Mem. S.A.It. Vol. 77, 629

Memorie della

The formation of the first black holes and their contribution to the reionization of the intergalactic medium

Z. Haiman[1]

Department of Astronomy, Columbia University, New York, NY 10027, USA
e-mail: `zoltan@astro.columbia.edu`

Abstract. The first massive astrophysical black holes likely formed at high redshifts ($z \gtrsim 10$) at the centers of low mass ($\sim 10^6$ M$_\odot$) dark matter concentrations. These black holes grow by mergers and gas accretion, evolve into the population of bright quasars observed at lower redshifts, and eventually leave the supermassive black hole remnants that are ubiquitous at the centers of galaxies in the nearby universe. The astrophysical processes responsible for the formation of the earliest seed black holes are poorly understood. The purpose of this review is to describe theoretical expectations for the formation and growth of the earliest black holes within the general paradigm of hierarchical cold dark matter cosmologies, and to summarize several relevant recent observations that have implications for the formation of the earliest black holes.

Key words. Cosmology: theory – Cosmology: observations

1. Introduction

It seems established beyond reasonable doubt that some supermassive black holes (SMBHs) were fully assembled early in the history of the universe. The handful of bright quasars at $z \gtrsim 6$ are likely powered by holes as massive as $\sim 10^9$ M$_\odot$, and the spectra and metallicity of these objects appear remarkably similar to their counterparts at moderate redshifts (Fan et al. 2003). Indeed, if one selects individual quasars with the same luminosity, their properties show little evolution with cosmic epoch.

This implies that the behavior of individual quasars is probably determined by local physics near the SMBH and is not directly coupled to the cosmological context in which the SMBH is embedded. However, it is clear that the quasar population as a whole does evolve over cosmic timescales. Observations from $0 \lesssim z \lesssim 6$ in the optical (e.g., the Anglo-Australian Telescope's Two Degree Field, or 2dF, and the Sloan Digital Sky Survey, or SDSS) and radio bands (Shaver et al. 1994) show a pronounced peak in the abundance of bright quasars at $z \approx 2.5$. Recent X-ray observations confirm the rapid rise from $z = 0$ towards $z \approx 2$ for X-ray luminous sources ($L_X > 10^{44}$ ergs s^{-1}; Barger et al. 2003) but have not shown evidence for a decline at still higher redshifts (Miyaji et al. 2000).

The cosmic evolution of quasar black holes between $0 \lesssim z \lesssim 6$ is likely driven by a mechanism other than local physics near the hole.

Send offprint requests to: Z. Haiman

In the cosmological context, it is tempting to link the evolution of quasars with that of dark matter halos condensing in a Cold Dark Matter (CDM) dominated universe, as the halo population naturally evolves on cosmic timescales (Efstathiou & Rees 1988). Indeed, this connection has proven enormously fruitful and has resulted in the following broad picture: the first massive astrophysical black holes appear at high redshifts ($z \gtrsim 10$) in the shallow potential wells of low mass ($\lesssim 10^8$ $M_\odot$) dark matter concentrations. These black holes grow by mergers and gas accretion, evolve into the population of bright quasars observed at lower redshifts, and eventually leave the SMBH remnants that are ubiquitous at the centers of galaxies in the nearby universe.

2. Cosmological Perturbations as the Sites of the First Black Holes

Recent measurements of the Cosmic Microwave Background (CMB) temperature anisotropies by the *Wilkinson Microwave Anisotropy Probe (WMAP)*, determinations of the luminosity distance to distant type Ia Supernovae, and other observations have led to the emergence of a robust "best fit" cosmological model with energy densities in CDM and "dark energy" of $(\Omega_M, \Omega_\Lambda) \approx (0.3, 0.7)$ (e.g., Spergel et al. 2003).

The growth of density fluctuations and their evolution into nonlinear dark matter structures can be followed in this cosmological model from first principles by semi-analytic methods (Press & Schechter 1974; Sheth et al. 2001). More recently, it has become possible to derive accurate dark matter halo mass functions directly in large cosmological N-body simulations (Jenkins et al. 2001). Structure formation in a CDM dominated universe is "bottom-up", with low mass halos condensing first. Dark matter halos with the masses of globular clusters (10^{5-6} $M_\odot$) are predicted to have condensed from $\sim 3\sigma$ peaks of the initial primordial density field as early as $\sim 1\%$ of the current age of the universe, or at redshifts of $z \sim 25$.

It is natural to identify these condensations as the sites where the first astrophysical objects, including the first AGN, were born.

3. Chemistry and Gas Cooling at High Redshifts

Baryonic gas that falls into the earliest nonlinear dark matter halos is shock heated to the characteristic virial temperatures of a few hundred Kelvin. It has long been pointed out (Rees & Ostriker 1977; White & Rees 1978) that such gas needs to lose its thermal energy efficiently (within about a dynamical time) in order to continue contracting, or in order to fragment. In the absence of any dissipation, it would simply reach hydrostatic equilibrium and would eventually be incorporated into a more massive halo further down the halo merger hierarchy. While the formation of nonlinear dark matter halos can be followed from first principles, the cooling and contraction of the baryons, and the ultimate formation of stars or black holes in these halos, is much more difficult to model ab initio.

The gas content of a cosmological perturbation can contract together with the dark matter only in dark halos above the cosmological Jeans mass, $M_J \approx 10^4\ M_\odot[(1+z)/11]^{3/2}$, in which the gravity of dark matter can overwhelm thermal gas pressure. In these early, chemically pristine clouds, radiative cooling is dominated by H_2 molecules. As a result, gas phase H_2 "astrochemistry" is likely to determine the epoch when the first AGN appear (the role of H_2 molecules for early structure formation was reviewed by Abel & Haiman 2001). Several papers have constructed complete gas-phase reaction networks and identified the two possible ways of gas-phase formation of H_2 via the H^- or H_2^+ channels. These were applied to derive the H_2 abundance under densities and temperatures expected in collapsing high redshift objects (Hirasawa 1969; Lepp & Shull 1984; Shapiro & Kang 1987). Studies that incorporate H_2 chemistry into cosmological models and that address issues such as non-equilibrium chemistry, dynamics, or radiative transfer have appeared relatively more recently. Haiman, Thoul, & Loeb (1996) used spherically symmetric simulations to study the masses and redshifts of the earliest objects that can collapse and cool via H_2. The first three dimensional cosmological simulations that incorporate H_2 cooling date back to Gnedin & Ostriker (1996, 1997) and Abel et al. (1997).

The basic picture that emerged from these papers is as follows. The H_2 fraction after recombination in the smooth "protogalactic" gas is small ($x_{H2} = n_{H2}/n_H \sim 10^{-6}$). At high redshifts ($z \gtrsim 100$), H_2 formation is inhibited, even in overdense regions, because the required intermediaries H_2^+ and H^- are dissociated by cosmic "microwave"background (CMB, but with the typical wavelength in the infrared) photons. However, at lower redshifts, when the CMB photons redshift to lower energies, the intermediaries survive, and a sufficiently large H_2 abundance builds up inside collapsed clouds ($x_{H2} \sim 10^{-3}$) at redshifts $z \lesssim 100$ to cause cooling on a timescale shorter than the dynamical time. Sufficient H_2 formation and cooling is possible only if the gas reaches temperatures in excess of ~ 200 K or masses of a few $\times\ 10^5\ M_{\odot}[(1 + z)/11]^{-3/2}$ (note that while the cosmological Jeans mass increases with redshift, the mass corresponding to the cooling threshold, which is well approximated by a fixed virial temperature, has the opposite behavior and decreases at high redshift). The efficient gas cooling in these halos suggests that the first nonlinear objects in the universe were born inside $\sim 10^5\ M_{\odot}$ dark matter halos at redshifts of $z \sim 20$ (corresponding to an $\sim 3\sigma$ peak of the primordial density peak).

The behavior of metal-free gas in such a cosmological "minihalo" is a well posed problem that has recently been addressed in three dimensional numerical simulations (Abel, Bryan, & Norman 2002; Bromm, Coppi, & Larson 2002). These works have been able to follow the contraction of gas to much higher densities than previous studies. They have shown convergence towards a temperature/density regime of $T \sim 200$ K, $n \sim 10^4\ cm^{-3}$, dictated by the critical density at which the excited states of H_2 reach equilibrium and cooling becomes less efficient (Galli & Palla 1998). The 3D simulations suggest that the mass of the gas does not fragment further into clumps below sizes of $10^2 - 10^3\ M_{\odot}$, but rather it forms unusually massive stars. Such stars would naturally leave behind black hole seeds, which can subsequently grow by mergers and accretion into the SMBHs. Interestingly, massive stars have an "either/or" behavior. Nonrotating stars with masses between $\sim 40 - 140\ M_{\odot}$ and above $\sim 260\ M_{\odot}$ collapse directly into a black hole without an explosion, and hence without ejecting their metal yields into the surrounding medium, whereas stars in the range $\sim 140 - 260\ M_{\odot}$ explode without leaving a remnant (Heger et al. 2003). This dichotomy is especially interesting because early massive stars are attractive candidates for polluting the IGM with metals at high redshifts (Madau, Ferrara, & Rees 2001; Wasserburg & Qian 2000). It is likely that the first stars had a range of masses, in which case they could contribute to both metal enrichment and to the seed black hole population, with a relative fraction that depends sensitively on their initial mass function (IMF).

4. Cosmological Reionization: Do the First Black Holes Contribute?

Perhaps the most conspicuous effect of the first generation of light sources, once they collectively reach a critical emissivity of ionizing radiation, is the reionization of the IGM. As has long been known, the absence of strong HI Lyα absorption (i.e., a so-called Gunn-Peterson trough, Gunn & Peterson 1965, hereafter GP) in the spectra of distant sources implies that the IGM is highly ionized (with volume averaged neutral fractions $\lesssim 10^{-4}$) at all redshifts $z \lesssim 6$. There have been two observational breakthroughs recently. On the one hand, there is evidence, from the strong absorption in the spectra of the highest redshift SDSS quasars, that the transition from a neutral to a highly ionized state of hydrogen in the IGM is occurring close to $z \sim 6$ (Becker et al. 2001; Fan et al. 2003; White et al. 2003). On the other hand, the recent detection of a large electron scattering optical depth by the *WMAP* satellite implies that significant ionization had taken place at much higher redshifts ($z \sim 15$, Spergel et al. 2003). There is currently a flurry of activity trying to interpret these results in the context of reionization models (see Haiman 2004 for a recent review). The electron scattering optical depth measured by *WMAP* still has a significant uncertainty, $\tau = 0.17 \pm 0.04$ (Kogut et al. 2003; Spergel et al. 2003). Nevertheless, these developments bring into sharp focus an interesting "old" question: could AGN have contributed to the reionization of the IGM? A natural follow-up question would then be, can we use reionization as a probe of the earliest AGN?

The current *WMAP* results are inconsistent at the 3σ level with a sudden percolation of HII bubbles occurring at $z \sim 6$, which would correspond to the low optical depth of $\tau = 0.04$. This discrepancy is reduced (to the $\sim 2\sigma$ level) even in the simplest models for reionization in which the ionizing emissivity traces the collapse of DM structures. With a reasonable choice of efficiency parameters in such a model, percolation indeed occurs around $z \sim 6$, satisfying the GP trough detections. In such models, there is a natural "tail" of partial ionization, extending to redshifts beyond the percolation epoch, which predicts the total $\tau \sim 0.08$ (Haiman & Holder 2003; Ciardi et al. 2003). However, if the high value of $\tau = 0.17$ is confirmed in future CMB polarization data (e.g., by several additional years of *WMAP* data), the implication will remain: there are additional sources of ionizing radiation at $z \sim 15$. Most importantly, with further improved CMB polarization measurements by *Planck*, the reionization history at high redshifts can be mapped to high precision (Kaplinghat et al. 2003; Holder et al. 2003).

The emissivity of the bright optical quasar population drops steeply at high redshifts ($z \gtrsim 3$; e.g., Fan et al. 2002). There is a hint that the evolution towards high redshifts is flatter in X-rays (Miyaji et al. 2000). While this could be explained if optical quasars were selectively more dust-obscured at high redshifts, this interpretation would fail to explain the sharp decline towards high redshifts that is also seen in the radio (Shaver et al. 1996). If the sharp decline is real, it is easy to show that quasars do not contribute significantly to the ionizing background at $z \gtrsim 6$ (Madau, Haardt, & Rees 1999; Haiman, Abel, & Madau 2001; Fan et al. 2001; Barger et al. 2003) and thus cannot account for the GP troughs detected in the SDSS quasars at this redshift.

However, there is, at least in principle, still room for AGN to contribute to reionization. First, the above ignores the possible presence of faint "miniquasars" (a terminology introduced by Haiman, Madau, & Loeb 1999) below the current detection thresholds. It has been pointed out (Haiman & Loeb 1998b; Haehnelt, Natarajan, & Rees 1998) that there could be a significant population of such faint quasars and that their expected abundance depends crucially on the duty cycle of quasar activity. If the quasar lifetime is short ($\lesssim 10^7$ years), then quasars must reside in intrinsically abundant, low mass halos in order to match their observed surface density on the sky. Conversely, if quasars are long-lived, they must be harbored by the rarer, more massive halos (for the same apparent abundance). The abundance of low mass halos declines less rapidly (and can even increase for $M_{\rm halo} < M_*$) towards high redshifts, and therefore if the quasar duty cycle is short, one expects a larger number of yet-to-be detected "miniquasars". Quasar lifetimes are currently uncertain but are constrained to lie in the range $10^6 - 10^8$ years (see the review by Martini 2004). A particularly relevant method to obtain the lifetime (and thus host halo mass) for the typical quasar at a fixed luminosity is to study the spatial clustering of quasars in large surveys such as 2dF or SDSS (Haiman & Hui 2001; Martini & Weinberg 2001). Current results from 2dF favor $t \lesssim 10^7$ years (see Croom et al. 2004).

In the simple models of Haiman & Loeb (1998) that assume a short quasar lifetime, quasars can reionize the IGM by $z \gtrsim 10$. That model was "calibrated" to reproduce the original observed relation between SMBH mass and bulge mass at $z \sim 0$ by Magorrian et al. (1998). However, the model runs into difficulties with more recent observations: (1) it overproduces the expected counts of faint X-ray sources in the CDFs, and (2) it is no longer consistent with the more recent local SMBH mass estimates (which are reduced by a factor of ∼4) and their steeper dependence on the velocity dispersion $M_\bullet \propto \sigma^{4-5}$ rather than $\propto \sigma^3$. Wyithe & Loeb (2003) recently presented an updated model satisfying these constraints. In their model, the abundance of faint quasars at high redshifts falls short of reionizing the universe at $z \sim 6$.

Despite the above conclusions, it is natural to ask whether the abundance of fainter miniquasars could be higher, and whether they could then significantly contribute to the reionization history. Ricotti & Ostriker (2004) show that such SMBHs can significantly ionize the universe if they contain a fraction $\gtrsim 10^{-5}$ of all baryons. Another example is a large population of intermediate ($\sim 100\,M_\odot$) black holes, which have a harder spectrum and are more efficient ionizers (Madau et al. 2004).

The hard spectra of quasars produce several distinguishing features for reionization (Oh 2001; Venkatesan et al. 2001). Because the mean free path is longer than the Hubble length for photons with energies $\gtrsim [(1+z)/10]^{1/2}$ keV, there is no sharp "edge" for the discrete HII regions surrounding the ionizing sources. As a result, the neutral fraction should decrease gradually throughout most of the entire IGM. This is in sharp contrast with the Swiss cheese topology of reionization by softer photons. Furthermore, X-ray photons deposit a significant fraction ($\sim 1/3$) of their energy into ionizations only, while the IGM is close to neutral. Once the ionized fraction reaches $\sim 30\%$, most of their energy is thermalized with the electrons (e.g., Shull & van Steenberg 1985). As a result, reionization by quasars would be quite different from the stellar case: the IGM would be gradually ionized to the ionized fraction of $\sim 30\%$ (as opposed to suddenly fully ionized). These features make it unlikely that quasars contributed significantly to the sudden elimination of the GP troughs at $z \sim 6$. However, the same features would be attractive in producing partial reionization at high redshifts, and thus would help in explaining the large optical depth measured by *WMAP* (Madau et al. 2004; Ostriker et al. 2003). Note that in this scenario, normal stars would "take over" and dominate the ionizing background at $z \sim 6$, causing the overlap of highly ionized regions. The stars would then concurrently heat the IGM to $\sim 2 \times 10^4$ K. Hui & Haiman (2003) have argued (see also Theuns et al. 2002) that the IGM could not be kept *fully* ionized continuously from $z = 15$ to $z = 4$ because it would then cool adiabatically to a temperature that is below the observed value at $z \sim 4$. The above scenario could naturally avoid this constraint.

We have therefore seen that the first AGN at $z > 6$ could, in principle, still be important contributors to reionization at high redshifts. To conclude this section, we point out yet another potential constraint. At energies above ~ 1 keV (rest frame at $z = 0$), there is little absorption, and whatever radiation was produced by the high redshift quasar population would add cumulatively to the present-day background. Most of the soft X-ray background has already been resolved into low redshift sources (Mushotzky et al. 2000; see also Wu & Xue 2001 and references therein). Dijkstra, Haiman, & Loeb (2004a) find that the putative high redshift quasars, if they are to fully reionize the IGM, would overproduce the soft X-ray background. However, distant miniquasars that produce enough X–rays to only partially ionize the IGM to a level of at most $x_e \sim 50\%$ are still allowed.

Acknowledgements. I am grateful to the organizers of the conference for their kind invitation and for hosting an enjoyable workshop. I also thank my recent collaborators Lam Hui and Eliot Quataert for many fruitful discussions. I acknowledge financial support by NSF through grants AST-0307200 and AST-0307291 and by NASA through grant NAG5-26029.

References

Abel, T., Anninos, P., Zhang, Y., Norman, M. L. 1997, NewA, 2, 181

Abel, T., Bryan, G. L., & Norman, M. L. 2002, Science, 295, 93

Abel, T., & Haiman, Z. 2001, in "Molecular Hydrogen in Space", Cambridge Contemporary Astrophysics, Eds. F. Combes & G. Pineau des Forrets. (Cambridge, U.K.: Cambridge University Press), p237

Barger, A. J., et al. 2003, ApJ, 584, L61

Becker, R. H., et al. 2001, AJ, 122, 2850
Bromm, V., Coppi, P. S., & Larson, R. B. 2002, ApJ, 564, 23
Ciardi, B., Ferrara, A., & White, S. D. M. 2003, MNRAS, 344, 7
Croom, S., et al. 2004, in ”AGN Physics with SDSS”, Eds. G. T. Richards & P. B. Hall, (astro-ph/0310533)
Dijkstra, M., Haiman, Z., & Loeb, A. 2004, ApJ, 613, 646
Efstathiou, G., & Rees, M. J. 1988, MNRAS, 230, 5
Fan, X., et al. 2003, AJ, 125, 1649
Galli, D., & Palla, F. 1998, A&A, 335, 403
Gnedin, N. Y., & Ostriker, J. P. 1996, ApJ, 472, 63
Gnedin, N. Y., & Ostriker, J. P. 1997, ApJ, 486, 581
Gunn, J. E., & Peterson, B. A. 1965, ApJ, 142, 1633
Haehnelt, M. G., Natarajan, P., & Rees, M.J. 1998, MNRAS, 300, 827
Haiman, Z. 2004, in “Coevolution of Black Holes and Galaxies”, Carnegie Observatories Astrophysics Series, Vol. 1, Ed. L. C. Ho. (Cambridge, U.K.: Cambridge University Press), in press (astro-ph/0304131)
Haiman, Z., & Holder, G. P. 2003, ApJ, 595, 1
Haiman, Z., & Hui, L. 2001, ApJ, 547, 27
Haiman, Z., & Loeb, A. 1998, ApJ, 503, 505
Haiman, Z., Madau, P., & Loeb, A. 1999, ApJ, 514, 535
Haiman, Z., Thoul, A. A., & Loeb, A. 1996, ApJ, 464, 523
Heger, A., et al. 2003, ApJ, 591, 288
Hirasawa, T. 1969, Prog. Theor. Phys., 42(3), 523
Hui, L., & Haiman, Z. 2003, ApJ, 596, 9
Jenkins, A., et al. 2001, MNRAS, 321, 372
Kaplinghat, M., Chu, M., Haiman, Z., Holder, G., Knox, L., & Skordis, C. 2003, ApJ, 583, 24
Kogut, A., et al. 2003, ApJS, 148, 161
Lepp, S., & Shull, J. M. 1984, ApJ, 280, 465
Madau, P., Ferrara, A., & Rees, M. J. 2001, ApJ, 555, 92
Madau, P., Haardt, F., & Rees, M. J. 1999, ApJ, 514, 648
Madau, P., et al. 2004, ApJ, in press (astro-ph/0310223)
Martini, P. 2004 in “Coevolution of Black Holes and Galaxies”, Carnegie Observatories Astrophysics Series, Vol. 1, Ed. L. C. Ho. (Cambridge, U.K.: Cambridge University Press), in press (astro-ph/0304009)
Martini, P., & Weinberg, D. H. 2001, ApJ, 547, 12
Miralda-Escudé, J., & Rees, M. J. 1993, MNRAS, 260, 617
Press, W. H., & Schechter, P. L. 1974, ApJ, 181, 425
Ricotti, M., & Ostriker, J. P. 2004, MNRAS, in press (astro-ph/0311003)
Rees, M. J., & Ostriker, J. P. 1977, ApJ, 179, 541
Shapiro, P. R., & Kang, H. 1987, ApJ, 318, 32
Shaver, P. A., et al. 1996, Nature, 384, 439
Sheth, R. K., Mo, H. J., & Tormen, G. 2001, MNRAS, 323, 1
Shull, J. M., & van Steenberg, M. E. 1985, ApJ, 298, 268
Spergel, D. N., et al. 2003, ApJS, 148, 175
Theuns, T., et al. 2002, ApJL, 567, 103
Wasserburg, G. J., & Qian, Y.-Z. 2000, ApJ, 538, L99
White, R. L., Becker, R. H., Fan, X., & Strauss, M. A. 2003, AJ, 126, 1
White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341
Wu, X., & Xue, Y. 2001, ApJ, 560, 544
Wyithe, S., & Loeb, A. 2003, ApJ, 595, 614