Mon. Not. R. Astron. Soc. **370,** 289–298 (2006) doi:10.1111/j.1365-2966.2006.10467.x

Formation of supermassive black holes by direct collapse in pre-galactic haloes

Mitchell C. Begelman,[1,2,3⋆] Marta Volonteri[3⋆] and Martin J. Rees[3⋆]

[1]*JILA, University of Colorado, Boulder, CO 80309-0440, USA*
[2]*Department of Astrophysical and Planetary Sciences, University of Colorado at Boulder, Boulder, CO, USA*
[3]*Institute of Astronomy, Madingley Road, Cambridge CB3 0HA*

Accepted 2006 April 21. Received 2006 April 21; in original form 2006 February 16

ABSTRACT

We describe a mechanism by which supermassive black holes (SMBHs) can form directly in the nuclei of protogalaxies, without the need for 'seed' black holes left over from early star formation. Self-gravitating gas in dark matter haloes can lose angular momentum rapidly via runaway, global dynamical instabilities, the so-called 'bars within bars' mechanism. This leads to the rapid build-up of a dense, self-gravitating core supported by gas pressure – surrounded by a radiation pressure-dominated envelope – which gradually contracts and is compressed further by subsequent infall. We show that these conditions lead to such high temperatures in the central region that the gas cools catastrophically by thermal neutrino emission, leading to the formation and rapid growth of a central black hole.

We estimate the initial mass and growth rate of the black hole for typical conditions in metal-free haloes with $T_{\rm vir} \sim 10^4$ K, which are the most likely to be susceptible to runaway infall. The initial black hole should have a mass of $\lesssim$20 $M_\odot$, but in principle could grow at a super-Eddington rate until it reaches $\sim 10^4$–10^6 $M_\odot$. Rapid growth may be limited by feedback from the accretion process and/or disruption of the mass supply by star formation or halo mergers. Even if super-Eddington growth stops at $\sim 10^3$–10^4 $M_\odot$, this process would give black holes ample time to attain quasar-size masses by a redshift of 6, and could also provide the seeds for all SMBHs seen in the present Universe.

Key words: accretion, accretion discs – black hole physics – hydrodynamics – instabilities – galaxies: formation – cosmology: theory.

1 INTRODUCTION

Several scenarios have been presented for the formation and growth of supermassive black holes (SMBHs) in the nuclei of galaxies. One possible route traces the black hole progenitors back to the first generation of stars. The first stars formed out of metal-free gas, with the lack of an efficient cooling mechanism possibly leading to a very top-heavy initial stellar mass function (Carr, Bond & Arnett 1984; Larson 1998). Numerical simulations of the fragmentation of primordial clouds in standard cold dark matter (CDM) theories suggest that Pop III stars were indeed very massive (Bromm, Coppi & Larson 1999; Abel, Bryan & Norman 2000; Bromm, Coppi & Larson 2002), and would have left behind black hole 'seeds' of anywhere from tens to several hundred solar masses. The main features of a plausible scenario for the hierarchical assembly, growth and dynamics of massive black holes from such seeds have been discussed most recently by Volonteri, Haardt & Madau (2003), Volonteri et al. (2005) and Volonteri & Rees (2005).

Another family of models for massive black hole formation is based on the collapse of supermassive objects formed directly out of dense gas (Haehnelt & Rees 1993; Umemura, Loeb & Turner 1993; Loeb & Rasio 1994; Eisenstein & Loeb 1995; Bromm & Loeb 2003; Koushiappas, Bullock & Dekel 2004). The main challenge for these models is the disposal of angular momentum. Eisenstein & Loeb (1995) and Koushiappas et al. (2004) investigated the formation of black holes from low angular momentum material, either in haloes with extremely low angular momentum (Eisenstein & Loeb 1995), or by considering only the low angular momentum tail of material in haloes with efficient gas cooling. But even in these models, as in all the others, substantial angular momentum transport is required in order for the gas to form a central massive object, which ultimately collapses as a result of the post-Newtonian gravitational instability. Various angular momentum transport mechanisms have been invoked, including radiation drag against the cosmic microwave

⋆E-mail: mitch@jila.colorado.edu (MCB); marta@ast.cam.ac.uk (MV); mjr@ast.cam.ac.uk (MJR)

background (at very high redshifts: Umemura et al. 1993), viscosity driven by magnetic fields or turbulence, Rossby waves (Colgate et al. 2003) and self-gravitational instabilities.

The scenario we investigate here is related to the second family of models, and focuses on the outcome of global dynamical instabilities driven by self-gravity, the so-called 'bars within bars' mechanism (Shlosman, Frank & Begelman 1989; Shlosman, Begelman & Frank 1990). Self-gravitating gas clouds become bar-unstable when the level of rotational support surpasses a certain threshold. A bar can transport angular momentum outwards on a dynamical time-scale via gravitational and hydrodynamical torques, allowing the radius to shrink. Provided that the gas is able to cool, this shrinkage leads to even greater instability, on shorter time-scales, and the process cascades. This mechanism is a very attractive candidate for collecting gas in the centres of haloes, because it works on a dynamical time and can operate over many decades of radius. In contrast to the formation of a supermassive 'star', with high entropy throughout, we show that the 'bars within bars' mechanism produces a 'quasi-star' with a very low specific entropy near the centre. As a result, the initial core collapse leading to black hole formation involves only a few solar masses, rather than the thousands of solar masses usually associated with direct collapse models. Despite this modest beginning, accretion from the envelope surrounding the collapsed core can build up a substantial black hole mass very rapidly – possibly at a highly super-Eddington rate.

The plan of the paper is as follows. In Section 2, we discuss the criterion for global gravitational instability and apply it to the gas in dark matter haloes with a realistic distribution of angular momentum parameters. If more than a few tenths of the baryonic matter fall towards the centre of the halo, then gravitational instability should be very common. But even an infalling fraction of $\sim$10 per cent can lead to an interesting number of unstable haloes. In Section 3, we specialize to haloes with virial temperatures $T_{\rm vir} \gtrsim 10^4$ K and metal-free gas. The bars within bars scenario makes specific predictions about the radial distribution of infalling gas and the associated circular velocity, which goes from constant in the outer parts of the inflow to quasi-Keplerian close-in. We first discuss the infall process neglecting star formation, and then show how the process is modified (but not necessarily halted) if a fraction of the infalling gas at each radius forms stars. The gravitational binding energy liberated by infalling gas increases steadily with decreasing radius, until the luminosity exceeds the Eddington limit, the infalling stalls and a radiation pressure-supported 'quasi-star' forms (Section 4). The radius of the quasi-star is a few astronomical units, a scale that does not change even as the quasi-star grows in mass.

We show that the quasi-star has a positive specific entropy gradient, and that gas pressure remains important in the core of the quasi-star. The temperature of this core steadily increases as matter piles on to the quasi-star (Section 5), eventually approaching 10^9 K, at which point it undergoes catastrophic cooling and collapse by thermal neutrino emission (Section 6). We argue that this leads to the formation of a black hole of $\sim$10–20 $M_\odot$, which may then grow at rate that greatly exceeds the Eddington limit (Section 7). This rapid growth could produce a black hole of several million solar masses, although feedback and depletion of the mass supply could quench the growth rate at an earlier stage. We discuss the co-evolution of the black holes and their hosts and the global impact of the black hole population in Section 8. We conclude by discussing the implications of this model for the interpretation of high-z quasars, the statistics of black hole masses in the local universe, and its relevance to other astrophysical situations where black holes could grow at a very rapid rate.

Unless otherwise stated, all results shown below refer to the currently favoured ΛCDM world model with $\Omega_{\rm M} = 0.3$, $\Omega_\Lambda = 0.7$, $h = 0.7$, $\Omega_{\rm b} = 0.045$, $\sigma_8 = 0.93$ and $n = 1$.

2 CONDITIONS FOR RUNAWAY COLLAPSE

We focus here mainly on the dynamical stability of the gas in haloes with virial temperatures $T_{\rm vir} \gtrsim 10^4$ K. Runaway collapse could also occur in smaller haloes, provided that molecular hydrogen cooling is efficient and gas can cool well below the virial temperature. In the absence of molecular hydrogen, gas in haloes with $T_{\rm vir} < 10^4$ K would tend to remain less dense than the dark matter; tidal forces would then prevent widespread collapse and fragmentation at this stage. Since cooling and collapse of the gas is more likely in large haloes, and the masses involved are larger, we henceforth refer to haloes with virial temperatures $T_{\rm vir} \gtrsim 10^4$ K, unless otherwise stated. We stress nevertheless that runaway collapse is not completely ruled out in smaller systems at early times, well before the first generation of stars created a photodissociating background.

Bromm & Loeb (2003) show that if molecular hydrogen formation is suppressed in haloes with $T_{\rm vir} > 10^4$ K, the gas tends to condense into massive clumps in the centre. The gaseous component of these haloes can cool even in the absence of H_2 via neutral hydrogen atomic lines to $\sim$8000 K, and contract nearly isothermally (Oh & Haiman 2002). These massive clumps do not fragment as long as molecular hydrogen remains unimportant. One way to hinder the formation of molecular hydrogen is the presence of a dissociating background (Haiman, Abel & Rees 2000; Oh & Haiman 2002; Bromm & Loeb 2003; but see Machacek, Bryan & Abel 2001). It is therefore possible that the formation of seed black holes in massive haloes follows an earlier epoch of star formation. We consider redshifts high enough that a large fraction of gas is still unenriched by metals, or very lightly polluted, so that metal line cooling is still unimportant (Santoro & Shull 2006).

We assume that the baryons preserve their specific angular momentum during collapse (Mo, Mao & White 1998), and settle into a rotationally supported disc at the centre of the halo (Mo et al. 1998; Oh & Haiman 2002). Flattened systems can be subject to dynamical and secular instabilities, even when embedded in external haloes (Fall & Efstathiou 1980). Several instability criteria have been investigated (e.g. Ostriker & Peebles 1973; Efstathiou, Lake & Negroponte 1982; Christodoulou, Shlosman & Tohline 1995), which determine the maximum rotational energy (or angular momentum) that a system can possess and still be stable against bar-like instabilities. Christodoulou et al. (1995) propose a simple, but robust, criterion for stability which can be expressed as

$$\alpha = \left(\frac{1}{2} f \frac{T}{|W|}\right)^{1/2} < 0.34, \qquad (1)$$

where T is the rotational kinetic energy, W is the gravitational potential energy and f is a parameter dependent on the geometry of the system, with $f = 1$ for discs.

Numerical simulations have not yet clarified the detailed dynamics of gaseous collapse in haloes, and we explore here three different models for self-gravitating gas discs. We assume that the disc has either constant circular velocity (Mestel discs: Mestel 1963) or constant angular velocity (rigid body rotation), or that the gas settles down into a classical exponential disc. We embed the gaseous discs into a halo of mass $M_{\rm h}$, virial radius $R_{\rm vir}$, and virial temperature $T_{\rm vir}$, described by a Navarro, Frenk & White (1997, hereafter NFW) dark matter density profile, with a spin parameter $\lambda_{\rm spin} (\equiv J_{\rm h} E_{\rm h}^{1/2} / G M_{\rm h}^{5/2}$,

where $J_{\rm h}$ is the total angular momentum and $E_{\rm h}$ is the binding energy). We recall that, within the spherical collapse model, the mass of a halo, at a given redshift of formation, scales with the virial temperature as $M_{\rm h} \simeq 10^4 \Delta_{\rm vir}^{-1/2} T_{\rm vir}^{-3/2}\,{\rm M}_\odot$, where $\Delta_{\rm vir}$ is the virial density in units of the critical density.

We determine the characteristics of the gaseous discs via a procedure similar to that of Mo et al. (1998). We then apply the criterion given by Christodoulou et al. (1995) in order to determine the stability of the modelled systems. Stability depends on two parameters: the halo spin parameter $\lambda_{\rm spin}$ and the fraction of baryonic matter that ends up in the disc, $f_{\rm d} = (\Omega_{\rm M}/\Omega_{\rm b})\,(M_{\rm disc}/M_{\rm h})$. The results can be understood qualitatively by approximating the disc kinetic energy as $T_{\rm disc} \approx 0.5\, M_{\rm disc}\, V_{\rm c}^2(R_{\rm disc})$, where $R_{\rm disc}$ is the scalelength of the disc,[1] which can be determined under the assumption that the collapsing baryons conserve angular momentum. If we ignore the contribution of the halo to the circular velocity, for the three cases we find

$$T_{\rm disc,Mestel} \approx \frac{\pi^2\, G\, M_{\rm disc}^2 f_{\rm d}(\Omega_{\rm b}/\Omega_{\rm M})}{32\,\lambda_{\rm spin}^2\, R_{\rm vir}}, \tag{2}$$

$$T_{\rm disc,rigid} \approx \frac{9\,\pi^2\, G\, M_{\rm disc}^2 f_{\rm d}(\Omega_{\rm b}/\Omega_{\rm M})}{40\,\lambda_{\rm spin}^2\, R_{\rm vir}}, \tag{3}$$

$$T_{\rm disc,exp} \approx \frac{G\, M_{\rm disc}^2}{\lambda_{\rm spin}\, R_{\rm vir}\, f_R}, \tag{4}$$

where

$$f_R \approx 1 - 3 f_{\rm d}\left(\frac{\Omega_{\rm b}}{\Omega_{\rm M}}\right) + 5.2 f_{\rm d}^2 \left(\frac{\Omega_{\rm b}}{\Omega_{\rm M}}\right)^2 \tag{5}$$

(see Mo et al. 1998, for the exact expression). At fixed $\lambda_{\rm spin}$ and $f_{\rm d}$, $T_{\rm disc,rigid} > T_{\rm disc,Mestel} > T_{\rm disc,exp}$. The total kinetic ($T$) and potential ($|W|$) energies of the systems, including the contribution and stabilizing effect of the NFW halo, increase with $\lambda_{\rm spin}$ due to the halo contribution within $R_{\rm disc}$, which increases with $\lambda_{\rm spin}$. The ratio $T/|W|$, nevertheless, decreases, due to an increasingly dominant halo contribution.

The stability results are summarized in Fig. 1, in which the maximum spin parameter $\lambda_{\rm spin,max}$ for which a disc is unstable is shown as a function of the fraction of baryons forming the disc, i.e. for every $f_{\rm d}$, discs are stable for $\lambda_{\rm spin} > \lambda_{\rm spin,max}$.

The distribution of spin parameters found in numerical simulations is well fit by a lognormal distribution in $\lambda_{\rm spin}$, with mean $\bar{\lambda}_{\rm spin} = 0.05$ and s.d. $\sigma_\lambda = 0.5$:

$$p(\lambda)\,{\rm d}\lambda = \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp\left[-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_\lambda^2}\right]\frac{{\rm d}\lambda}{\lambda}. \tag{6}$$

This function is a good fit to the N-body results of Warren et al. (1992). Similar results were found in later investigations (e.g. Cole & Lacey 1996; Bullock et al. 2001; van den Bosch et al. 2002). With this assumption we can estimate the fraction of discs subject to dynamical instability, as a function of $f_{\rm d}$, for each of the three disc models (Fig. 1).

3 STRUCTURE AND EVOLUTION OF COLLAPSING GAS

The unstable conditions described in Section 2 are expected to lead to runaway infall, provided that the gas remains cooler than the local virial temperature as it collapses. For the densities expected in pre-galactic haloes, the cooling time to $\lesssim 10^4$ K is much smaller than the dynamical time; this ordering is preserved as the collapse proceeds. At the initial disc radius, the gravitational potential due to the gas is already appreciable compared to that of the dark matter. As we will see below, collapse leads to a mean gas density profile at least as steep as r^{-2}, implying a virial temperature that remains roughly constant or increases with decreasing r, whereas the dark matter density is expected to increase only as $\propto r^{-1}$ in the inner parts of the halo (NFW). The relative dominance of gas self-gravity over the dark matter potential thus increases as the gas collects towards the centre, implying that the conditions for large-scale gravitational instability intensify with decreasing radius. Conditions are therefore ideal for the 'bars within bars' instability. We will henceforth neglect the dark matter.

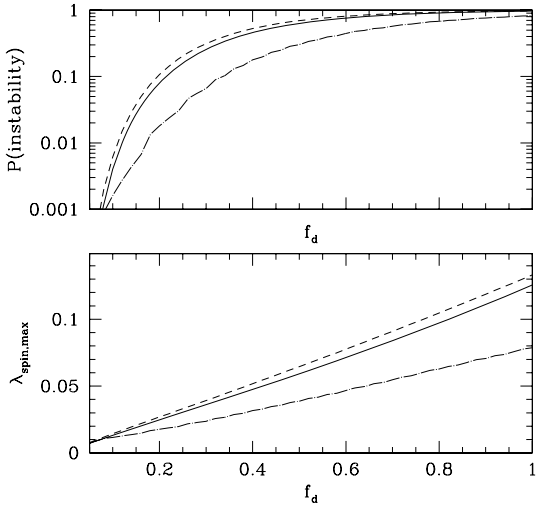

Figure 1. Bottom panel: maximum spin parameter, $\lambda_{\rm spin,max}$, for disc instability as a function of the gas fraction ending up in the disc. Discs are stable for $\lambda_{\rm spin} > \lambda_{\rm spin,max}$. Solid line: Mestel disc, dashed line: rigid disc, dot–dashed line: exponential disc. Upper panel: fraction of unstable discs, for each of the three disc models.

The collapse of a self-gravitating, isothermal gas cloud has been analysed in both the non-rotating (Larson 1969; Penston 1969; Shu 1977) and rotating but inviscid (Saigo & Hanawa 1998, and references therein) limits. In all cases, the outer part of the flow evolves towards the density profile of a singular isothermal sphere, $\rho \propto r^{-2}$. Mineshige & Umemura (1997), analysing the case of a shrinking, self-gravitating accretion disc subject to an α-viscosity, found analogous behaviour, i.e. a surface density distribution $\Sigma \propto r^{-1}$ in the outer regions. Using smoothed particle hydrodynamics simulations, Englmaier & Shlosman (2004) show that a rotating, self-gravitating gas cloud can decouple dynamically from a larger stellar bar, thus verifying the basic 'bars within bars' picture. They find that the shrinking gas bar develops a strong density gradient with radius, and an angular pattern speed $\Omega \propto a^{-1}$, where a is the semimajor axis of the gaseous bar. These features are also consistent with an r^{-2} radial density distribution.

The $\rho \propto r^{-2}$ behaviour can be understood as follows. Suppose the initial mass and radius of the cloud are M_0 and r_0, respectively, corresponding to a virial speed $v_0 = (GM_0/r_0)^{1/2} \sim 10 v_{10}\,{\rm km\,s^{-1}}$. At large radii and early times, the evolutionary time-scale of the flow is set by the free-fall time at r_0, $t_0 \sim r_0/v_0$. At smaller radii the free-fall time is shorter, but matter is being fed into these radii on the much longer time-scale t_0. Therefore, the mass flux through

[1] Note that both the Mestel disc and the rigid disc are defined only for $R < R_{\rm disc}$ (Mestel 1963).

these regions must be roughly constant and independent of radius,

$$\dot{M}(r) = \dot{M}_0 \sim \frac{M_0}{t_0} \sim \frac{v_0^3}{G} = 0.2\, v_{10}^3\ \mathrm{M_\odot\ yr^{-1}}. \quad (7)$$

On average, the self-gravitating gas at every radius gets rid of its angular momentum on the local free-fall time, $t_{\rm ff}(r) \sim [GM(r)/r^3]^{-1/2}$, where $M(r)$ is the mass contained within a radius r. The mass flow rate is therefore $\dot{M}_0 \sim M(r)/t_{\rm ff}(r) \propto [M(r)/r]^{3/2}$. Since $\dot{M}_0$ is constant, we have $M(r) \propto r$ and $\rho \propto r^{-2}$.

This behaviour can change if the collapsing gas fragments and forms stars. A self-gravitating cluster of collisionless particles would increase its velocity dispersion in response to the bar potential, thus quenching the instability. Although the cooling time-scale of the gas is shorter than the dynamical time, and therefore violates the Gammie (2001) criterion for avoiding fragmentation, we do not expect fragmentation to deplete a large fraction of the inflow. This is because the circular speed in the potential of the self-gravitating gas remains roughly constant at $r < r_0$, and corresponds to a temperature that is close to the virial temperature of the halo. Since we consider only haloes with $T_{\rm vir} \gtrsim 10^4$ K, the gas never cools very far below $T_{\rm vir}$ and therefore does not form a very thin sheet (i.e. the Toomre 1964 parameter Q does not become extremely small). The Jeans mass under such conditions is only a few times smaller than $M(r)$, suggesting that fragmentation will be inefficient. A corollary of this argument is that the efficient collection of gas in the centre of a halo might occur only for a relatively narrow range of $T_{\rm vir} \gtrsim 10^4$ K, and only under metal-free conditions where the formation of H_2 is inhibited (these arguments were discussed at length by Bromm & Loeb 2003), or for gas enriched below the critical metallicity threshold for fragmentation (Santoro & Shull 2006).

As noted in many earlier works, the r^{-2} density profile cannot persist all the way to the centre of the collapsing cloud (Shu 1977; Mineshige & Umemura 1997; Saigo & Hanawa 1998). After a time t, gas collecting at a rate $\dot{M}_0$ will dominate the potential out to a radius $r_1(t) \sim (t/t_0)\, r_0$. An r^{-2} density distribution, with $M(r) \propto r$, is not globally self-gravitating inside r_1, and therefore cannot drive the bars within bars instability. In non-rotating collapse (Shu 1977) the accumulated gas behaves like a point mass and the density at $r < r_1$ scales as $\rho \propto r^{-3/2}$, as in Bondi (1952) accretion. In inviscid collapse of a rotating fluid (Saigo & Hanawa 1998), the centrifugal barrier forces the density to be roughly constant at small radii.

Our model involves effective transport of angular momentum, and in this respect is closer to the case considered by Mineshige & Umemura (1997). Assuming an α-viscosity with fixed α, they find that the surface density profile in the inner region steepens to $\Sigma \propto r^{-5/3}$, corresponding to a steep mean density distribution $\rho \propto r^{-8/3}$. The corresponding inflow rate scales as $\dot{M} \propto r^{1/3}$, and the inflow speed, $v \propto r$, falls far below the free-fall speed ($\propto r^{-1/3}$). We also expect self-similar settling of gas at a mean speed $v(r) \sim r/t \ll v_0$ for $r \ll r_1$. If we assume that angular momentum transfer is governed by global self-gravitational instabilities, instead of an α-viscosity, then the gas must adopt a configuration where these instabilities are nearly quenched in order to transfer angular momentum on a time-scale much longer than the dynamical time.

The conditions for global instability depend on the details of the gravitational potential as well as the radial distributions of density, pressure and angular momentum. Because of the rapid cooling, gas pressure is negligible and we expect the system to remain strongly unstable at all radii where it is substantially self-gravitating (Shlosman et al. 1989). The only way to suppress the instability, it seems, is for the density distribution to become sufficiently centrally concentrated that a large fraction of the gravitational potential at each r is generated by the gas at much smaller radii. This requires the mean surface density distribution to steepen to $\Sigma \propto r^{-2}$, corresponding to a mean density $\rho \propto r^{-3}$, so that there are roughly equal amounts of mass per decade of radius. Note that this is slightly steeper than the density distribution obtained by Mineshige & Umemura (1997).

The above argument assumes that angular momentum continues to be transported by global gravitational instabilities, *and* that fragmentation continues to be unimportant. The latter assumption is much less secure at $r < r_1$ than it is further out. If the gas remains isothermal at $\sim 10^4$ K, the Toomre Q-parameter would decrease $\propto r^{1/2}$ and the inflowing gas would form a thin disc. Thus, fragmentation could seriously hamper the bars within bars instability at $r < r_1$. If fragmentation is highly efficient, then the inflowing gas might simply lay down an isothermal stellar potential with a constant velocity dispersion $\sim v_0$. Little gas would be deposited inside $r_1(t)$.

It seems unlikely, however, that the outcome is this extreme. Fragmentation should not deplete the gas density much below the threshold for *local* gravitational instability, $Q \sim 1$. Gas with the corresponding density, and a sound speed $\sim v_0$, would continue to accrete at a rate $\alpha \dot{M}_0$, where α is a viscosity parameter (Shlosman et al. 1990). Even where fragmentation is suppressed, local non-axisymmetric gravitational instabilities could continue to drive angular momentum transport with an effective $\alpha \sim O(1)$. Other sources of angular momentum transport probably operate as well, such as turbulence stirred up by the fragmentation and star formation itself. Therefore, although we are not able to predict exactly how much gas makes it all the way to the central region of the halo, we are probably justified in parametrizing the surviving mass flux as $\alpha \dot{M}_0$, with α assumed to be $\sim O(1)$.

4 CREATION AND GROWTH OF A 'QUASI-STAR'

Disc accretion persists as long as the infalling gas is able to radiate away the liberated binding energy. Given an accumulated mass of $M_*(t) \sim \alpha \dot{M}_0 t$ at $r \ll r_1$, we find that the luminosity generated outside a radius r is given by

$$L(r,t) \sim \alpha \dot{M}_0 \frac{GM_*}{r} \sim \alpha^2 \frac{v_0^5}{G} \left[\frac{r}{r_1(t)}\right]^{-1}. \quad (8)$$

Within a certain radius this radiation is trapped, the pressure builds up and the gas inflates into a pressure-supported cloud, which we dub a 'quasi-star'. Since rotational support no longer dominates, we assume that the self-gravitational instabilities are finally quenched. The condition for radiation trapping is given by

$$L(r,t) > L_{\rm Edd}(t) \left(1 + \frac{p_{\rm gas}}{p_{\rm rad}}\right)^{-1}, \quad (9)$$

where $L_{\rm Edd}(t) \sim 4\pi\alpha c G \dot{M}_0 t/\kappa$ is the Eddington limit, given the appropriate opacity κ, for the accumulated mass $M_*(t)$. $p_{\rm gas}$ and $p_{\rm rad}$ are the gas and radiation pressure, respectively, in the quasi-star. Once the quasi-star mass exceeds a few solar masses (i.e. very early in its growth, since the mass is growing at $\sim 0.2\, \alpha v_{10}^3\ \mathrm{M_\odot\ yr^{-1}}$), the mean local thermodynamic equilibrium (LTE) radiation pressure exceeds the gas pressure; we will henceforth assume $p_{\rm rad} \gg p_{\rm gas}$. The interior of the quasi-star is hot enough to ionize hydrogen, allowing us to assume that the opacity is dominated by electron scattering. Substituting $r_1(t) \sim (t/t_0)\, r_0$, we find that the radius of

the quasi-star is time independent, and is given by

$$r_* \sim \frac{\alpha\kappa v_0^3}{4\pi G c} = 1.6 \times 10^{13} \alpha v_{10}^3 \text{ cm}. \quad (10)$$

Thus, the quasi-star that is going to give rise to a SMBH has a radius of 1 au, for $\alpha v_{10}^3 \sim 1$.

Conditions in the interior of the quasi-star are extremely sensitive to the mass inflow rate, which we express through its dependence on α and v_0. Denoting the quasi-stellar mass by $M_* = m_* \, \mathrm{M}_\odot \propto t$, we find the mean density $\rho_* \sim 10^{-7} \alpha^{-3} v_{10}^{-9} \, m_* \, \mathrm{g \, cm^{-3}}$, mean pressure $p_* \sim 10^6 \alpha^{-4} v_{10}^{-12} m_*^2 \, \mathrm{erg \, cm^{-3}}$, and mean temperature (in LTE) $T_* \sim 1.5 \times 10^5 \alpha^{-1} v_{10}^{-3} \, m_*^{1/2}$ K. These estimates justify our assumptions that $p_{\rm rad} \gg p_{\rm gas}$ for $m_* >$ a few, and that the opacity is dominated by electron scattering.

5 INTERIOR STRUCTURE AND EVOLUTION OF THE QUASI-STAR

The characteristic specific entropy of the matter joining the quasi-star increases with time, $s_* \equiv p_*/\rho_*^{4/3} \propto M_*^{2/3} \propto t^{2/3}$. Since hydrostatic equilibrium demands $p \propto \rho^2 r^2 \propto \rho^{4/3} M(r)^{2/3}$ in the quasi-stellar interior, where $M(r)$ is the mass contained within a radius r, we conclude that each layer of matter added to the quasi-star approximately conserves its specific entropy as the quasi-star grows. The positive entropy gradient, $s(r) \propto M(r)^{2/3} \propto r^{2/3}$, stabilizes the quasi-star against convection, which would otherwise tend to homogenize the entropy. This implies that each layer of the stellar interior is compressed by overlying material until the radiative diffusion time across the layer, $t_{\rm diff}(r) \sim \rho\kappa r^2/c$, is roughly equal to the age of the quasi-star.

The interior structure of the quasi-star is therefore characterized by a density profile $\rho \sim \rho_*(r/r_*)^{-2}$, pressure profile $p \sim p_*(r/r_*)^{-2}$, and temperature profile $T \sim T_*(r/r_*)^{-1/2}$.

These scalings apply as long as radiation pressure exceeds gas pressure. However, the decreasing specific entropy towards the centre implies that the ratio of radiation pressure to gas pressure decreases with decreasing r,

$$\frac{p_{\rm rad}}{p_{\rm gas}} \sim m_*^{1/2} \left(\frac{r}{r_*}\right)^{1/2}, \quad (11)$$

implying that $p_{\rm gas} \sim p_{\rm rad}$ at small enough radii. The gas pressure-dominated core comprises the earliest material laid down during the growth of the quasi-star. It has a radius $r_{\rm c} \sim r_* \, m_*^{-1} \sim 1.5 \times 10^{13} \alpha v_{10}^3 m_*^{-1}$ cm, temperature $T_{\rm c} \sim 1.5 \times 10^5 \alpha^{-1} v_{10}^{-3} \, m_*$ K, and density $\rho_{\rm c} \sim 10^{-7} \alpha^{-3} v_{10}^{-9} \, m_*^3 \, \mathrm{g \, cm^{-3}}$. The core mass is independent of M_*, α and v_{10}, and is roughly $1 \, \mathrm{M}_\odot$.

Nuclear burning commences when the core temperature reaches $T_{\rm c} \sim 10^6$–10^7 K, for a quasi-star mass $m_* \sim$ (10–100)αv_{10}^3. At this point the core density is not that dissimilar to densities inside main sequence stars, so the burning time-scales are likely to be similar as well. The evolution time-scale due to infall is much shorter, $t_{\rm ev} = M_*/\dot{M}_0 \sim 5\alpha^{-1} v_{10}^{-3} m_*$ yr, so we do not expect nuclear burning to progress very far until the core temperature approaches $\sim 10^8$ K, for $m_* \sim 10^3 \, \alpha v_{10}^3$. At this point the gravitational binding energy of the quasi-star is $\sim 10^{13} \alpha^{-1} v_{10}^{-3} \, m_* \, \mathrm{erg \, g^{-1}}$ while the available nuclear binding energy is $\lesssim 6 \times 10^{18} m_*^{-1} \, \mathrm{erg \, g^{-1}}$, if burning is confined to the core. In order for the nuclear energy release to overpower the gravitational binding energy of the quasi-star as a whole, the mass must satisfy $m_* < 700 \, \alpha^{1/2} v_{10}^{3/2}$. Thus, by the time that nuclear reactions are able to run to completion, the available energy is probably insufficient to seriously affect the outer layers of the quasi-star.

It is the ultrahigh infalling rate, squeezing the core and raising its temperature beyond the thermostatic set points of thermonuclear reactions, which distinguishes the evolution of the quasi-star from that of a normal Pop III star. At best, nuclear burning can provide a brief hiatus in the contraction of the core, which ultimately reaches temperatures $\gtrsim 10^9$ K where neutrino losses become important.

6 CORE COLLAPSE AND INITIAL GROWTH OF BLACK HOLE

Continued compression by infalling matter prevents the core from losing energy radiatively and collapsing or becoming degenerate. At sufficiently high temperatures, however, neutrino losses lead to core collapse and the formation of a black hole. At $T_{\rm c} \lesssim 10^9$ K, the dominant neutrino loss mechanism is the Urca process, which is $\sim$300 times faster than photoneutrino production (Itoh et al. 1989; Qian & Woosley 1996; Dutta et al. 2004; see also Koers & Wijers 2005 for a summary of principal rates). At higher temperatures, pair annihilation becomes competitive with the Urca process, but because $\rho \propto T^3$ in the core, both mechanisms scale similarly with $T_{\rm c}$. Therefore, we may approximate the core cooling rate by $Q_{\rm c} \sim 3 \times 10^{15} (T_{\rm c}/10^9 \, \mathrm{K})^9 \, \mathrm{erg \, s^{-1} \, cm^{-3}}$ and the cooling time-scale by $t_{\rm cool} \sim 4 p_{\rm c}/Q_{\rm c}$. Setting this equal to $t_{\rm ev}$, we find that the core collapses when $m_* \sim 3600 \alpha v_{10}^3$ and $T_{\rm c} \sim 5 \times 10^8$ K.

Details of the collapse depend on the angular momentum in the core as well as the precise core mass at the time of collapse. If angular momentum is initially unimportant, the core should collapse at roughly constant temperature. As the specific entropy decreases due to cooling, gas pressure begins to exceed radiation pressure and neutrino losses are dominated by the Urca process. Because the core mass is rather low, collapse could get hung up by electron degeneracy pressure, but infalling matter from the quasi-star envelope – which continues to cool via neutrino losses – would quickly drive the mass over the Chandrasekhar limit. Continued infall similarly circumvents neutron degeneracy, with the result that a black hole of a several solar masses forms in a few times the core free-fall time.

If the angular momentum of the core and surrounding material is too large to permit direct collapse to a black hole, then neutrino cooling will lead to a rotationally supported disc. As material joins the disc, self-gravity will trigger a new round of 'bars within bars' instability, which will generate additional entropy (enhancing neutrino cooling) as well as facilitating collapse by removing angular momentum.

The amount of matter that falls promptly into the black hole depends on the distribution of angular momentum in the $\rho \propto r^{-2}$ envelope of the quasi-star. The black hole will immediately swallow all the matter in the quasi-stellar envelope with a specific angular momentum $j = \Omega r^2 \lesssim G M_{\rm BH}/c$. At one extreme, the specific angular momentum at each radius could be a fixed fraction of the Keplerian value, implying $j(M) \propto M$ as a function of the enclosed mass $M(r)$. In this case, the black hole does not grow immediately much beyond its initial mass. At the opposite extreme, the internal redistribution of angular momentum within the quasi-star could have led to solid body rotation, implying that $j \propto M^2$. In the latter case, the black hole mass could quickly swallow a fraction $\sim (G M_*/r_* c^2)^{1/2}$ of the mass of the envelope, assuming that the rotation rate reaches approximately the Keplerian value at r_*. However, this amounts to only about $20 \alpha v_{10}^3 \, \mathrm{M}_\odot$; therefore the prompt black hole mass is unlikely to exceed several tens of solar masses.

7 SUBSEQUENT BLACK HOLE GROWTH

After the initial collapse and prompt accretion phase, the growth of the black hole is regulated by angular momentum transport. The envelope continues to accumulate mass from infall, and the binding energy per unit mass increases with time, $v_*(t)^2 \sim GM_*(t)/r_*$, where r_* initially remains constant. The gravitational sphere of influence of the black hole has a radius $r_{\rm BH} \sim GM_{\rm BH}(t)/v_*(t)^2 \sim (M_{\rm BH}/M_*)r_*$. If angular momentum were unimportant, then the black hole would grow at the Bondi rate, $\dot{M}_{\rm Bondi} \sim v_*^3/G$, and would swallow the quasi-star in a free-fall time. Thereafter it would grow at the infall rate, $\alpha\dot{M}_0$.

Since angular momentum is important, a fraction of the binding energy released close to the black hole, $\epsilon\dot{M}_{\rm BH}c^2$, where $\epsilon \sim 0.1$ is the accretion efficiency, is transported outwards by the torque. If it is not radiated away from close to the hole – by neutrino losses, since there is no free surface from which a wind can emerge – this energy must react back on the inflow, slowing down the accretion. Let us assume that neutrino losses are negligible. Initially, the region affected by the feedback energy is confined to the interior of the quasi-star. The total energy liberated by the time the black hole reaches mass $M_{\rm BH}$, $E_{\rm BH} \sim \epsilon M_{\rm BH}c^2$, affects the density profile inside the quasi-star out to a radius $r_a \sim E_{\rm BH}/(\rho_* r_*^2 v_*^2)$. Note that the liberated energy is trapped inside the quasi-star, rather than flowing through it in a quasi-steady state, because the radiative diffusion time-scale at every radius is comparable to the age of the quasi-star and the growth time of the black hole (Section 5).

When r_a reaches r_*, the liberated energy equals the binding energy of the quasi-star and the latter begins to expand. This happens very early in the angular momentum-dominated growth phase, when the black hole mass has increased by an amount $\Delta M_{\rm BH} \sim GM_*^2/\epsilon c^2 r_* \lesssim O(M_{\rm BH})$. To show this, we estimate the rate at which the black hole swallows matter from the quasi-star envelope. The rate at which mass is *supplied* to the sphere of influence of the black hole can be taken to be proportional to the Bondi rate, $\dot{M}_{\rm sup} \sim 4\pi\alpha_{\rm BH}\rho(r_{\rm BH})v_*(t)r_{\rm BH}^2$, where $\rho(r_{\rm BH})$ is the density evaluated at the black hole radius of influence and $\alpha_{\rm BH} < 1$ is a parameter that describes the inefficiency of mass capture, e.g. due to a finite rate of angular momentum transport ($\alpha_{\rm BH}$ need not be the same as α). We expect the pressure and density distributions to be rather flat between $r_{\rm BH}$ and r_a because of the extra energy injection, and therefore use v_* to estimate both $r_{\rm BH}$ and the free-fall speed at the radius of influence. We also take $\rho(r_{\rm BH}) \sim \rho(r_a) \sim \rho_*(r_*/r_a)^2$. Note, however, that even if the density increases $\propto r^{-\beta}$ at $r < r_a$, our prescription gives a lower limit to $\dot{M}_{\rm sup}$ provided that $\beta < 14/5$.

The rate at which matter actually reaches the black hole is suppressed by a further factor, due to the back reaction of the energy flux inside the radius of influence (Gruzinov 1998; Blandford & Begelman 1999; Narayan, Igumenshchev & Abramowicz 2000; Quataert & Gruzinov 2000). In the absence of a wind that removes energy and/or angular momentum, the accretion rate is reduced to $\dot{M}_{\rm BH} \sim \epsilon^{-1}(v_*/c)^2\dot{M}_{\rm sup}$ (Blandford & Begelman 1999, 2004). We then find that the accretion rate is

$$\dot{M}_{\rm BH} \sim 3\frac{\alpha_{\rm BH}}{\epsilon^3}\frac{c^3}{G}\left(\frac{v_*}{c}\right)^9. \tag{12}$$

Since $v_* \propto M_*^{1/2}$, $\dot{M}_{\rm BH} \propto M_*^{9/2}$ is a steeply increasing function of M_*. Comparing it to the inflow rate on to the quasi-star, we find that the feedback energy equals the binding energy of the quasi-star before the black hole mass has doubled. Thus Bondi accretion, even modified by feedback and a finite rate of angular momentum transport, should quickly bring the quasi-star to the point where its evolution is driven by feedback from the black hole.

The feedback flux does not blow apart the quasi-star, since this would stop the growth of the black hole and therefore the feedback. Instead the quasi-star expands gradually, allowing the black hole accretion rate to adjust so that the feedback energy flux equals the Eddington limit for the instantaneous quasi-star mass, $\dot{M}_{\rm BH} \sim 2\times10^{-3}(\epsilon/0.1)^{-1}(m_*/10^5)\,{\rm M}_\odot\,{\rm yr}^{-1}$. The feedback energy flux exceeds the Eddington limit for the black hole by a factor of $M_*/M_{\rm BH}$; thus, the black hole grows at a super-Eddington rate as long as $M_* > M_{\rm BH}$. This configuration requires most of the feedback flux to be carried by convective motions inside the quasi-star, since the enclosed mass at $r_{\rm BH} < r < r_*$ is a steeply increasing function of radius. However, one can show that the required convective velocity, while larger than the mean inflow speed, is much smaller than the local free-fall speed at all r. If the quasi-star mass continues to increase at the constant rate $\alpha\dot{M}_0$, then the black hole mass evolves according to

$$M_{\rm BH}(t) \sim 4\times10^5\alpha v_{10}^3\left(\frac{\epsilon}{0.1}\right)^{-1}\left(\frac{t}{10^7\ {\rm yr}}\right)^2 {\rm M}_\odot, \tag{13}$$

i.e. $M_{\rm BH} \propto M_*^2$.

To determine the evolution of the structure of the quasi-star in response to feedback, we estimate $\dot{M}_{\rm BH}$ using the modified Bondi rate discussed above. If we assume the density to be roughly uniform within the quasi-star (outside $r_{\rm BH}$), we have

$$\dot{M}_{\rm BH} \sim 3\alpha_{\rm BH}\frac{c^3}{\epsilon G}\left(\frac{M_{\rm BH}}{M_*}\right)^2\left(\frac{v_*}{c}\right)^4. \tag{14}$$

Equating this to the Eddington-limited rate (assuming electron scattering opacity) and using equation (13) and the assumed infall rate on to the quasi-star, we obtain

$$r_* \sim 2\times10^{15}\alpha^{-1}v_{10}^{-3}\left(\frac{\alpha_{\rm BH}}{0.01}\right)^{1/2}\left(\frac{\epsilon}{0.1}\right)^{-1}\left(\frac{m_*}{10^5}\right)^{3/2}{\rm cm}. \tag{15}$$

Neutrino losses are rapidly quenched by the expansion of the quasi-star. Radiation pressure dominates throughout the envelope, and the temperature (in LTE) decreases linearly with time (and with M_*), $T_* \sim 4\times10^5(\alpha_{\rm BH}/0.01)^{-1/2}\alpha v_{10}^3(\epsilon/0.1)\,(m_*/10^5)^{-1}$ K. Within the radius of influence of the black hole, the pressure varies $\propto r^{-3/2}$ (not $\propto r^{-5/2}$, as in ordinary Bondi accretion, because of the feedback), and $T \propto r^{-3/8}$ can exceed T_* by a factor of as large as $(\epsilon c^2/v_*^2) \sim 40\ (\alpha_{\rm BH}/0.01)^{3/16}(\alpha v_{10}^3)^{-3/8}(m_*/10^5)^{3/16}$, close to the black hole. Such temperatures are inadequate to produce a significant neutrino flux when the black hole grows much beyond its initial mass.

The effective temperature of the photosphere of the quasi-star is also expected to decrease, implying that the quasi-star is unlikely to be a significant source of hard ultraviolet radiation when it has grown beyond $\sim10^4\,{\rm M}_\odot$. Up to this point, the rate of production of ionizing photons is very high, of the order of 10^{55} photons s^{-1}; but since this phase lasts for $\lesssim10^5$ yr, the total output falls far short of the requirement for reionizing the Universe. Similarly, the quasi-star produces $\simeq10^{50}$ photons s^{-1} in the Lyman–Werner band, but can keep the molecular hydrogen in its surroundings photodissociated only for $\lesssim10^5$ yr. These estimates correspond to a spherical photosphere at r_*, but we note that photospheric temperatures could be even lower if the photosphere is strongly flattened by rotation.

The above estimates are valid only as long as $T_* \gtrsim 10^4$ K, corresponding to

$$M_* < 4\times10^6\alpha v_{10}^3\left(\frac{\alpha_{\rm BH}}{0.01}\right)^{-1/2}\left(\frac{\epsilon}{0.1}\right){\rm M}_\odot \tag{16}$$

and

$$M_{\rm BH} < 9 \times 10^5 \alpha v_{10}^3 \left(\frac{\alpha_{\rm BH}}{0.01}\right)^{-1} \left(\frac{\epsilon}{0.1}\right) {\rm M}_\odot. \quad (17)$$

For the fiducial parameters, the black hole mass at this stage is almost as large as that of the quasi-star; further growth can occur at the Eddington limit of the black hole. However, we emphasize the uncertainty in parameters such as $\alpha_{\rm BH}$ and α. (We use different fiducial estimates of α and $\alpha_{\rm BH}$ because the latter represents viscous transport of angular momentum while the former represents a removal of gas from the inflow due to fragmentation and star formation).

At lower temperatures, the Planck mean opacity (which is relevant for calculating the radiation force in LTE, and therefore the Eddington limit) becomes very sensitive to temperature (Mayer & Duschl 2005), increasing sharply at $T_* \lesssim 10^4$ K and then decreasing to several orders of magnitude below the electron scattering opacity as the temperature declines further. The sharp decrease in opacity would affect the photosphere at an even earlier stage in the evolution of the quasi-star.

Finally, we note the existence of an alternative evolutionary scenario in which $\alpha_{\rm BH}$ is so small that feedback does not regulate the structure of the quasi-star. If $\alpha_{\rm BH}$ is essentially zero (in practice, $\ll 10^{-6}$ when $m_* \sim 10^5$), then the centrifugal barrier forms a wall within the quasi-star at $r \sim r_{\rm BH}$. r_* is once again constant and the temperature at $r_{\rm BH}$ scales as $M_*/M_{\rm BH}^{1/2}$. If this ratio increases with time, following the initial collapse and prompt accretion phase, then neutrino losses remain important within the sphere of influence of the black hole. We are then justified in assuming that self-gravitational instabilities transport angular momentum effectively. Provided that nearly all of the liberated binding energy is carried away by neutrinos, we deduce that the black hole grows at such a rate that $M_{\rm BH} \propto M_*^2 \propto t^2$. Adopting $T(r_{\rm BH}) = 10^9 T_9$ K as the threshold for rapid neutrino cooling, we find that the black hole mass (in solar units) grows according to

$$m_{\rm BH} \sim 225 \alpha^{-2} v_{10}^{-6} T_9^{-2} \left(\frac{m_*}{10^5}\right)^2 \sim 9 \times 10^4 T_9^{-2} \left(\frac{t}{10^7\ {\rm yr}}\right)^2. \quad (18)$$

This rate is not much smaller than the Eddington-limited rate, equation (13), in the presence of feedback. Given the convergence of these two extreme estimates, we are reasonably confident that rapid black hole growth up to masses $\sim 10^4$–10^6 M$_\odot$ is possible under the conditions postulated here.

The discussion in Sections 3–7 can be generalized to haloes with $T_{\rm vir} < 10^4$ K. If molecular hydrogen cools the gas down to $\sim$200 K, then runaway collapse without fragmentation could occur in haloes with correspondingly low virial temperatures. The characteristic infall speed is then $v_{10} \lesssim 0.2$, implying inflow rates of a few thousandths of a solar mass per year. Nevertheless, a quasi-star with a hot dense core will eventually develop, and will ultimately collapse to form a black hole due to runaway neutrino cooling. The mass of the prompt black hole is insensitive to v_{10}, and therefore is still $\lesssim 20\,{\rm M}_\odot$. However, the mass of the quasi-star at this stage scales as v_{10}^3; thus the quasi-star mass is only a few times that of the black hole. Moreover, the black hole could not reach more than a few thousand solar masses before the growth rate becomes sub-Eddington, and limited by the inflow of gas into the quasi-star.

8 EVOLUTION OF THE BLACK HOLE POPULATION

How large a population of black holes is likely to result from gravitational instability of gas discs in high-redshift haloes? Given the threshold of $T_{\rm vir} > 10^4$ K for efficient cooling, and therefore for 'bars within bars' instability, we can trace the co-evolution of the black holes and their hosts. A $T_{\rm vir} \sim 10^4$ K halo has a mass between 10^7 M$_\odot$ and 10^9 M$_\odot$ at redshift $6 < z < 20$. The black hole forming in such a host could grow at the super-Eddington rate given by equation (13) until it reaches $\sim 10^6$ M$_\odot$, at which point its mass would approach that of the quasi-star. Thereafter it could grow, by Eddington-limited accretion (or the infall rate, if smaller), to an even larger mass. However, the growth of the black hole can also be terminated earlier by lack of fuel, i.e. by using up all the available gas in the unstable disc (if $f_{\rm d}$ is much smaller than unity) and/or if star formation depletes the inflow. Therefore, we do not assume that black holes grow rapidly to the maximum allowed mass. Indeed, we will see below that our mechanism can provide the seeds for all present-day SMBHs, even if its efficiency is quite low.

More massive haloes, with $T_{\rm vir} \gg 10^4$ K, are probably prone to fragmentation and star formation, which would inhibit instability and therefore the formation of a black hole by this process. Using the Press–Schechter formalism (Sheth & Tormen 1999), we estimate that the number density of haloes with virial temperature $T_{\rm vir} > 2 \times 10^4$ K ($T_{\rm vir} > 5 \times 10^4$ K) is about 10 per cent (1 per cent) of the density of haloes with 10^4 K $< T_{\rm vir} < 2 \times 10^4$ K (10^4 K $< T_{\rm vir} < 5 \times 10^4$ K). More massive haloes make an even smaller contribution. Since the contribution of haloes with $T_{\rm vir} \gg 10^4$ K is negligible, we can estimate the black hole density by integrating over all host haloes with $T_{\rm vir} > 10^4$ K.

Among all haloes with $T_{\rm vir} > 10^4$ K, only those with a low enough spin parameter (given $f_{\rm d}$) will host a disc unstable to 'bars within bars' instability. Assuming a seed black hole mass of 20 M$_\odot$, we plot the comoving mass density of black hole seeds in Fig. 2. The mass density is small, but this process can nevertheless seed most of the systems which will evolve into the local galaxies where SMBHs have been found. For a given local halo, the extended Press–Schechter formalism can be used to estimate the average number of progenitors with $T_{\rm vir} \sim 10^4$ K at $z > 10$. The probability of black hole formation depends also on the amount of gas which condenses to form a disc (see Section 2). If the fraction of gas typically ending up in the disc is $f_{\rm d} \approx 0.5$, we find that this process needs

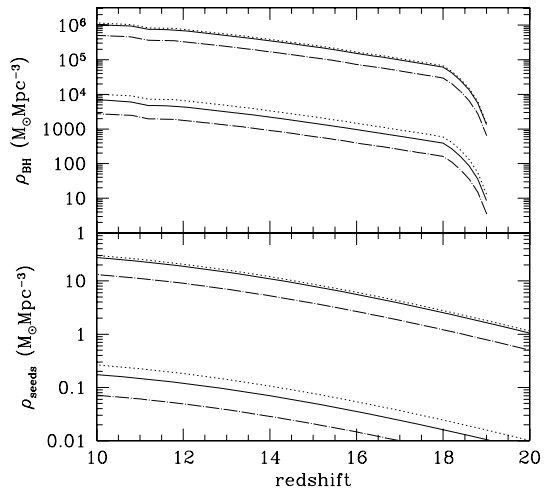

Figure 2. Lower panel: comoving density of 20 M$_\odot$ black hole seeds as a function of redshift. Upper panel: comoving density of black holes, assuming continuous formation and growth of seeds to 10^6 M$_\odot$ according to equation (13), starting from $z = 20$. Solid line: Mestel disc, dotted line: rigid disc, dot–dashed line: exponential disc. The upper set of lines assumes $f_{\rm d} = 0.5$, the lower set assumes $f_{\rm d} = 0.1$.

to operate only until $z \simeq 18$ in order to supply seeds to all present-day haloes of Milky Way, or larger, size. If black hole formation proceeded to $z \simeq 14$, then all haloes with mass $>10^{11}\,\mathrm{M}_\odot$ today could have been seeded (see Barth, Greene & Ho 2005). If f_d was lower, the black hole formation process would have had to continue for longer, in order to form seeds in the progenitors of most local galaxies. These constraints would be eased by an early generation of black hole seeds, formed in a small fraction of haloes with $T_\mathrm{vir} \gtrsim 200\,\mathrm{K}$, before H_2 is photodissociated by the first stars.

The process of seed formation can be widespread enough to account for subsequent SMBH evolution, even if the subsequent black hole growth (Section 7) is inefficient in most high-redshift haloes. We note, in fact, that black holes with masses well below $\sim 10^6\,\mathrm{M}_\odot$ are expected in local faint active galactic nuclei (Barth et al. 2005). The growth of black hole seeds up to $\sim 10^4$–$10^6\,\mathrm{M}_\odot$ cannot therefore happen in all high-redshift systems.

We can estimate the upper limit to the total black hole mass density predicted by our model by assuming continuous formation of seed black holes, and adopting equation (13) to estimate their early growth. We let black holes grow until either they have consumed all the gas, or the infall of matter on to the black hole greatly decreases, at $M_\mathrm{BH} \approx 10^6\,\mathrm{M}_\odot$. The comoving density of haloes is estimated using the Sheth & Tormen (1999) halo mass function. The black hole mass density shown in Fig. 2 must be compared to the density that we observe at low z, i.e. $\rho_\mathrm{BH}(z=0) \approx 3$–$5 \times 10^5\,\mathrm{M}_\odot\,\mathrm{Mpc}^{-3}$ (Aller & Richstone 2002; Yu & Tremaine 2002; Fabian & Iwasawa 1999; Elvis, Risaliti & Zamorani 2002; Marconi et al. 2004) and $\rho_\mathrm{BH}(z=3) \approx 4$–$5 \times 10^4\,\mathrm{M}_\odot\,\mathrm{Mpc}^{-3}$ (Merloni 2004). At $z = 6$, a lower limit to the black hole density is obtained by integrating the observed luminosity function, $\rho_\mathrm{BH}(z=6) \approx \mathrm{few} \times \mathrm{M}_\odot\,\mathrm{Mpc}^{-3}$ (Fan et al. 2004). This density, however, includes only black holes with masses $\gtrsim 10^9\,\mathrm{M}_\odot$. To obtain a rough estimate of the total black hole density at $z = 6$, we adopt the Sołtan (1982) argument. We extrapolate the luminosity function up to $z = 6$, assuming the redshift dependence given by Richards et al. (2005), integrate the emitted quasar luminosity, and convert it into a black hole density by normalizing to the $z = 0$ SMBH density. We do not correct here for obscured quasars, or assume a radiative efficiency, and therefore consider the density that we find, $\rho_\mathrm{BH}(z=6) \approx 1$–$3 \times 10^3\,\mathrm{M}_\odot\,\mathrm{Mpc}^{-3}$, more of an indication than a robust estimate.

The densities shown in Fig. 2 must be regarded as upper limits to the black hole density, as we have not included any effect that can interrupt or disturb black hole formation and growth. We discuss in the next section how the hierarchical framework for structure formation can modify this simple picture. We also recall that, eventually, efficient star formation occurs in these haloes, competing for the gas supply and possibly limiting the mass available for black hole accretion.

9 DISCUSSION AND CONCLUSIONS

We have presented a scenario for the accumulation of gas in the centres of dark matter haloes with $T_\mathrm{vir} \gtrsim 10^4\,\mathrm{K}$, the initial collapse of the gas to form seed black holes, and the subsequent early growth of SMBHs. This mechanism can lead naturally to the super-Eddington growth of black holes up to masses $\sim 10^6\,\mathrm{M}_\odot$, as early as redshifts 10–20. Given additional growth to $\sim 10^9\,\mathrm{M}_\odot$ at close to the Eddington rate, the model can account for the population of quasars observed at $z \sim 6$ (Fan et al. 2004). Even without significant growth after the formation phase, this mechanism could produce the seeds for all SMBHs inferred to exist in the local Universe.

We argue that global self-gravity triggers the 'bars within bars' instability (Shlosman et al. 1989, 1990), under certain conditions, as gas forms a rotationally supported thick disc in the centre of the halo. On scales much smaller than the disc radius, and times shorter than the free-fall time, quasi-steady inflow is a better representation of the infall than a monolithic collapse. Local, or quasi-local sources of 'viscosity', such as those due to magnetic fields, turbulence or radiation drag, are not required to transport the angular momentum that inhibits black hole formation. In metal-free haloes with little molecular hydrogen, this behaviour is possible once the virial temperature exceeds $\sim 10^4\,\mathrm{K}$ (Oh & Haiman 2002). Under these conditions, gravitational instabilities can transport angular momentum effectively from scales of several kpc down to scales initially as small as ~ 1 au, at a fraction of a solar mass per year (for a characteristic infalling speed $v_0 \sim 10\,\mathrm{km\,s}^{-1}$). We suggest that inflow is most efficient in haloes where T_vir does not exceed a few times $10^4\,\mathrm{K}$, since fragmentation of the infalling gas is unlikely to be efficient in this case (Bromm & Loeb 2003). As the mass in the centre builds up, global instabilities may be quenched in the inner regions, but local gravitational instabilities could continue to drive a substantial inflow, even if a certain amount of star formation occurs.

Instability occurs only where the halo spin parameter, λ_spin, falls below a threshold value that depends on f_d, the fraction of gas that forms the disc. For $f_\mathrm{d} \gtrsim 0.5$ the threshold is comparable to the mean spin parameter predicted by simulations, and >20 per cent of all $T_\mathrm{vir} \sim 10^4\,\mathrm{K}$ haloes should exhibit instability. Even a value of f_d as small as 0.1 leads to instability in $\gtrsim$1 per cent of haloes and a significant seed population of black holes.

In haloes with the low angular momentum required to trigger black hole formation, the rapid infall of gas leads to a mass accumulation much larger than that expected in a mini-halo with an average spin parameter. The formation of a 'standard' Pop III star (Bromm et al. 1999; Abel et al. 2000; Bromm et al. 2002) is therefore suppressed in favour of a massive 'quasi-star' (Section 4). We suggest that the much smaller mini-haloes (with virial temperature well below $10^4\,\mathrm{K}$) that form the first stars are distinct from the haloes that form the seeds of SMBHs, although the former may be precursors of the latter. Photodissociation of molecular hydrogen, possibly by a small population of Pop III stars, would suppress fragmentation of the infalling gas. It is therefore possible that the formation of seed black holes follows an earlier epoch of star formation, as the 'quasi-star' itself is not a significant source of photodissociating photons for long. The epoch of black hole formation must happen early enough, however, that the Universe is not highly metal enriched – later episodes of star formation would enrich the environments of seed black holes.

The most important conclusion of our model is that the 'quasi-star' formed by the accumulating gas has a low-entropy, gas pressure-dominated core surrounded by a radiation pressure-dominated envelope. As matter piles on to the quasi-star, the core is squeezed until its temperature approaches $10^9\,\mathrm{K}$ (typically when the envelope mass reaches a few thousand $\mathrm{M}_\odot$). Cooling by thermal neutrinos then leads to core collapse and the formation of a seed back hole of $\sim$10–20 $\mathrm{M}_\odot$. This is a novel application of neutrino cooling, which has been invoked previously in connection with hyperaccretion on to neutron stars in supernovae (Colgate 1971; Chevalier 1989; Houck & Chevalier 1991) and common envelope binaries (Chevalier 1993; Brown, Lee & Bethe 2000), and on to black holes in gamma-ray bursts (Narayan, Paczyński & Piran 1992; Woosley 1993; Popham, Woosley & Fryer 1999; Narayan, Piran & Kumar 2001). It is difficult to set up the necessary conditions for efficient neutrino cooling, since radiation pressure generally prevents the

accretion rate from reaching the required level from below, unless the viscosity parameter α is extremely small (Chevalier 1996). Previous models have circumvented this problem by invoking strong radiation trapping in a steady-state (or slowly varying) accretion flow. In our case the inflow sets up favourable conditions in a time-dependent fashion by establishing a steep positive entropy gradient in the quasi-star, with only mild radiation trapping.

The subsequent evolution of the black hole can be very fast, with growth to more than a million solar masses possible in less than a Salpeter time-scale. Even taking account of strong energy feedback driven by angular momentum transport, we conclude that black holes can accrete at the Eddington rate *for the quasi-star mass*, which exceeds the Eddington rate for the black hole by a factor of $M_*/M_{\rm BH}$. For steady infall on to the quasi-star, this corresponds to a black hole mass increasing with time as $M_{\rm BH} \propto t^2$.

If black hole growth (equation 13) proceeded undisturbed in all haloes satisfying the instability criterion with $T_{\rm vir} > 10^4$ K, then the total mass density in SMBHs would become comparable to the local one already at high redshift. However, a number of effects can limit this initial phase of rapid growth. Limitations intrinsic to the halo include the overall mass supply that participates in the infall, as well as removal of matter from the inflow by star formation. Moreover, the haloes and their embedded black holes do not grow in isolation. Haloes susceptible to the 'bars within bars' instability represent high peaks in the field of density fluctuations (Tegmark et al. 1997; Madau et al. 2004). Therefore, they should experience an enhanced number of major mergers with respect to 'average' haloes at the same redshift. Halo major mergers can modify our basic results in two ways. First, cosmological simulations show that the spin parameter of a halo typically increases after a major merger (Vitvitska et al. 2002). On the other hand, the spin parameter decreases after a long series of minor mergers. Major mergers, therefore, should delay the triggering of instabilities, at least until a sufficient number of minor mergers has lowered the spin parameter again. Secondly, a major merger could destroy the coherence of the 'bars within bars' process. By interfering with the infall of matter on to the quasi-star, a violent encounter could hasten the depletion of the mass supply well before the upper limits discussed above are reached. Such a disturbance is unlikely to modify the interior structure of the existing quasi-star or suddenly stop the growth of the black hole, however, since these involve only the very core of the system.

Seeding of larger haloes by black hole mergers could also be limited by the 'gravitational rocket' effect, the recoil due to the net linear momentum carried away by gravitational waves in the coalescence of two black holes (Haiman 2004; Madau et al. 2004; Yoo & Miralda-Escudé 2004). The recoil velocity still has large uncertainties, but can easily exceed $\sim 100\,{\rm km\,s^{-1}}$, comparable to the escape velocity from shallow halo potentials. Volonteri & Rees (2006) estimate that up to 50 per cent of black holes merging in high-redshift haloes can be ejected due to the gravitational rocket effect.

Despite these potential sources of inefficiency, the mechanism we have outlined could be the principal route leading to SMBH formation in galactic nuclei. The main elements of the model – particularly the cascading infall via 'bars within bars' instability, and the formation and evolution of the quasi-star with runaway neutrino cooling – should be testable via numerical simulations. We hope that such simulations can be undertaken shortly.

ACKNOWLEDGMENTS

This work was supported in part by NASA Beyond Einstein Foundation Science grant NNG05G192G, NSF grant AST-0307502, and the University of Colorado Council on Research and Creative Work. MCB thanks the Institute of Astronomy and the Master and Fellows of Trinity College, Cambridge, for their hospitality.

REFERENCES

Abel T., Bryan G., Norman M., 2000, ApJ, 540, 39
Aller M. C., Richstone D., 2002, AJ, 124, 3035
Barth A. J., Greene J. E., Ho L. C., 2005, ApJ, 619, L151
Blandford R. D., Begelman M. C., 1999, MNRAS, 303, L1
Blandford R. D., Begelman M. C., 2004, MNRAS, 349, 68
Bondi H., 1952, MNRAS, 112, 195
Bromm V., Loeb A., 2003, ApJ, 596, 34
Bromm V., Coppi P. S., Larson R. B., 1999, ApJ, 527, L5
Bromm V., Coppi P. S., Larson R. B., 2002, ApJ, 564, 23
Brown G. E., Lee C.-H., Bethe H. A., 2000, ApJ, 541, 918
Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001, ApJ, 555, 240
Carr B. J., Bond J. R., Arnett W. D., 1984, ApJ, 277, 445
Chevalier R. A., 1989, ApJ, 346, 847
Chevalier R. A., 1993, ApJ, 411, L33
Chevalier R. A., 1996, ApJ, 459, 322
Christodoulou D. M., Shlosman I., Tohline J. E., 1995, ApJ, 443, 563
Cole S., Lacey C., 1996, MNRAS, 281, 716
Colgate S. A., 1971, ApJ, 163, 221
Colgate S. A., Cen R., Li H., Currier N., Warren M. S., 2003, ApJ, 598, L7
Dutta S. I., Ratkovic S., Prakash M., 2004, Phys. Rev. D, 69, 023005
Efstathiou G., Lake G., Negroponte J., 1982, MNRAS, 199, 1069
Eisenstein D. J., Loeb A., 1995, ApJ, 443, 11
Elvis M., Risaliti G., Zamorani G., 2002, ApJ, 565, L75
Englmaier P., Shlosman I., 2004, ApJ, 617, L115
Fabian A. C., Iwasawa K., 1999, MNRAS, 303, 34
Fall S. M., Efstathiou G., 1980, MNRAS, 193, 189
Fan X. et al., 2004, AJ, 128, 515
Gammie C. F., 2001, ApJ, 553, 174
Gruzinov A., 1998, unpublished manuscript, preprint (astro-ph/9809265)
Haehnelt M. G., Rees M. J., 1993, MNRAS, 263, 168
Haiman Z., 2004, ApJ, 613, 36
Haiman Z., Abel T., Rees M. J., 2000, ApJ, 534, 11
Houck J. C., Chevalier R. A., 1991, ApJ, 376, 234
Itoh N., Adachi T., Nakagawa M., Kohyama Y., Munakata H., 1989, ApJ, 339, 354
Koers H. B. J., Wijers R. A. M. J., 2005, MNRAS, 364, 934
Koushiappas S. M., Bullock J. S., Dekel A., 2004, MNRAS, 354, 292
Larson R. B., 1969, MNRAS, 145, 271
Larson R. B., 1998, MNRAS, 301, 569
Loeb A., Rasio F. A., 1994, ApJ, 432, 52
Machacek M. E., Bryan G. L., Abel T., 2001, ApJ, 548, 509
Madau P., Rees M. J., Volonteri M., Haardt F., Oh S. P., 2004, ApJ, 606, 484
Marconi A., Risaliti G., Gilli R., Hunt L. K., Maiolino R., Salvati M., 2004, MNRAS, 351, 169
Mayer M., Duschl W. J., 2005, MNRAS, 358, 614
Merloni A., 2004, MNRAS, 353, 1035
Mestel L., 1963, MNRAS, 126, 553
Mineshige S., Umemura M., 1997, ApJ, 480, 167
Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319
Narayan R., Paczyński B., Piran T., 1992, ApJ, 395, L83
Narayan R., Igumenshchev I. V., Abramowicz M. A., 2000, ApJ, 539, 798
Narayan R., Piran T., Kumar P., 2001, ApJ, 557, 949
Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493 (NFW)
Oh S. P., Haiman Z., 2002, ApJ, 569, 558
Ostriker J. P., Peebles P. J. E., 1973, ApJ, 186, 467
Penston M. V., 1969, MNRAS, 144, 425
Popham R., Woosley S. E., Fryer C., 1999, ApJ, 518, 356
Qian Y.-Z., Woosley S. E., 1996, ApJ, 471, 331
Quataert E., Gruzinov A., 2000, ApJ, 539, 809
Richards G. T. et al., 2005, MNRAS, 360, 839

Saigo K., Hanawa T., 1998, ApJ, 493, 342
Santoro F., Shull J. M., 2006, ApJ, 643, 26
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Shlosman I., Frank J., Begelman M. C., 1989, Nat, 338, 45
Shlosman I., Begelman M. C., Frank J., 1990, Nat, 345, 679
Shu F. H., 1977, ApJ, 214, 488
Sołtan A., 1982, MNRAS, 200, 115
Tegmark M., Silk J., Rees M. J., Blanchard A., Abel T., Palla F., 1997, ApJ, 474, 1
Toomre A., 1964, ApJ, 139, 1217
Umemura M., Loeb A., Turner E. L., 1993, ApJ, 419, 459
van den Bosch F. C., Abel T., Croft R. A. C., Hernquist L., White S. D. M., 2002, ApJ, 576, 21
Vitvitska M., Klypin A. A., Kravtsov A. V., Wechsler R. H., Primack J. R., Bullock J. S., 2002, ApJ, 581, 799
Volonteri M., Rees M. J., 2005, ApJ, 633, 624
Volonteri M., Rees M. J., 2006, ApJ, submitted
Volonteri M., Haardt F., Madau P., 2003, ApJ, 582, 559
Volonteri M., Madau P., Quataert E., Rees M. J., 2005, ApJ, 620, 69
Warren M. S., Quinn P. J., Salmon J. K., Zurek W. H., 1992, ApJ, 399, 405
Woosley S. E., 1993, ApJ, 405, 273
Yoo J., Miralda-Escudé J., 2004, ApJ, 614, L25
Yu Q., Tremaine S., 2002, MNRAS, 335, 965

This paper has been typeset from a TEX/LATEX file prepared by the author.