosmological test of the comparison project. The same test has been performed at several resolution and executed on an Opteron 2.2 GHz and a GeForce 8800 GTX, which are comparable in terms of generation (2005–2006). A significant acceleration close to 80 was observed on GPU compared to a monocore run on CPU. Both calculations were performed using single float precision and no difference could be seen between the calculations at such precision. Again, this acceleration is possible because of the initial choice of an explicit scheme, and does not hint of the speedup that could be achieved for an implicit scheme. However, since such techniques involve usually large sparse matrix solver, which are non-trivial to parallelize, it is likely that they would not benefit from the GPU acceleration at the same levels.

ATON is able to run on configurations with multiple graphical devices. It implies that GPUs should communicate in order to exchange boundary conditions. This is simply done by adding an MPI layer over the GPU inner parallelization. Once a GPU has updated its subgrid the following sequence happens at each time step.

1. A GPU-buffer is created on each GPU to gather the data to be passed at each time step (here, namely, the radiative energy and fluxes).
2. This GPU-buffer is transmitted to the host into a CPU-buffer.
3. The CPU-buffers are exchanged using regular MPI-based instructions.
4. The updated CPU-buffers are transmitted to the GPUs into the GPU-buffer.
5. The data inside the buffer are distributed into the correct radiative variables.

Considering the coalescence and alignment constraints depicted above, the natural parallelization for multi-GPU calculations is "slab-based" and, for example, a 512 × 512 × 512 calculation would be divided in four calculations 512 × 512 × 128 on 4 GPUs. The reason is that at each time step, the radiative energy and fluxes at the boundaries should be passed to the neighboring GPU by collecting them into a buffer and a slab-based configuration implies that the collection is performed in a coalescent manner (and the distribution as well). However, we choose to stick to subcube-based parallel configuration in the prospect of coupling with N-Body+hydro integrators (such as RAMSES) which parallel configuration is closer to "subcube" segmentation than "slab-based" ones. Furthermore, the slab-based decomposition cannot be naively applied for large problems because of hardware limitation such as the amount of memory per kernel (16KB) or the number of threads per block (512). Conversely, it implies that non-coalescent access is performed during the gathering/distribution phases (see Figure 25). Finally it should be noted that such communications require systematic transfers between the hosts and the GPUs through the PCI bus, which act therefore as a bottleneck in the communications. In Figure 26 we present the average duration of the time steps for several multi-GPU configuration and problem sizes and the acceleration as a function of the number of GPUs. Even though the implemented parallelization is simple, the speedup trends are quite optimal and the amount of time spent into the communication remains reasonable at levels of 10%–15%. This number is

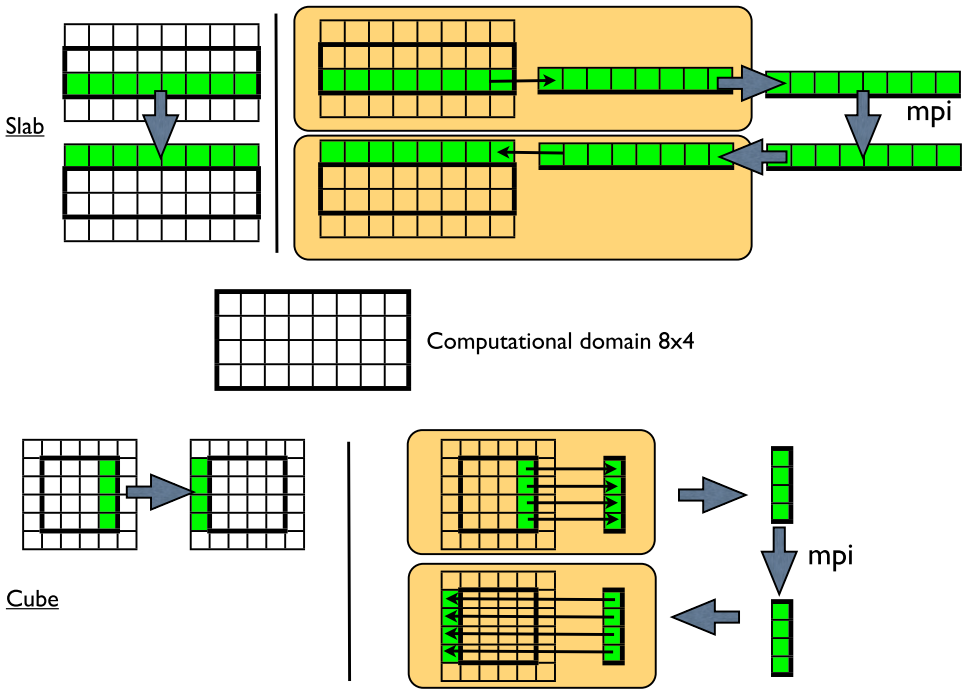

Figure 25. Two communication strategies for multi-GPU calculations. The coalescent direction is assumed to be the horizontal one. Top: slab-based communication. Bottom: cube-based communication. For the cube-based decomposition, some communications involve gathering and dispatching data in a non-coalescent manner. The cube-based technique has nevertheless been chosen for `ATON` to assess large problems, to reduce the shared memory usage and in the prospect of coupling `ATON` to integrators with cube-based decomposition.

(A color version of this figure is available in the online journal.)

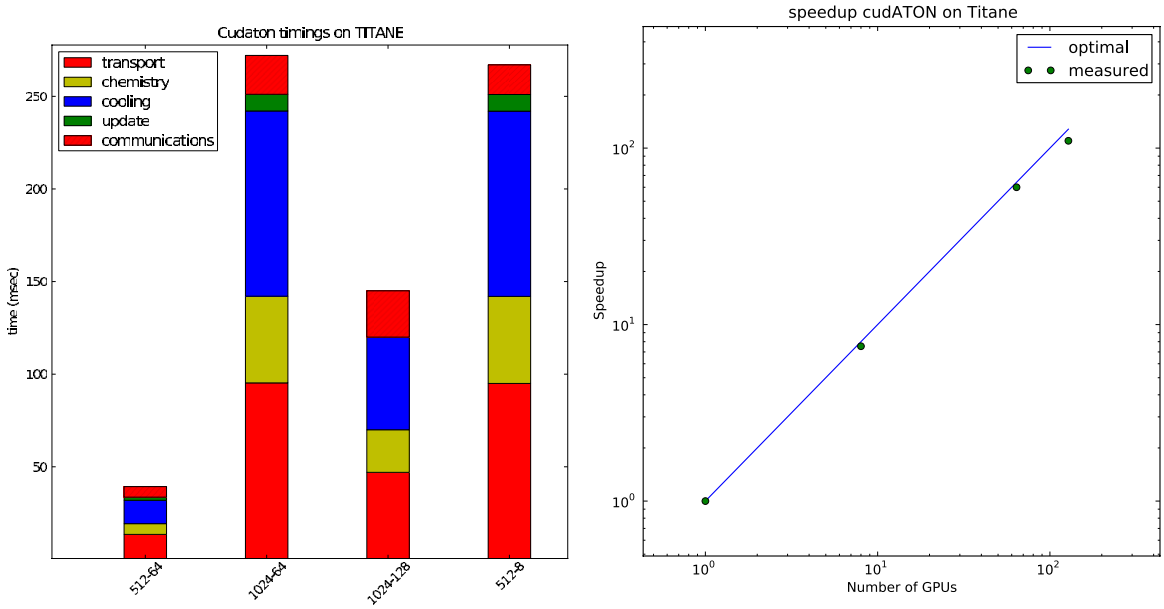

Figure 26. Timings of multi-GPU calculations. Left: average duration of a time step for a typical cosmological field used in this work and for several parallel configurations. The first integer stands for the total number of cells along one direction and the second stands for the number of GPUs. For example 512–8 means a 512^3 radiative transfer calculation distributed over 8 GPUs. Right: acceleration as a factor of the number of GPUs compared to a mono-GPU calculation, the dot stands for the actual measurement while the straight line stands for the perfect acceleration trend. Measurements were performed on the Titane supercomputer (CCRT-CEA) using Tesla C1060 GPUs.

(A color version of this figure is available in the online journal.)

quite large by standards of parallel high performance computing but compared to an initial acceleration of a few tens (compared to the CPU), this overhead remains small enough to assess large problems.

REFERENCES

Abel, T., Norman, M. L., & Madau, P. 1999, ApJ, 523, 66
Altay, G., Croft, R. A. C., & Pelupessy, I. 2008, MNRAS, 386, 1931

Alvarez, M. A., Bromm, V., & Shapiro, P. R. 2006, ApJ, 639, 621
Anninos, P., Zhang, Y., Abel, T., & Norman, M. L. 1997, New Astron., 2, 209
Aubert, D., Amini, M., & David, R. 2009, Lecture Notes in Comp. Science, 5544, 874
Aubert, D., & Teyssier, R. 2008, MNRAS, 387, 295
Baek, S., Di Matteo, P., Semelin, B., Combes, F., & Revaz, Y. 2009, A&A, 495, 389
Barkana, R., & Loeb, A. 2001, Phys. Rep., 349, 125
Bertschinger, E. 1998, ARA&A, 36, 599
Bolton, J. S., & Haehnelt, M. G. 2007, MNRAS, 382, 325
Cen, R. 1992, ApJS, 78, 341
Ciardi, B., Ferrara, A., Marri, S., & Raimondo, G. 2001, MNRAS, 324, 381
Dubroca, B., & Feugeas, J.-L. 1999, C. R. Acad. Sci. Paris, 329, 915
Efstathiou, G., Davis, M., White, S. D. M., & Frenk, C. S. 1985, ApJS, 57, 241
Fan, X., Narayanan, V. K., Strauss, M. A., White, R. L., Becker, R. H., Pentericci, L., & Rix, H. 2002, AJ, 123, 1247
Fan, X., et al. 2006, AJ, 132, 117
Finlator, K., Özel, F., & Davé, R. 2009a, MNRAS, 393, 1090
Finlator, K., Özel, F., Davé, R., & Oppenheimer, B. D. 2009b, MNRAS, 400, 1049
Fromang, S., Hennebelle, P., & Teyssier, R. 2006, A&A, 457, 371
Furlanetto, S. R. 2006, MNRAS, 371, 867
Furlanetto, S. R., Zaldarriaga, M., & Hernquist, L. 2004, ApJ, 613, 1
Gnedin, N. Y. 2000, ApJ, 542, 535
Gnedin, N. Y., & Abel, T. 2001, New Astron., 6, 437
Gnedin, N. Y., & Fan, X. 2006, ApJ, 648, 1
Gnedin, N. Y., & Hui, L. 1998, MNRAS, 296, 44
Gnedin, N. Y., Kravtsov, A. V., & Chen, H. 2008, ApJ, 672, 765
González, M., Audit, E., & Huynh, P. 2007, A&A, 464, 429
Governato, F., et al. 2009, MNRAS, 398, 312
Governato, F., et al. 2010, Nature, 463, 203
Hernquist, L., Bouchet, F. R., & Suto, Y. 1991, ApJS, 75, 231
Hoeft, M., Yepes, G., Gottlöber, S., & Springel, V. 2006, MNRAS, 371, 401
Iliev, I. T., Mellema, G., Pen, U., Merz, H., Shapiro, P. R., & Alvarez, M. A. 2006a, MNRAS, 369, 1625
Iliev, I. T., Mellema, G., Shapiro, P. R., & Pen, U. 2007, MNRAS, 376, 534
Iliev, I. T., Shapiro, P. R., & Raga, A. C. 2005, MNRAS, 361, 405
Iliev, I. T., & The Cosmological Radiative Transfer Comparison Project Collaboration 2009, Mem. Soc. Astron. Ital., 80, 415
Iliev, I. T., et al. 2006b, MNRAS, 371, 1057
Katz, N., Weinberg, D. H., & Hernquist, L. 1996, ApJS, 105, 19
Kohler, K., Gnedin, N. Y., & Hamilton, A. J. S. 2007, ApJ, 657, 15
Komatsu, E., et al. 2009, ApJS, 180, 330
Maselli, A., Ferrara, A., & Ciardi, B. 2003, MNRAS, 345, 379
Mayer, L., Governato, F., & Kaufmann, T. 2008, Adv. Sci. Lett., 1, 7
McQuinn, M., Lidz, A., Zahn, O., Dutta, S., Hernquist, L., & Zaldarriaga, M. 2007, MNRAS, 377, 1043
Miralda-Escudé, J., Haehnelt, M., & Rees, M. J. 2000, ApJ, 530, 1
Oh, S. P., & Furlanetto, S. R. 2005, ApJ, 620, L9
Pawlik, A. H., & Schaye, J. 2008, MNRAS, 389, 651
Petkova, M., & Springel, V. 2009, MNRAS, 396, 1383
Prunet, S., Pichon, C., Aubert, D., Pogosyan, D., Teyssier, R., & Gottloeber, S. 2008, ApJS, 178, 179
Rasera, Y., & Teyssier, R. 2006, A&A, 445, 1
Razoumov, A. O., Norman, M. L., Abel, T., & Scott, D. 2002, ApJ, 572, 695
Schaye, J., & Vecchia, C. D. 2008, MNRAS, 383, 1210
Sengupta, S., Harris, M., Zhang, Y., & Owens, J. D. 2007, Graphics Hardware (New York: ACM), 97
Shin, M., Trac, H., & Cen, R. 2008, ApJ, 681, 756
Shull, J. M., & Venkatesan, A. 2008, ApJ, 685, 1
Songaila, A. 2004, AJ, 127, 2598
Springel, V., & Hernquist, L. 2003, in IAU Symp. 208, Astrophysical Supercomputing using Particle Simulations, ed. J. Makino & P. Hut (Cambridge: Cambridge Univ. Press), 273
Stinson, G., Seth, A., Katz, N., Wadsley, J., Governato, F., & Quinn, T. 2006, MNRAS, 373, 1074
Susa, H. 2006, PASJ, 58, 445
Teyssier, R. 2002, A&A, 385, 337
Teyssier, R., Fromang, S., & Dormy, E. 2006, J. Comput. Phys., 218, 44
Toro, E. F., Spruce, M., & Speares, W. 1994, Shock Waves, 4, 25
Trac, H., & Cen, R. 2007, ApJ, 671, 1
Trac, H., Cen, R., & Loeb, A. 2008, ApJ, 689, L81
Trac, H., & Gnedin, N. Y. 2009, arXiv:0906.4348
Wise, J. H., & Cen, R. 2009, ApJ, 693, 984
Yepes, G., Kates, R., Khokhlov, A., & Klypin, A. 1997, MNRAS, 284, 235