Bursting and quenching in massive galaxies without major mergers or AGNs

Yuval Birnboim,[★] Avishai Dekel[★] and Eyal Neistein[★]

Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel

ABSTRACT

We simulate the build-up of galaxies by spherical gas accretion through dark matter haloes, subject to the development of virial shocks. We find that a uniform cosmological accretion rate turns into a rapidly varying disc build-up rate. The generic sequence of events (Shocked-Accretion Massive Burst and Shutdown, SAMBA) consists of four distinct phases: (i) continuous cold accretion while the halo is below a threshold mass $M_{sh} \sim 10^{12}\,M_\odot$, (ii) tentative quenching of gas supply for $\sim 2\,\mathrm{Gyr}$, starting abruptly once the halo is $\sim M_{sh}$ and growing a rapidly expanding shock, (iii) a massive burst due to the collapse of $\sim 10^{11}\,M_\odot$ gas in $\sim 0.5\,\mathrm{Gyr}$, when the accumulated heated gas cools and joins new infalling gas and (iv) a long-term shutdown, enhanced by a temporary shock instability in late SAMBAs, those that quench at $z \sim 2$, burst at $z \sim 1$ and end up quenched in $10^{12-13}\,M_\odot$ haloes today. The quenching and bursting occur at all redshifts in galaxies of baryonic mass $\sim 10^{11}\,M_\odot$ and involve a substantial fraction of this mass. They arise from rather smooth accretion, or minor mergers, which, unlike major mergers, may leave the disc intact while being built in a rapid pace. The early bursts match observed maximum starbursting discs at $z \gtrsim 2$, predicted to reside in $\lesssim 10^{13}\,M_\odot$ haloes. The late bursts resemble discy luminous infrared galaxies (LIRGs) at $z \lesssim 1$. On the other hand, the tentative quenching gives rise to a substantial population of $\sim 10^{11}\,M_\odot$ galaxies with a strongly suppressed star formation rate at $z \sim 1$–3. The predicted long-term shutdown leads to red and dead galaxies in groups. A complete shutdown in more massive clusters requires an additional quenching mechanism, as may be provided by clumpy accretion. Alternatively, the SAMBA bursts may trigger the active galactic nucleus (AGN) activity that couples to the hot gas above M_{sh} and helps the required quenching. The SAMBA phenomenon is predicted based on a spherical model that does not simulate star formation and feedback – it is yet to be investigated using detailed cosmological simulations.

Key words: accretion, accretion discs – shock waves – galaxies: evolution – galaxies: formation – galaxies: haloes – dark matter.

1 INTRODUCTION

Observations reveal a puzzling phenomenon of 'maximum starbursting' in massive galaxies at high redshift in cases where major mergers are ruled out. Examples include luminous infrared galaxies (LIRGs) at $z \lesssim 1$ (Hammer et al. 2005) and maximum bursters at $z \gtrsim 2$ (Förster Schreiber et al. 2006; Genzel et al. 2006). In these cases, galaxies of stellar masses $\sim 10^{11}\,M_\odot$ seem to be forming a large fraction of their stars at a star formation rate (SFR) ~ 100–$200\,M_\odot\,\mathrm{yr}^{-1}$ within $\sim 0.5\,\mathrm{Gyr}$. This is much shorter than the

age of the universe then, indicating coherent star formation from $\sim 10^{11}\,M_\odot$ of gas in the latest stages of its collapse/assembly rather than in its earlier smaller progenitors. Furthermore, the burst duration is significantly shorter than the time implied by the typical cosmological accretion rate on to haloes of $\sim 10^{12}\,M_\odot$ at the relevant epochs, and is as short as the time for streaming at the virial velocity from the virial radius to the centre. Gaseous major mergers, which could have provided such rapid bursts, are unlikely in many of these cases, where the galaxies are detected to be relatively smooth, rotating, thick discs that could not have survived a major merger and show no trace of such an event. At the low end of the SFR distribution, most galaxies of stellar mass $M_d \geqslant 10^{11}\,M_\odot$ tend to be quenched red and dead galaxies, and a substantial fraction of such galaxies show strongly suppressed SFRs already at high

★E-mail: yuval@phys.huji.ac.il (YB); dekel@phys.huji.ac.il (AD); eyal_n@phys.huji.ac.il (EN)

redshifts, $0.5 < z < 2.7$ (Kriek et al. 2006; Noeske et al. 2007). We seek an explanation for these phenomena at the two extreme regimes of SFR.

We report here on new insight gained from spherical modelling of gas accretion on to massive galaxies, which could be the basis for understanding the above phenomena and other central issues in galaxy formation.

Following the classic work on the interplay between cooling and dynamical times in galaxy formation (Binney 1977; Rees & Ostriker 1977; Silk 1977; White & Rees 1978; Blumenthal et al. 1984), and earlier one-dimensional (1D) simulations (Thoul & Weinberg 1995; Forcada-Miro & White 1997; Ciotti & Ostriker 2001), we performed analytic calculations accompanied by hydrodynamical simulations to study the evolution of a virial shock in a spherical configuration (Birnboim & Dekel 2003; Dekel & Birnboim 2006). We predicted the existence of a threshold halo mass for the presence of a virial shock, at $M_{\text{sh}} \sim 10^{12}\,\text{M}_\odot$, roughly independent of redshift. Less massive haloes do not permit a stable shock, as rapid radiative cooling prevents the hypothetical post-shock gas pressure from supporting the shock against global gravitational collapse into the halo centre. The gas accreting through the virial radius of such haloes flows cold ($\sim 10^4\,\text{K}$) into the inner halo, where it may eventually shock, feed a disc and efficiently form stars. Once the halo grows above M_{sh}, a stable shock emerges from the inner halo and rapidly propagates toward the virial radius, halting the infalling gas and creating a hot medium in quasi-static equilibrium at the halo virial temperature. We found that the transition to stability occurs when the standard radiative cooling time at the assumed metallicity equals the time for compression behind the shock, $t_{\text{comp}} = (21/5)\rho/\dot{\rho} \simeq (4/3)R/V$, where ρ is the gas density behind the shock, R is the shock radius and V is the infall velocity into the shock.

Based on these findings, it has been proposed that the presence of hot, dilute gas allows the suppression of gas supply to the disc, possibly assisting active galactic nucleus (AGN) feedback, and thus leads to quenching of star formation (Binney 2004; Dekel & Birnboim 2006). This mass threshold has proven to be the key for understanding the robust division of galaxies into blue, star-forming discs versus red and dead spheroids and their basic properties (Cattaneo et al. 2006; Croton et al. 2006; Dekel & Birnboim 2006).

The success of the spherical model in predicting the threshold mass as seen in cosmological hydrodynamical simulations (Kereš et al. 2005; Dekel & Birnboim 2006) is demonstrated in Fig. 1, which summarizes the results of Birnboim et al. (in preparation). It shows as a function of halo mass the fraction of cold gas within the virial radius and outside the disc in dark matter haloes from cosmological simulations by A. Kravtsov (described in Kravtsov 2003; Kravtsov & Gnedin 2005, where they were used for other purposes). At all redshifts, there is a transition from cold-dominated haloes to hot-dominated haloes near $5 \times 10^{11}\,\text{M}_\odot$. This is compatible with the threshold mass predicted by Dekel & Birnboim (2006) for realistic halo metallicities that gradually grow to $Z \simeq 0.1$ solar today.

This success of the spherical model motivates our present, more detailed and daring investigation of the spherical configuration. We study how the development of the virial shock converts an otherwise uniform accretion through the virial radius into a rapidly varying accretion rate on to the inner disc. The resultant disc build-up, and the induced SFR, evolve through a generic sequence of events involving two quenching episodes and a massive burst, which always starts at the crossing of the halo threshold mass, namely in galaxies of $M_{\text{d}} \sim 10^{11}\,\text{M}_\odot$ at any redshift. We term this phenomenon SAMBA, for Shocked-Accretion Massive Burst and Shutdown.
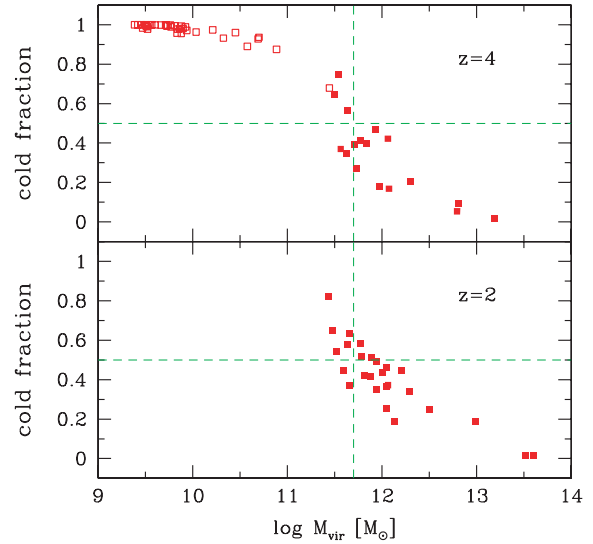

Figure 1. The threshold halo mass for virial shock heating in cosmological hydro simulations, at different redshifts, based on Birnboim et al. (in preparation). Each point represents the fraction of cold gas within $0.1R$-R in a different halo, for haloes above a minimum mass due to numerical resolution. The smaller haloes (open symbols) are drawn from a high-resolution simulation in a box of comoving size $6\,h^{-1}$ Mpc, which was terminated at $z = 4$ and therefore do not show counterparts at later redshifts. The more massive haloes (solid symbols) come from a simulation of a lower resolution in a box of $80\,h^{-1}$ Mpc. Cold gas is defined here by $T < 10^4 + 0.4\,(T_{\text{vir}} - 10^4)$, e.g. $T < 2.5 \times 10^5$ K for $M \simeq 10^{12}\,\text{M}_\odot$, where $T_{\text{vir}} \simeq 6 \times 10^6$ K. The transition from cold dominance to hot dominance occurs near $5 \times 10^{11}\,\text{M}_\odot$ at all redshifts (vertical line), as predicted by the spherical model for similar metallicities (Dekel & Birnboim 2006).

Note that we loosely refer to the massive, rapid gas collapse event as a 'burst', having in mind the star formation event that is expected to follow but is not properly simulated here. The cold gas accretion rate could possibly be interpreted as an upper limit to the SFR.

In Section 2 we describe the simulation method. In Section 3 we present the four phases of the SAMBA phenomenon. In Section 4 we dare a tentative comparison to observations. In Section 5 we describe SAMBA recipes for semi-analytic modelling. In Section 6 we discuss our results. In Appendix A, related to Section 2, we provide an extended Press–Schechter (EPS) estimate of the average virial accretion rate on to dark haloes. In Appendix B, related to Section 5, we specify the SAMBA recipes for semi-analytic models (SAMs).

2 METHOD OF SIMULATIONS

2.1 The spherical hydrodynamical code

Our accurate 1D Lagrangian hydrodynamical code (based on the code described in Birnboim & Dekel 2003) simulates a spherical gravitating system consisting of dark matter and a fraction f_{b} of gas with a constant metallicity Z. The initial smooth density-perturbation profile is constructed such that it produces the desired average accretion rate at the virial radius, Section 2.2. The dissipationless dark matter shells detach from the cosmological expansion, collapse and oscillate into virial equilibrium such that they deepen the potential well attracting the dissipating gas shells. The gas is cooling radiatively based on the atomic cooling function computed by Sutherland & Dopita (1993), and contracts dissipatively into the

inner halo. The collapse of each gas shell is stopped at ∼5 per cent of the virial radius by an artificial centrifugal force which mimics the formation of a central disc.

The main improvements to the original code described in Birnboim & Dekel (2003) are as follows. The numerical scheme for time evolution now computes the pressure and energy of the gas shells within the forth-order Runge Kutta solution rather than externally. A force corresponding to the cosmological constant Λ has been added to the force equations, making the calculation fully consistent with the Λ cold dark matter (ΛCDM) cosmology, though with only minor effects on the evolution inside the haloes. The simulations described here consist of 1000 gas shells and 5000 dark matter shells, corresponding to a spatial resolution of ∼50 pc in the inner halo growing to ∼1 kpc near the virial radius. The energy conservation level, set by the time-steps and spatial resolution, is better than 2 per cent over today's Hubble's time. Convergence tests of some of the runs with twice the spatial resolution yielded very similar results. The baryonic fraction f_b is assumed to be 10 per cent throughout the simulations presented here, crudely taking into account certain gas mass loss due to feedback effects. Simulations with $f_b = 0.05$ and 0.15 per cent yield consistent results.

2.2 Cosmological accretion rate

The average fractional accretion rate on to haloes of mass M at time t is estimated in Appendix A using the EPS formalism, following Neistein, van den Bosch & Dekel (2006). Equation (A5) reads

$$\frac{\dot{M}}{M}(M, t) = s(M)\frac{\delta_c}{D}\frac{\dot{D}}{D}, \tag{1}$$

$$s(M) \equiv \left[\frac{2/\pi}{\sigma^2(M/q) - \sigma^2(M)}\right]^{1/2}. \tag{2}$$

Here $\sigma(M)$ is the rms linear density fluctuation on the scale corresponding to mass M, linearly extrapolated to $z = 0$ using the linear growth function $D(t)$ normalized to unity at t_0. The constants are $\delta_c \simeq 1.68$ from the spherical collapse model, and $q \simeq 2.2$ reflecting an intrinsic uncertainty in the EPS formalism. The accretion time $(\dot{M}/M)^{-1}$ is shown later on in Fig. 8 for haloes of $10^{12}\,M_\odot$ in the ΛCDM cosmology. A practical approximation for haloes of ∼$10^{12}\,M_\odot$ in ΛCDM is

$$\frac{\dot{M}}{M}(M, t) = 0.04\,\text{Gyr}\left(\frac{M}{10^{12}\,M_\odot}\right)^{0.15}(1 + z)^{2.25}. \tag{3}$$

The mass growth history of the main progenitor of a halo that ends up with mass M_0 at time t_0 is given by equation (A9):

$$M(t|M_0) = F^{-1}\left[\delta_c(D^{-1} - 1) + F(M_0)\right], \tag{4}$$

where F^{-1} is the inverse function of

$$F(M) = \int_M^\infty \frac{\text{d}m}{m\,s(m)}. \tag{5}$$

We specify in Appendix A the details of this calculation in ΛCDM.

The initial conditions of the simulation specify the mean density profile of the spherical perturbation $\bar{\delta}_i(M)$, referring to spheres of mass M at an initial time t_i in the linear regime. In order to end up with a halo mass M_0 at t_0, and have the proper average virial accretion rate throughout the halo history, we invert $M(t)$ from equation (4) to a collapse time for mass M, $t_c(M)$, and then derive the initial profile from

$$\bar{\delta}_i(M) = \delta_c \frac{D(t_i)}{D[t_c(M)]}. \tag{6}$$

3 FOUR PHASES OF ACCRETION

3.1 Generic SAMBA

Figs 2, 3 and 4 present the SAMBA phenomenon in three simulations, spanning a range of final halo masses. They display the radii of gas shells as they fall through the virial radius and eventually accrete on to the inner disc. One may tentatively and very crudely identify the disc build-up rate with maximum SFR. We see that every SAMBA consists of four distinct phases, as follows.

(i) *Cold accretion*. As long as the halo mass is below M_{sh} there is no stable shock in the halo. The gas flows cold and unperturbed toward the disc following the cosmological accretion rate at the virial radius, which is constructed here to be rather uniform. The resultant disc accretion rate is growing gradually with time, while the specific $\dot{M}_d/M_d$ is declining. For example, in the late SAMBA shown in Fig. 2, the cold accretion phase lasts till $t \simeq 4.6$ Gyr, with the temperature map showing no heating and the disc accretion rate rather uniform.

(ii) *Tentative quenching*. A shock forms near the disc upon crossing M_{sh}, and it propagates outward very rapidly. The gas that falls through the shock during its rapid expansion is heated and halted to a stop (sometimes even pulled back) before it cools and rains down. A first phase of suppressed disc accretion lasts for 1–2 Gyr, e.g. between $t \simeq 2.7$ and 4.3 Gyr in Fig. 3.

(iii) *Massive burst*. The gas that was slowed down in the previous quenching phase now joins new accreting gas in a big crunch on to the disc. The accumulation of shells in a massive collapse results
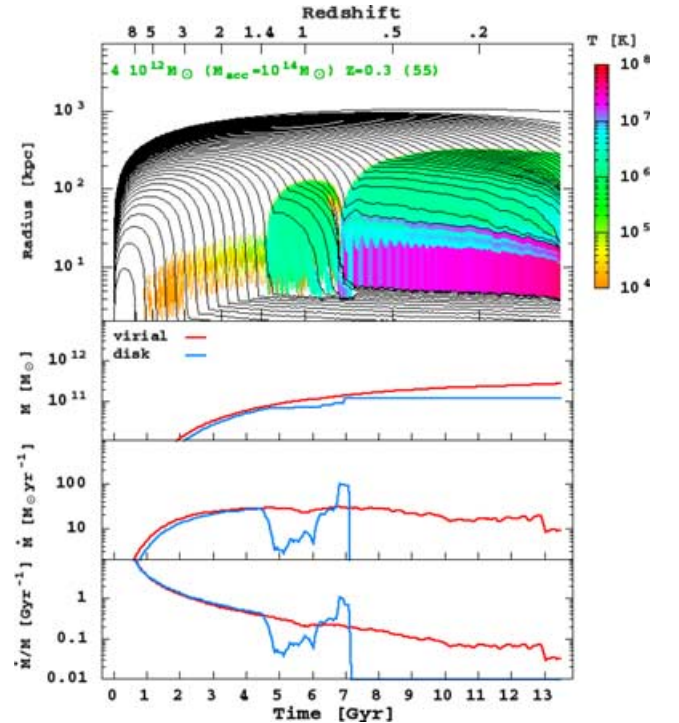

Figure 2. Spherical accretion on to the outer dark halo and the inner 'disc'. Top: infalling gas shells, with the temperature highlighting the evolution of the virial shock. Bottom: baryonic mass growth, accretion rate and specific accretion rate (smoothed over a virial crossing time), at the virial radius (red curve) and at the 'disc' (blue curve). A late SAMBA: today's halo mass $M_0 = 4 \times 10^{12}\,M_\odot$, metallicity $Z = 0.3$ and a reduced accretion rate, leading to a tentative quenching onset at $z_1 \simeq 1.1$, a burst at $z_b \simeq 0.7$ with efficiency $B \simeq 5$, followed by very effective long-term quenching.
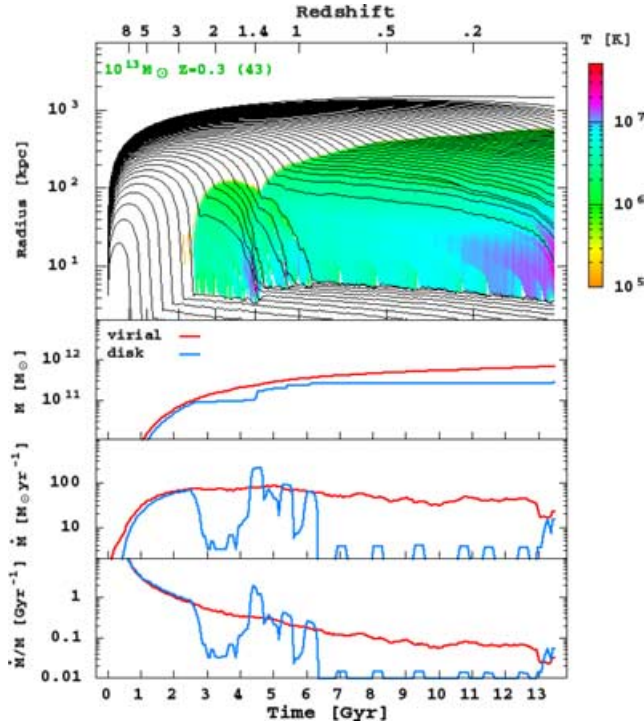
Figure 3. Same as Fig. 2 for an intermediate SAMBA: $M_0 = 10^{13}$ M$_\odot$, $Z = 0.3$, typical accretion rate, $z_1 \simeq 2.5$, $z_b \simeq 1.4$ with $B \simeq 6$, followed by effective long-term quenching.
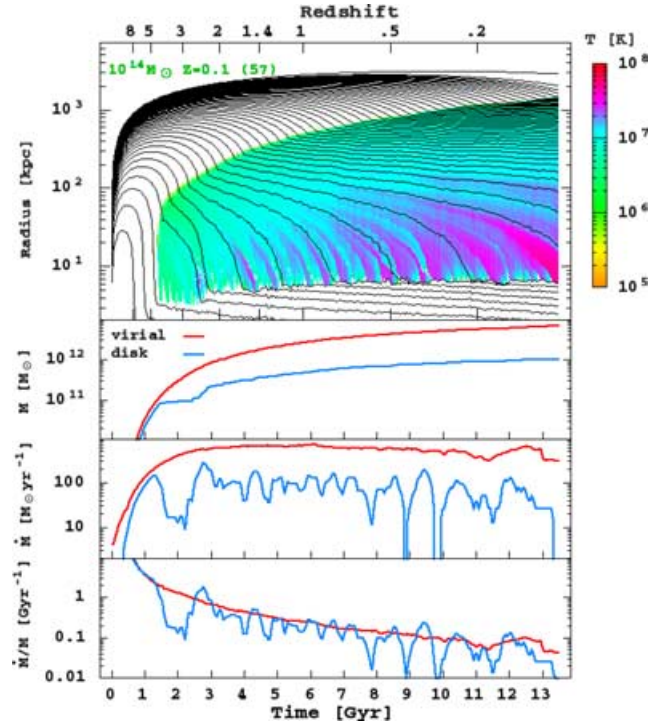

Figure 4. Same as Fig. 2 for an early SAMBA: $M_0 = 10^{14}$ M$_\odot$, $Z = 0.1$, typical accretion rate, $z_1 \simeq 4.8$, $z_b \simeq 2.5$ with $B \simeq 2.3$, followed by ineffective quenching.

from the longer delay suffered by gas that falls in through the shock earlier, when the shock expands faster. The collapse is enhanced further when the shock develops a temporary instability causing it to contract back to subvirial radii. Thus, $\sim 10^{11}$ M$_\odot$ of gas enter the disc during $\sim 0.5 - 1$ Gyr, producing a burst of $\dot{M}_d \sim 100 - 200$ M$_\odot$ yr^{-1}, or $\dot{M}_d / M_d \sim 1 - 2$, typically twice the specific virial accretion rate.

(iv) *Post-burst slowdown or shutdown.* The halo mass is safely above the threshold, with the cooling rate significantly slower than the compression rate, allowing the shock to join the gradual expansion of the virial radius. Bursts that occur prior to $z \sim 1.4$ are followed by a long period of gradually declining specific accretion rate, which is comparable to the specific virial accretion rate. Bursts that occur after $z \sim 1.4$ are followed by an effective long-term shutdown from $z \sim 0.7$ and on, lasting for several Gyr. The infall is halted effectively while crossing the shock during its second expansion phase.

The timing of the SAMBA events scales with the onset time, t_1. A relevant characteristic time-scale at that epoch is the virial crossing time, the time it takes to stream at the virial velocity from the halo virial radius to the centre. Based on the spherical collapse model, it is independent of mass and given by[1]

$$t_{\rm vir} \equiv \frac{R_{\rm v}}{V_{\rm v}} \simeq 2.49 \, {\rm Gyr} \, (1 + z)^{-3/2}. \tag{7}$$

In the Einstein–de Sitter phase of cosmological evolution (valid to a good approximation prior to $z \sim 1$), $t_{\rm vir} \simeq 0.18 t_{\rm Hubble}$.

[1] For a proper virial radius, $(1 + z)^{-1}$ should be replaced by $(\Delta_{200}\Omega_{\rm m0.3}h_{0.7}^2)^{-1/3}(1 + z)^{-1}$. In a flat universe $\Delta(a) \simeq [18\pi^2 - 82\Omega_\Lambda(a) - 39\Omega_\lambda(a)^2]/\Omega_{\rm m}(a)$, with $a = (1 + z)^{-1}$, $\Omega_{\rm m}(a)$ given in Appendix A and $\Omega_\Lambda = 1 - \Omega_{\rm m}$.

We define the efficiency of disc accretion by the ratio of specific accretion rate into the disc and into the virial radius,

$$B \equiv \frac{\dot{M}_d/M_d}{\dot{M}/M}. \tag{8}$$

A smoothed version of this quantity over a virial crossing time, $B_{\rm v}$, which can be read as the ratio of the two curves shown at the bottom panels of Figs 2–4, serves us for an automatic identification of the four phases. This quantity is about unity during the cold-flow phase. We mark by t_1 the time when $B_{\rm v}$ starts its first drop below unity due to the shock onset. Then t_2 and t_3 mark the subsequent crossings of unity upward and downward, which we define as the beginning and the end of the burst. Finally, we identify the following upward crossing of unity, at t_4, as the end of the long-term quenching phase, when such a phase exists.

There are several robust SAMBA characteristics that only weakly depend on the redshift of the event. The onset of the SAMBA is always at the crossing of the threshold halo mass, $M_{\rm sh} \sim 10^{12}$ M$_\odot$, thus involving a stellar mass $M_d \sim 10^{11}$ M$_\odot$. The burst peaks about half a Hubble time ($1.5 - 2.5$ Gyr) later, and involves $\lesssim 10^{11}$ M$_\odot$ of rapidly accreting gas in about a virial crossing time $t_{\rm vir}(\sim 0.5 - 1$ Gyr). The average burst efficiency is $B_{\rm v} \simeq 2$, with a peak value of $B \sim 6$.

3.2 Systematic redshift dependence

By simulating the accretion histories of haloes that end up with different masses M_0 today, we obtain a sequence of SAMBAs with a range of starting times, t_1, given that the SAMBA onset is determined by the crossing of a fixed halo mass threshold. Our fiducial case assumes the average virial accretion history on to haloes M_0 and a constant metallicity $Z = 0.3$ solar; we obtain somewhat earlier
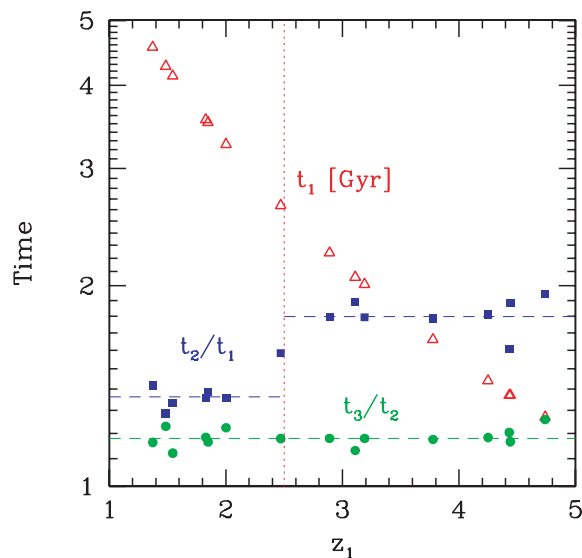
Figure 5. Characteristic times of simulated SAMBAs as a function of the onset redshift z_1, when $M = M_{sh}$. The Hubble time at the onset of tentative quenching is t_1 (open red triangles). The ratio t_2/t_1 (filled blue squares) refers to the duration of the tentative quenching. The ratio t_3/t_2 (filled green circles) refers to the duration of the burst, with t_3 the beginning of long-term quenching in late SAMBAs. The vertical line marks the transition from early to late SAMBAs. The horizontal lines mark the fits used for the SAM recipe in Section 5.

Figure 6. SAMBA properties as a function of redshift: disc accretion properties. Red triangles refer to the tentative quenching, at z_1 (top panels) or $z_q \equiv 0.5(z_1 + z_2)$ (bottom panels). Blue squares refer to the burst at $z_b \equiv 0.5(z_2 + z_3)$. Vertical lines separate early from late SAMBAs. From top to bottom: disc mass, fractional disc growth, accretion rate and specific accretion rate.

(later) onset times by lowering (raising) Z or by raising (lowering) the virial accretion rate at the relevant epoch. It turns out that our current fiducial choice of $Z = 0.3$ at all times does not follow the gradual cosmological growth of metallicity with time as estimated in Dekel & Birnboim (2006). This results in an artificial overestimate of M_{sh} at high z, by a factor of ~ 2, which we should scale out when considering the cosmological redshift dependence of the SAMBAs.

The characteristic times of 15 simulated SAMBAs are shown in Fig. 5. These times naturally scale with the Hubble time. The burst duration is always $\Delta t_b \simeq 0.18\, t_2 \simeq t_{vir}$. The duration of the tentative quenching is $\Delta t_q \simeq 0.8\, t_1$ for $z_1 > 2.5$ and $\Delta t_q \simeq 0.36 t_1$ for $z_1 \leqslant 2.5$. The systematic variations with the onset time are primarily driven by the gradual decline with time of the specific virial accretion rate, which induces a drastic change between SAMBAs that start prior to $z \simeq 2.5$ and those that start later. This is a distinction between the central galaxies of today's clusters of $> 10^{13}\, M_\odot$ haloes and the more isolated galaxies in today's groups of $\leqslant 10^{13}\, M_\odot$ haloes. The former burst before $z \sim 1.4$, when $\dot{M}/M > 1\, \mathrm{Gyr}^{-1}$, while the latter burst later, when $\dot{M}/M < 1\, \mathrm{Gyr}^{-1}$. In the late SAMBAs, the rapidly expanding shock overshoots to outside the virial radius, where the low accretion rate fails to provide enough post-shock compression for support against contraction. The resultant temporary instability makes the shock drop to subvirial radii before the pressure behind it can push it back to the virial radius. Apparently, this instability appears once the specific virial accretion rate drops below $1\, \mathrm{Gyr}^{-1}$.

The effect of the virial accretion rate on the tentative shock instability is clearly demonstrated in Figs 2–4. The strength of the instability as measured by the tentative shrinkage of the shock to subvirial radii, as seen in the top panels, is correlated with the specific virial accretion rate at shock onset, which could be read from the red curve at the bottom panels or from the inverse of the accretion time shown in Fig. 8. This trend is valid for the whole suite of simulated SAMBAs.
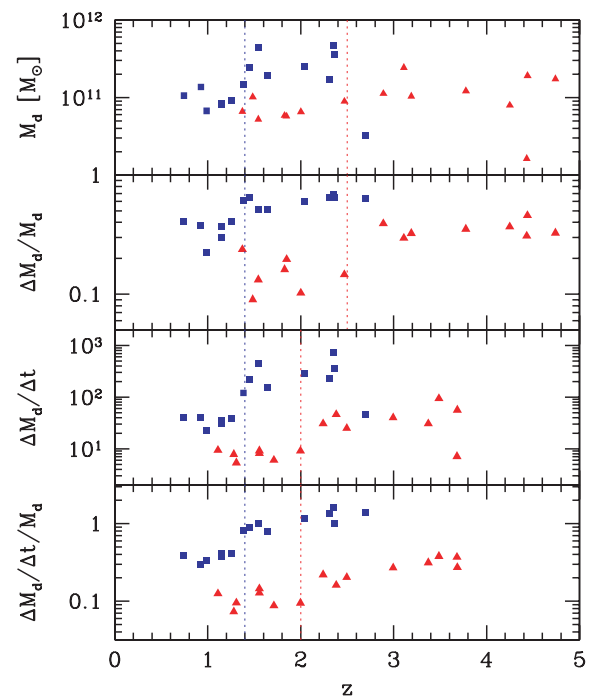
Other quantities characterizing the tentative quenching and the subsequent burst are summarized in Figs 6 and 7. Most of the burst characteristics vary gradually with the burst time $t_b \equiv 0.5(t_2 + t_3)$, and show a bimodality to early and late bursts separated at $z \simeq 1.4$, associated with the shock instability. While in the late bursts the disc mass is $M_d \sim 10^{11}\, M_\odot$, it is typically twice as large in the early burst, where the specific accretion rate is higher and the preceding quenching is less efficient. The fraction of disc mass involved in the burst between t_2 and t_3, $\Delta M_d/M_d$, varies from ~ 0.7 in the early bursts to ~ 0.3–0.4 in the late ones. The mean specific accretion rate, $(\Delta M_d/\Delta t)/M_d$, varies from ~ 1–1.5 to $\sim 0.4\, \mathrm{Gyr}^{-1}$. This corresponds to a drop of the average accretion rate across the burst from $\dot{M}_d \sim 100$–$400\, M_\odot\, \mathrm{yr}^{-1}$ in the early bursts to ~ 30–$40\, M_\odot\, \mathrm{yr}^{-1}$ in the late bursts. The average burst efficiency over the burst is always $B_b = (\Delta M_d/M_d)/(\Delta M/M) \simeq 2$, while the peak efficiency changes from ~ 2 in the early bursts to 3–6 in the late bursts.

The ratio of virial to disc mass at the end of the burst is bi-modal: it is high for the early bursts, $M/M_d \sim 25$–50, and low for the late bursts, $M/M_d \sim 10$–15. The high values in the early bursts is because the high virial accretion rate there leads to a rapid halo growth during the quenching phase combined with the failure of the subsequent burst to consume most of the accumulated hot gas. In the late SAMBAs, on the other hand, the halo growth is slow and the burst consumes all the available gas, leading to a low M/M_d. The early bursts are thus predicted to reside in very massive haloes, of $M \lesssim 10^{13}\, M_\odot$, while the late bursts are to be found in $\sim 10^{12}\, M_\odot$ haloes.

The characteristics of the tentative quenching phase also vary gradually with t_1. At the onset of all SAMBAs, $M \sim 10^{12}\, M_\odot$ and
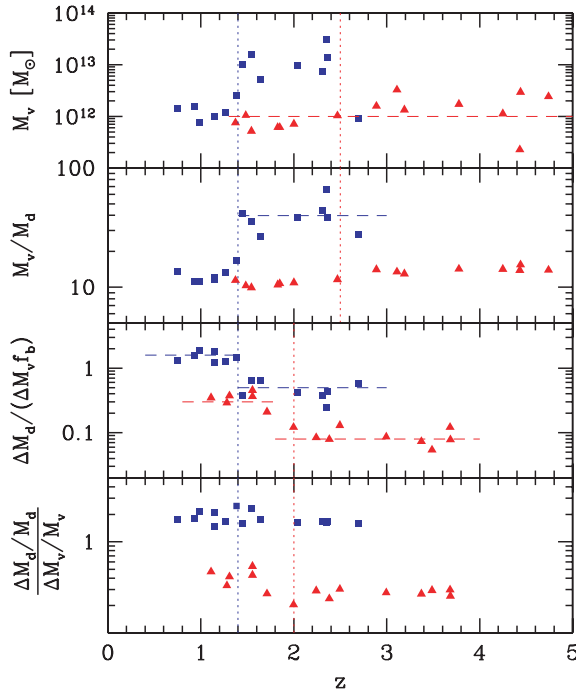
Figure 7. SAMBA properties as in Fig. 6: disc versus virial accretion. From top to bottom: virial mass, virial-to-disc mass ratio, disc-to-virial accretion ratio and efficiency B of specific-accretion ratio. Horizontal lines mark the fits used for SAM recipes in Section 5.

$M_d \sim 10^{11} \, M_\odot$, and we should scale out the apparent slight decline with time of these masses, which is largely an artefact of the assumed constancy of metallicity with time. The typical quenching efficiency is $B \sim 0.3$ for the early SAMBAs and $B \sim 0.4$ for the late ones. The ratio of disc to virial gas accretion rate, $\Delta M_d/(\Delta M f_b)$ varies from $\lesssim 0.1$ to ~ 0.3 in the early and late SAMBAs, respectively, separated at onset redshift $z \simeq 2.5$. The tentative quenching can make the galaxy appear red for ~ 1 Gyr before bursting.

The most pronounced effect of the shock instability occurring in the late SAMBAs is a drastic shutdown of disc accretion for several Gyr after the burst, starting at $z \simeq 1.2$ or later. The infalling gas is halted and pushed back after crossing the shock during its second rapid expansion phase, and the fact that the halo is already well above M_{sh} makes the cooling and infall time long. These galaxies, in haloes $\leqslant 10^{13} \, M_\odot$ today, will appear red and dead at $z \leqslant 1$, without the help of any additional feedback effect. On the other hand, the post-burst phase of the early SAMBAs is characterized by a disc specific accretion rate that fluctuates about the virial rate, $B \sim 1$. For these central cluster galaxies today to become red and dead one should appeal to another quenching mechanism that could provide long-term maintenance.

4 COMPARISON WITH OBSERVATIONS

4.1 Maximum bursting discs

Two relevant examples of observed 'maximum bursters' stand out. First, rest-frame UV/optically selected star-forming galaxies at $z \gtrsim 2$ (termed BX and BzK). A high-resolution adaptive-optics infrared (IR) study of BzK-15504 at $z \simeq 2.4$ (Genzel et al. 2006) reveals $M_d \simeq 8 \times 10^{10} \, M_\odot$ plus $M_{gas} \simeq 4 \times 10^{10} \, M_\odot$ forming stars at $\sim 100-200 \, M_\odot \, yr^{-1}$. The stellar population has a typical age of only

~ 0.5 Gyr, less than 20 per cent of the Hubble time at that epoch. The morphology and kinematics fit an exponential disc of scale radius $\simeq 4.5$ kpc rotating at $V_{rot} \simeq 230 \, km \, s^{-1}$, and a recent major merger is ruled out. The velocity dispersion of $\sigma \simeq 60-110 \, km \, s^{-1}$ indicates a rapid accretion on to the disc, again of ~ 0.5 Gyr. Several galaxies with similar properties were detected in a lower resolution study (Förster Schreiber et al. 2006).

Other examples are found among the LIRGs at $z \sim 0.6-0.9$. Hammer et al. (2005) have detected at least six galaxies of $M_d \sim 10^{11} \, M_\odot$ forming stars at $\simeq 120-200 \, M_\odot \, yr^{-1}$ while their morphologies and kinematics resemble spiral galaxies with no trace of major mergers. More than half of their IR luminous galaxies are spirals. The compilation of SFR at $z \sim 1$ by Noeske et al. (2007) reveals some massive galaxies that form stars at $\sim 200 \, M_\odot \, yr^{-1}$, while only ~ 7 per cent of the galaxies show evidence for ongoing major mergers (Lotz et al. 2006), and this fraction is not higher for the galaxies with the highest SFR (private communication).

In summary, the observed maximum bursters show the following robust features: (a) the baryonic mass is on the order of $M_d \sim 10^{11} \, M_\odot$; (b) a substantial fraction of the baryonic mass is bursting; (c) the burst is short compared to the Hubble time and the cosmological mean accretion time, and is comparable to the virial crossing time; (d) the star formation occurs in an extended configuration of a few kpc and (e) in many cases there are morphological and kinematical indications for a thick rotating disc with no trace of major mergers.

A distinct feature of the maximum bursts is that most of the stars of the $M_d \sim 10^{11} \, M_\odot$ galaxies have formed over a short period of time just prior to the observed epoch, ~ 0.5 Gyr at $z \gtrsim 2$ and ~ 1 Gyr at $z \lesssim 1$. This is compared in Fig. 8 to several relevant characteristic time-scales. The burst duration is shorter than the age of the universe at that time by a factor ~ 6. This implies that most of the stars have formed in a coherent starburst of the whole $\sim 10^{11} \, M_\odot$ of gas in its late stages of collapse or assembly rather than in the smaller
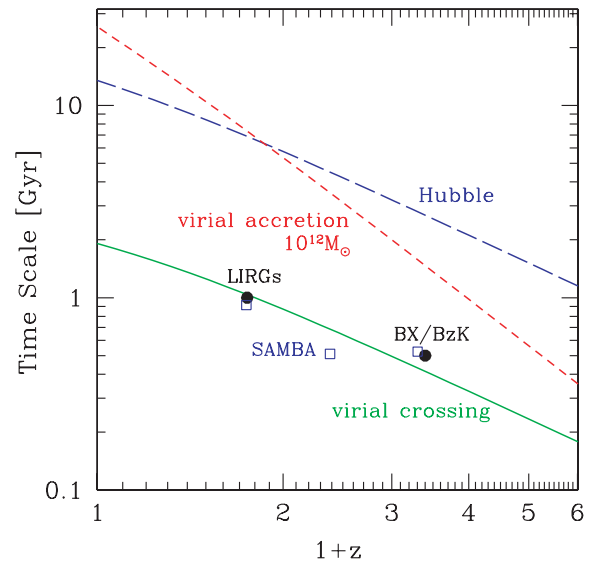

Figure 8. Characteristic time-scales as a function of redshift. The curves refer to (a) the Hubble time (blue, long dash), (b) the average cosmological accretion time on to $M = 10^{12} \, M_\odot$ haloes (red, short dash) and (c) rapid streaming from the virial radius to the centre, R/V (green, solid). The estimated duration of the observed maximum starbursts are represented by black circles. The predicted duration of three SAMBA bursts are marked by blue open squares.

progenitors during the preceding period of hierarchical build-up. The implied suppression of SFR in small haloes and its possible origin will be discussed in Section 6.

The burst duration is also shorter than the characteristic time for cosmological accretion on to dark haloes of $\sim 10^{12}\,\mathrm{M_\odot}$, the likely hosts of the $M_d \sim 10^{11}\,\mathrm{M_\odot}$ maximum bursters, as given by the inverse of the average $\dot{M}/M$ in equation (1). At $z = 0.75$, one has $t_{\mathrm{acc}}(10^{12}\,\mathrm{M_\odot}) \simeq 7\,\mathrm{Gyr}$, much longer than the indicated observed duration of $\Delta t \sim 1\,\mathrm{Gyr}$. At $z = 2.4$, one has $t_{\mathrm{acc}}(10^{12}\,\mathrm{M_\odot}) \simeq 1.6\,\mathrm{Gyr}$, which is about three times the observed $\Delta t \sim 0.5\,\mathrm{Gyr}$, though if the relevant haloes are more massive, $\lesssim 10^{13}\,\mathrm{M_\odot}$, the accretion time is ~ 50 per cent shorter. As seen in Fig. 8, the burst durations are rather comparable to the shortest characteristic time of the system, t_{vir}.

Gas-rich major mergers, which could in principle produce such rapid, massive bursts, fail to match the robust features (a), (d) and (e). In particular, major mergers are expected to destroy the discs and be associated with perturbed clumpy morphology and kinematics rather than a smooth rotating disc (Mihos & Hernquist 1996; Cox 2004; Lotz et al. 2006) (though a secondary disc may regrow after the merger is complete Springel & Hernquist 2005; Naab, Jesseit & Burkert 2006a). This is because the dynamical disturbance, involving tidal effects and violent relaxation, is expected to be pushed to an extreme during the $\sim 100\,\mathrm{Myr}$ of the close passages. Furthermore, the induced star formation is expected to be concentrated in a central cusp rather than an extended disc (Cox et al. 2006). It is worth noting that the mergers are only weakly associated with a characteristic mass, the non-linear clustering scale M_*, which varies rapidly with time – they are not associated with galaxies of any specific fixed mass.

The SAMBA bursts, as described by our naive spherical model above, have the potential of reproducing the observed features one by one in a straightforward way. In particular, they have the potential of keeping the disc intact while doubling its mass and puffing up its thickness. This is because the disc build-up, even if it involves filamentary and clumpy cold flows, could be rather uniformly stretched over the $\sim 0.5-1\,\mathrm{Gyr}$ duration of the burst, and therefore involve only moderate perturbative effects on the dynamics (yet to be studied in detail). In addition, the SAMBA bursts are predicted to be specifically associated with one characteristic baryonic mass of $\sim 10^{11}\,\mathrm{M_\odot}$ at all redshifts.

The early SAMBA bursts, which may be associated with the observed maximum bursters at $z \gtrsim 2$, are predicted to be in $M_d \sim 2 \times 10^{11}\,\mathrm{M_\odot}$ discs embedded in rather massive haloes of $M \sim 6 \times 10^{12}\,\mathrm{M_\odot}$. The comoving number density of haloes of $M > 5 \times 10^{12}\,\mathrm{M_\odot}$ at $z = 2.2$ is $n \simeq 10^{-4}\,(h^{-1}\,\mathrm{Mpc})^{-3}$, consistent with the number density of observed maximum bursters (Genzel et al. 2006).[2] The predicted halo virial velocity is $\sim 400\,\mathrm{km\,s^{-1}}$ and the virial radius is $\sim 200\,\mathrm{kpc}$. The low concentration expected for such haloes at $z \sim 2.2$, $C \sim 4$ (Bullock et al. 2001b), implies that the circular velocity at $\sim 10\,\mathrm{kpc}$ should be $\sim 200\,\mathrm{km\,s^{-1}}$, compatible with the observed bursters. The large virial radius for that epoch, with an assumed typical spin parameter, implies an exponential disc radius of $\sim 5\,\mathrm{kpc}$ (Bullock et al. 2001a), which explains the observed extended discs.

[2] This estimate is for the *WMAP*3 (*Wilkinson Microwave Anisotropy Probe*) parameters of the ΛCDM cosmology, $\Omega_m = 0.24$, $\Omega_\Lambda = 0.76$, $h = 0.73$ and $\sigma_8 = 0.74$ (Spergel et al. 2007). A similar number density is obtained for haloes twice as massive if one adopts instead $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $h = 0.7$ and $\sigma_8 = 0.9$.
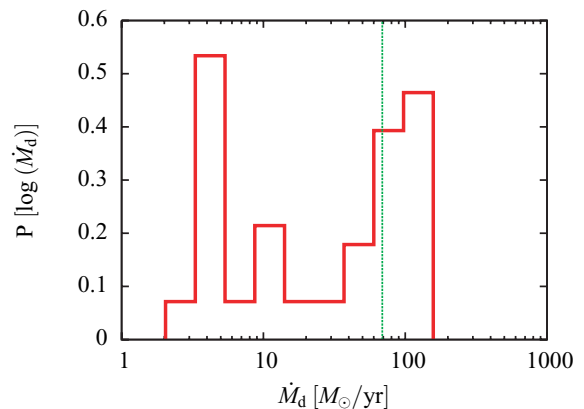

Figure 9. Predicted distribution of log disc accretion rate for galaxies in the disc-mass bin $(0.5-2) \times 10^{11}\,\mathrm{M_\odot}$ near $z \sim 2$. The vertical green line marks the mean virial accretion rate. About half the galaxies are expected to show a high accretion rate into the disc, while the other half is predicted to suffer a strongly suppressed accretion rate.

Thus, the burst properties predicted by the idealized spherical model do surprisingly well in matching the observed bursts.

4.2 Suppressed star formation rate

Despite the apparent vigour of the burst event, the most robust feature of the SAMBAs is the quenching. We note that during the SAMBA sequence of events, from the onset of the shock till the end of the burst about a Hubble time later, each galaxy spends more time in a phase of suppressed accretion rate rather than in a rapid accretion mode. Fig. 9 shows a crude estimate of the predicted distribution of disc accretion rate for galaxies in the stellar mass bin $(0.5-2) \times 10^{11}\,\mathrm{M_\odot}$ near $z \sim 2$. This crude estimate is based on the fraction of time spent at each accretion rate in the simulated SAMBA shown in Fig. 3, assuming that it is a representative SAMBA at that epoch. While a fraction of the galaxies are expected to show a high accretion rate into the disc at about twice the average virial level, more than half the galaxies are predicted to present a strongly suppressed disc accretion rate, an order of magnitude below the virial accretion rate or less. The associated distribution of SFR may be qualitatively similar.

Preliminary indications for a substantial population of $\sim 10^{11}\,\mathrm{M_\odot}$ galaxies with strongly suppressed SFR already exist. A bimodality in the SFR distribution is apparent at $z \lesssim 1$ in Noeske et al. (2007, fig. 1), where the typical SFR is $> 10\,\mathrm{M_\odot\,yr^{-1}}$ while about half the galaxies have an upper limit of SFR $< 1\,\mathrm{M_\odot\,yr^{-1}}$. At $z \sim 2-2.7$, Kriek et al. (2006) report that a significant fraction of the $\sim 10^{11}\,\mathrm{M_\odot}$ galaxies have surprisingly low upper limits for the specific SFR at $< 0.1\,\mathrm{Gyr^{-1}}$, while the typical is $\sim 1\,\mathrm{Gyr^{-1}}$.

The SAMBAs predict that the SFR histories of galaxies should show two peaks of SFR separated by a couple of Gyr, but this might be hard to detect observationally.

4.3 Downsizing

The SAMBA phenomenon provides a simple explanation for the observed 'downsizing' of elliptical galaxies. An 'archaeological' analysis of stellar ages reveals that today's more massive ellipticals have formed their stars earlier and over a shorter time-span (Thomas et al. 2005). From their SFR histories as a function of

look-back time one learns that elliptical galaxies of present-day stellar mass $\simeq 10^{11} - 10^{12}\,M_{\odot}$ have formed most of their stars at $z \sim 2-3$ in a maximum burst over $1-0.5$ Gyr with $\dot{M}_d/M_d \sim 2$ (or $z \sim 1.2-1.6$ in lower density environments). The progenitors of today's ellipticals, prior to their subsequent growth via dry mergers, seem to be consistent with the maximum bursters seen at high redshift, and with the SAMBA predictions.

The constancy of M_{sh} with redshift, which implies that all galaxies burst when they are $M_d \sim 10^{11}\,M_{\odot}$ and quench immediately after automatically implies downsizing in both the peak of SFR and the subsequent quenching. Galaxies of larger stellar mass today are likely to be embedded in more massive haloes, which in turn have crossed the threshold mass for the onset of SAMBA at an earlier time. They bursted earlier, for a shorter duration, and shut down earlier accordingly. This is demonstrated in a cosmological semi-analytic simulation by Cattaneo, Dekel & Faber (in preparation).

The SAMBAs predict yet another kind of downsizing phenomenon, concerning halo mass at burst: the early bursts at $z > 1.4$ are predicted to reside in massive haloes only slightly smaller than $M \sim 10^{13}\,M_{\odot}$, while the late bursts are to be found in smaller haloes, $M \sim 10^{12}\,M_{\odot}$.

5 RECIPES FOR SEMI-ANALYTIC MODELS

In order to study the potential observable implications of the SAMBA phenomenon in more detail, one can translate the predictions from the spherical model into a recipe for gas cooling and accretion rate on to the disc. This recipe is to be incorporated in SAMs, which follow the baryonic physics in given dark-halo merger trees. The SAMBA recipe should replace the standard recipe adopted in SAMs (White & Frenk 1991), which assumes that the halo gas interior to a 'cooling radius' is accreted on to the disc. The cooling radius within a (spherical) halo of a given mass at a given time is traditionally determined by the position where the cooling time equals a certain dynamical time e.g. the Hubble time. The SAMBA phenomenon suggests a deterministic modification of the disc accretion rate, based on our spherical modelling, with no free parameters. However, an optimal match to observations may suggest a certain fine-tuning of the model parameters, compensating for deviations from spherical symmetry and uniform accretion.

5.1 Following halo evolution in SAM

The proposed SAMBA recipe for the accretion rate on to the disc is specified in Appendix B (Section B1). In summary, given the mass growth of each halo in the merger tree, $M(t)$, and its derivative $\dot{M}$, the SAMBA onset time t_1 (redshift z_1) is identified by $M(t_1) = M_1 = 10^{12}\,M_{\odot}$, and the SAMBA is classified as 'early type' or 'late type' according to whether $z_1 > 2.5$ or $z_1 \leqslant 2.5$, respectively. Then the times t_2 and t_3, the beginning and the end of the burst, are computed using the fits shown in Fig. 5. The algorithm then specifies the growth rates of cold and hot gas mass during each of the four SAMBA phases defined by the above times. The numerical factors used for the tentative quenching and for the burst are read from straightforward fits to $\Delta M_d/(\Delta M f_b)$ in Fig. 7.

The disc mass growth according to the proposed SAMBA recipe for SAM is shown in Fig. 10 for typical early and late SAMBAs. It is compared to the standard disc growth rate following the evolution of a 'cooling radius'. The latter has a characteristic drop when the cooling radius first becomes smaller than the virial radius. The cooling-radius recipe can be somewhat improved once the Hubble time used in its definition is replaced by $t_{vir} = 0.18\,t_{Hubble}$. Yet, it
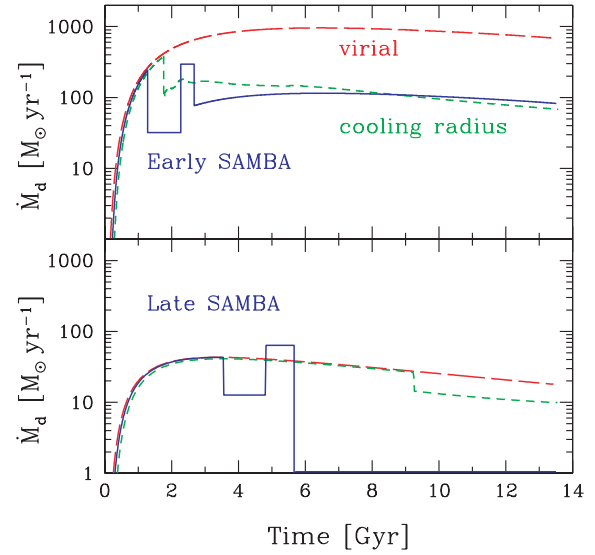

Figure 10. Disc growth rate in the SAMBA recipe for SAM (blue, solid) versus the smooth virial baryonic accretion rate (red, long dash) and the SAM standard (green, short dash) based on the evolution of the 'cooling radius' (White & Frenk 1991). Top: an early SAMBA ($z_1 \simeq 4.8$, $M_0 \simeq 10^{14}\,M_{\odot}$). In this case, long-term quenching has to originate from another process. Bottom: a late SAMBA ($z_1 \simeq 1.8$, $M_0 \simeq 4 \times 10^{12}\,M_{\odot}$), where long-term quenching is a natural outcome of uniform accretion.

fails to capture the depth of the tentative quenching, the burst and the long-term quenching of late SAMBAs.

Recall that the early SAMBAs, ending up in $M \geqslant 10^{13}\,M_{\odot}$ haloes today, should be quenched after t_3 by another mechanism, such as clumpy accretion or AGN feedback. If such a process is not explicitly included in the SAM, one can simply set $\dot{M}_d = 0$ in the post-burst phase for all haloes (as in Cattaneo et al. 2006).

5.2 SFR as a function of halo mass: FHM

An alternative, quick and crude way to learn about the implications of SAMBAs is by providing a recipe for the SFR as a function of halo mass and redshift, to be incorporated in a given halo population in a partial SAM while bypassing the detailed baryonic processes of cooling, star formation and feedback [formation history modelling (FHM); Croton & Dekel, in preparation The fact that the SAMBA phenomenon is driven by halo mass allows us to provide a concrete recipe of this sort, specified in Appendix B (Section B2).

In summary, one starts with $M_1 = 10^{12}\,M_{\odot}$ at t_1, and computes the other SAMBA characteristic halo masses M_2 and M_3, corresponding to the characteristic times t_2 and t_3, using fits to our simulation results. Given the virial accretion rate in the provided merger tree, $\dot{M}$, the disc accretion rate, which could be interpreted as an upper limit for the SFR, is given in each of the mass ranges, corresponding to the four SAMBA phases. The numerical values are read from crude fits to the simulation results for $\Delta M_d/(\Delta M f_b)$, as shown in Fig. 7. When implementing the recipe, one can either use the actual mass growth rate of the haloes drawn from merger-tree realizations, or the average virial accretion rate as provided in equation (3).

The SAMBA recipe for disc accretion rate is shown in Fig. 11, compared to the standard rate based on a 'cooling radius'. Again, the cooling-radius recipe can be somewhat improved once the Hubble time is replaced by t_{vir}, but it would still fail to recover the burst and the deep quenching features of the SAMBAs.
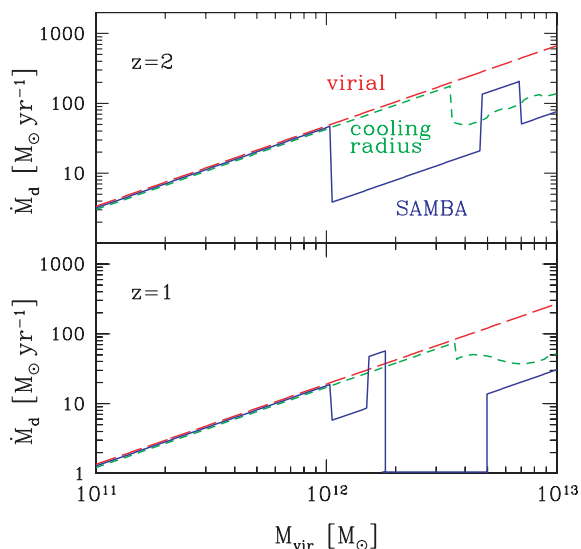
Figure 11. Disc growth in the SAMBA recipe for FHM (blue, solid) versus the virial baryonic accretion rate (red, long dash) and the standard based on 'cooling radius' (green, short dash). Top: at $z = 2$. Bottom: at $z = 1$. Quenching in haloes $\sim 10^{13}\,M_\odot$ and larger has to be due to another process.

The main feature of the SAMBA FHM is the quenching of SFR above a threshold halo mass, $M_{sh} \sim 10^{12}\,M_\odot$, as already implemented in SAM simulations (Cattaneo et al. 2006; Croton et al. 2006). The interesting new feature predicted by the model is the high SFR in a second range of masses above M_{sh}. At low redshifts, this is a slight increase of the effective threshold mass for quenching to $\sim 2 \times 10^{12}\,M_\odot$. At $z \geqslant 1.3$, the high SFR is predicted to extend to haloes as large as $\sim 7 \times 10^{12}\,M_\odot$. This should allow the appearance of massive blue galaxies at high redshifts. In the special range $1.24 < z < 1.53$, the high SFR is expected over an especially broad range of masses, $1.5-7 \times 10^{12}\,M_\odot$ (not shown in Fig. 11). Given that $10^{12}\,M_\odot$ is the typical mass for haloes that form at this redshift range, this should lead to a peak in the history of cosmological SFR density (the 'Madau' plot).

6 DISCUSSION AND CONCLUSION

A straightforward spherical model of uniform gas accretion on to haloes predicts a robust, rich sequence of events in the accretion on to the central discs, SAMBA, which could be the major driver of star formation history in galaxies. An initial cold accretion phase ends abruptly upon the birth of a shock in the inner regions of a halo once its mass exceeds a threshold mass of $\sim 10^{12}\,M_\odot$ and the disc contains $\sim 10^{11}\,M_\odot$. The infall of gas through the expanding shock creates a period of tentative quenching, where the accretion on to the disc is suppressed for a couple of Gyr. The accumulating hot gas joins new incoming gas in a big collapse, leading to a massive, rapid, burst involving $\sim 10^{11}\,M_\odot$ of gas in a few hundred Myr. The subsequent expansion of the shock to the virial radius results in a long-term shutdown, which is especially efficient if the burst occurs after $z \sim 1.4$, corresponding to today's haloes of $\leqslant 10^{13}\,M_\odot$. The predicted SAMBA bursts seem to match the observed maximum bursters of $M_d \sim 10^{11}\,M_\odot$ at $z \lesssim 1$ and at $z \gtrsim 2$, as well as the indicated population of $M_d \sim 10^{11}\,M_\odot$ galaxies with suppressed SFR. This rich phenomenon is a generic consequence of uniform accretion, before including the effects of major mergers or AGN feedback.

The implications of the SAMBA phenomenon should be taken with a grain of salt because the relevance of the spherical uniform gas-accretion model to the actual SFR history in real galaxies is yet to be demonstrated. This is an immediate challenge for cosmological simulations. The actual accretion could be rather filamentary and clumpy to begin with, and one should verify the extent to which the SAMBA behaviour remains valid under such conditions. The massive, rapid collapse of cold gas, which we loosely referred to as 'burst', is actually only an upper limit to the subsequent SFR, which depends on the details of the star formation process and can be slowed down by feedback processes (e.g. Ciotti & Ostriker 2007). One could be encouraged though by the success of the spherical model in predicting the threshold mass for virial shock heating (Kereš et al. 2005; Dekel & Birnboim 2006; Birnboim et al., in preparation). First clues for natural long-term quenching of the sort predicted here in a $\sim 10^{12}\,M_\odot$ halo may be seen in high-resolution simulations of galaxies with quiet histories (Naab et al. 2007; Libeskind & Dekel, in preparation). The following discussion of possible implications is pursued under the tentative, daring assumption that other robust SAMBA features will also be reproduced in cosmological simulations of the formation of such galaxies. The apparent match of the spherical predictions with several different observations makes this discussion worthwhile.

Using the fact that the accretion on to the disc is a straightforward function of halo mass, we propose a simple accretion recipe for SAMs to replace the standard recipe based on a 'cooling radius'. A realistic study of the SFR that is associated with this galaxy build-up, both in the case of smooth accretion of minor mergers and under major mergers and feedback effects, is necessary for a reliable comparison with observations. Still, the naive interpretation of the disc accretion rate as the maximum possible SFR provides a preliminary indication for a surprisingly good match to observations.

An obvious necessary condition for SAMBAs (which is actually automatically implied by the observed maximum bursts independent of any model) is that a substantial fraction of the accreting material on to haloes of $\sim 10^{12}\,M_\odot$ is gaseous rather than stellar. Star formation could have been suppressed in haloes smaller than a certain threshold due to different processes, such as stellar feedback or a density threshold for star formation. Given such a threshold mass, we can estimate the gas fraction in the accretion by evaluating the mass in all accreting haloes below that threshold compared to the total accreting mass (Neistein & Dekel, in preparation). We find using the EPS formalism specified in Appendix A that about one half of the accreting mass on to a halo of $10^{12}\,M_\odot$ is in haloes smaller than $\sim 10^{11}\,M_\odot$ (or 'smooth' accretion), independent of time. This should therefore be the minimum halo mass for star formation in order to have a gas-dominated accretion. If the suppression is due to supernova feedback, the minimum mass for star formation corresponds to haloes of virial velocity $\sim 120\,km\,s^{-1}$ (Dekel & Silk 1986; Dekel & Birnboim 2006, fig. 2), for which the accretion is almost all gaseous at $z = 0$, it is about 90 per cent at $z \simeq 0.7$, and about 67 per cent at $z \simeq 2.4$. This hints that supernova feedback could be the suppression mechanism allowing the accretion on to $\sim M_{sh}$ haloes to be predominantly gaseous at all relevant redshifts.

The dramatic variations in the disc accretion rate imply that each galaxy is likely to evolve through several subsequent phases of very high and very low SFR. This means that the expected overall trend of evolution from the blue cloud to the red sequence of galaxies is not a one-way track. In particular, some of the red and dead galaxies observed at high redshift may actually become blue again at later redshifts. The lesson is that the interpretation of the observed

evolution of the blue and red luminosity functions may involve non-trivial bookkeeping.

The late SAMBAs offer a very efficient post-burst shutdown that lasts for 6–7 Gyr. This can naturally give rise to red and dead galaxies in today's small-group haloes of $\leqslant 10^{13}$ M$_\odot$, perhaps S0 or elliptical galaxies (as simulated by Naab et al. 2007; Libeskind & Dekel, in preparation).

In the early SAMBAs, the accretion rate into the disc is suppressed compared to the virial rate even during the burst phase, but the post-burst accretion rate is not negligible. For these central galaxies in today's cluster haloes of $>10^{13}$ M$_\odot$ to end up as large red and dead ellipticals one needs a quenching mechanism beyond the one provided by spherical accretion. We show elsewhere (Dekel & Birnboim, 2007) that this could be provided by clumpy accretion, in which gas clumps of $10^7 - 10^9$ M$_\odot$ transfer their gravitational energy via drag and local shocks into heating and puffing up of the inner-halo gas of $\geqslant 10^{13}$ M$_\odot$ haloes.

Alternatively, a fashionable scenario suggests that the required quenching is provided by AGN feedback (Cattaneo et al. 2006; Croton et al. 2006). It has been noticed that the shock heating of the gas in haloes $>10^{12}$ M$_\odot$ enables the required coupling of the AGN energy with the halo gas (Dekel & Birnboim 2006), but we now realize that the SAMBA burst may be the actual trigger for AGN activity just above this critical mass. The SAMBAs may thus be responsible for both, the energy source and the coupling mechanism.

If the SAMBAs are responsible for triggering AGN activity, recall that they do so only in galaxies of $\sim 10^{11}$ M$_\odot$, a couple of Gyr after their haloes became larger than $\sim 10^{12}$ M$_\odot$. This may explain the otherwise surprising result that the bright quasars (near L_* of their Schechter luminosity function) seem to reside at all redshifts in haloes of $\sim 2 \times 10^{12}$ M$_\odot$ (Croom et al. 2005). A straightforward calculation using the Press–Schechter formalism shows that the evolution of number density of quasars at $z > 1$ can be reproduced by assuming quasar onset in every halo when it first becomes $\sim 2 \times 10^{12}$ M$_\odot$ and allowing a fixed short lifetime for each quasar.

The SAMBA events offer a complementary alternative to wet major mergers in producing rapid, massive bursts, with the important difference that it can, in principle, leave the disc intact. The survival of a disc configuration under a SAMBA collapse is yet to be investigated in detail. The associated shutdown leads to red and dead galaxies in haloes above the threshold mass. However, while major mergers also automatically produce spheroids, what could be the mechanism responsible for spheroid formation in association with the old, gas-poor stellar population according the halo shutdown scenario? First, the SAMBA burst provides the gas necessary for bar instability in the disc, which then transfers angular momentum to the outer disc and halo and produces a bulge. Second, the SAMBA quenching makes any subsequent mergers dryer, so they could lead to the large, boxy, non-rotating ellipticals, which cannot be reproduced by wet major mergers (Naab, Khochfar & Burkert 2006b; Novak et al., in preparation). Third, the minor mergers associated with the continuous smooth accretion of SAMBA could themselves lead to the formation of a spheroid (Naab et al. 2007; Bournaud et al., in preparation; Libeskind & Dekel, in preparation).

We argued in Dekel & Birnboim (2006) that discs could be built and form stars by cold streams, which could persist as narrow filaments even within the shock-heated medium slightly above $M_{\rm sh}$, especially at high redshifts. Streaming at the virial velocity, these flows could indeed give rise to rapid starbursts, on a time-scale comparable to the virial crossing time, but it is not clear that these flows could bring in mass in a rate that is much more efficient than the cosmological accretion rate. The SAMBAs provide such more efficient bursts in a natural way.

All the above is yet to be confirmed in cosmological simulations. It would be interesting to learn how the SAMBA phenomenon manifests itself when the spherical accretion of minor mergers is accompanied by filamentary streams and major mergers, and when star formation and feedback processes are properly simulated.

REFERENCES

Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Binney J., 1977, ApJ, 215, 483
Binney J., 2004, MNRAS, 347, 1093
Birnboim Y., Dekel A., 2003, MNRAS, 345, 349
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nat, 311, 517
Bullock J. S., Dekel A., Kolatt T. S., Kravtsov A. V., Klypin A. A., Porciani C., Primack J. R., 2001a, ApJ, 555, 240
Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001b, MNRAS, 321, 559
Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, MNRAS, 370, 1651
Ciotti L., Ostriker J. P., 2001, ApJ, 551, 131
Ciotti L., Ostriker J. P., 2007, Ap J, submitted (astro-ph/0703057)
Cox T. J., 2004, PhD thesis, UC Santa Cruz
Cox T. J., Dutta S. N., Di Matteo T., Hernquist L., Hopkins P. F., Robertson B., Springel V., 2006, ApJ, 650, 791
Croom S. M. et al., 2005, MNRAS, 356, 415
Croton D. J. et al., 2006, MNRAS, 365, 11
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Dekel A., Birnboim Y., 2007, preprint (arXiv: 0707.1214)
Dekel A., Silk J., 1986, ApJ, 303, 39
Forcada-Miro M. I., White S. D. M., 1997, preprint (astro-ph/9712204)
Förster Schreiber N. M. et al., 2006, ApJ, 645, 1062
Genzel R. et al., 2006, Nat, 442, 786
Hammer F., Flores H., Elbaz D., Zheng X. Z., Liang Y. C., Cesarsky C., 2005, A&A, 430, 115
Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2
Kravtsov A. V., 2003, ApJ, 590, L1
Kravtsov A. V., Gnedin O. Y., 2005, ApJ, 623, 650
Kriek M. et al., 2006, ApJ, 649, L71
Lacey C., Cole S., 1993, MNRAS, 262, 627
Lotz J. M. et al., 2006, preprint (astro-ph/0602088)
Mihos J. C., Hernquist L., 1996, ApJ, 464, 641
Mo H. J., White S. D. M., 2002, MNRAS, 336, 112
Naab T., Jesseit R., Burkert A., 2006a, MNRAS, 372, 839
Naab T., Khochfar S., Burkert A., 2006b, ApJ, 636, L81
Naab T., Johansson P. H., Ostriker J. P., Efstathiou G., 2007, Ap J, 658, 710
Neistein E., van den Bosch F. C., Dekel A., 2006, MNRAS, 372, 933
Noeske K. G. et al., 2007, ApJ, 660, L43
Rees M. J., Ostriker J. P., 1977, MNRAS, 179, 541
Silk J., 1977, ApJ, 211, 638
Spergel D. N. et al., 2007, ApJS, 170, 377
Springel V., Hernquist L., 2005, ApJ, 622, L9
Sugiyama N., 1995, ApJS, 100, 281
Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253
Thomas D., Maraston C., Bender R., de Oliveira C. M., 2005, ApJ, 621, 673
Thoul A. A., Weinberg D. H., 1995, ApJ, 442, 480
van den Bosch F. C., 2002, MNRAS, 331, 98

Wechsler R. H., Bullock J. S., Primack J. R., Kravtsov A. V., Dekel A., 2002, ApJ, 568, 52

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

APPENDIX A: MAIN PROGENITOR HISTORY

Following Neistein et al. (2006), we use the EPS formalism (Lacey & Cole 1993) to derive an analytic estimate for the main progenitor history. We track time backwards via the variable $\omega = \delta_c/D(t)$, where $D(t)$ is the cosmological growth factor of linear density fluctuations normalized to $D(t_0) = 1$ (see below), and $\delta_c \simeq 1.68$ (ignoring small variations due to the actual cosmology used). Mass enters via the monotonically decreasing function $\sigma(M)$, the standard deviation of the initial density fluctuations, as derived from their power spectrum, smoothed on a scale that corresponds to mass M and linearly extrapolated to t_0 (see below). According to EPS, a halo of mass M_0 at ω_0 has the following average number of progenitors of mass in the range $(M, M + dM)$ at the earlier time $\omega_0 + \Delta\omega$:

$$dN = \frac{1}{\sqrt{2\pi}} \frac{\Delta\omega}{(\Delta\sigma^2)^{3/2}} \exp\left[-\frac{(\Delta\omega)^2}{2\Delta\sigma^2}\right] \left|\frac{d\sigma^2}{dM}\right| \frac{M_0}{M} dM, \quad (A1)$$

where $\Delta\omega = \omega - \omega_0$ and $\Delta\sigma^2 = \sigma^2(M) - \sigma^2(M_0)$.

A1 Accretion rate on to M at t: a small time-step

We wish to compute the main progenitor history, i.e. the conditional average $M(\omega|\omega_0, M_0)$. The main progenitor is guaranteed to also be the most massive progenitor when $\Delta\omega$ is small enough, and this allows us to compute $\dot{M}$ by evaluating $M(\omega_0 + \Delta\omega|\omega_0, M_0)$ for small $\Delta\omega$.

For $M \geq M_0/2$, the probability density $P(M, \omega | M_0, \omega_0)$ is the same as the total progenitor distribution dN/dM given by equation (A1). For $M < M_0/2$, however, the only condition is $P \leq dN/dM$, and it is not sufficient to constrain P. As a first approximation we assume that the main progenitor always has a mass $M \geq M_0/2$. Then, the average mass of the main progenitor is

$$M(\omega_0 + \Delta\omega|\omega_0, M_0) = \int_{M_0/2}^{M_0} \frac{dN}{dM} M \, dM$$
$$= M_0 \left[1 - \text{erf}\left(\frac{\Delta\omega}{\sqrt{2\sigma_q^2 - 2\sigma_0^2}}\right)\right], \quad (A2)$$

where $\sigma_q = \sigma(M_0/q)$ with $q = 2$, and $\sigma_0 = \sigma(M_0)$. The rate of change is then

$$\frac{dM}{d\omega} = \lim_{\Delta\omega \to 0} \frac{M(\omega_0 + \Delta\omega|\omega_0, M_0) - M_0}{\Delta\omega}$$
$$= -M_0 \lim_{\Delta\omega \to 0} \frac{1}{\Delta\omega} \text{erf}\left(\frac{\Delta\omega}{\sqrt{2\sigma_q^2 - 2\sigma_0^2}}\right). \quad (A3)$$

Using the limit $\text{erf}(x) \to 2x/\sqrt{\pi}$ when $x \to 0$ we get

$$\frac{1}{M} \frac{dM}{d\omega} = -\left(\frac{2/\pi}{\sigma_q^2 - \sigma^2}\right)^{1/2}. \quad (A4)$$

Thus, the mean specific accretion rate on to haloes of mass M at t is

$$\frac{\dot{M}}{M}(M, t) = -s(M)\,\dot{\omega}(t), \quad (A5)$$

with

$$s(M) \equiv \left[\frac{2/\pi}{\sigma^2(M/q) - \sigma^2(M)}\right]^{1/2}, \quad (A6)$$

$$\dot{\omega}(t) = -\frac{\delta_c}{D}\frac{\dot{D}}{D}. \quad (A7)$$

When $\Delta\omega$ is not negligibly small, the probability for $M < M_0/2$ is not negligible, as it grows in proportion to $\Delta\omega$. Neistein et al. (2006) showed that this effect is rather small and can be bounded by well-defined limits on q, though the exact value of q is not fully specified by the EPS formalism. They showed that equation (A5) is still a valid approximation, but with q slightly larger than 2, and limited to the range 2–2.3 for a flat ΛCDM cosmology with $0.1 < \Omega_m < 0.9$ and $10^8 < M < 10^{15}$ M$_\odot$. We adopt $q = 2.2$ for all practical purposes.

As a sanity check, note that the obtained specific accretion rate at a fixed mass is self-similar in time, consistent with the fact that the only relevant mass scale at any given time is the non-linear clustering mass of the Press–Schechter formalism, M_*, defined by $\sigma[M_*(t)] = \delta_c/D(t)$. This is because the time dependence in equation (A5) is via $\dot{\omega} = -\sigma[M_*(t)]\dot{D}/D$. Once $\dot{M}/M$ is divided by the universal growth rate $\dot{D}/D$, the time dependence enters only through $M_*(t)$. Any viable approximation for the cosmological accretion rate must obey a self-similarity of this sort.

We also note in equation (2) that at a fixed t, $\dot{M}/M \to \infty$ when $M \to \infty$, but recall that haloes of $M \gg M_*$ are rare, such that most of the accreted mass is not on to arbitrarily massive haloes.

A2 Mass growth history of M_0 at t_0

The mass growth history of the main progenitor of a halo of mass M_0 at t_0 is obtained by integrating equation (A4) over $d\omega$ and over dM in the two sides of the equation:

$$\omega - \omega_0 = F(M) - F(M_0), \quad F(M) \equiv \int_M^\infty \frac{dm}{m\,s(m)}. \quad (A8)$$

The desired average mass of the main progenitor at t can be extracted from the above equation:

$$M(t|M_0) = F^{-1}[\omega - \omega_0 + F(M_0)], \quad w = \frac{\delta_c}{D(t)}, \quad (A9)$$

where F^{-1} is the inverse function of F, to be evaluated either explicitly or via numerical interpolation.

A3 A power-law power spectrum

A simple, fully self-similar example is provided by a power-law power spectrum, $P_k \propto k^n$, for which

$$\sigma = \sigma_R M^{-\alpha}, \quad \alpha \equiv (n + 3)/6. \quad (A10)$$

Here σ_R is a normalization constant, and M is measured in units of M_R, the average mass originally contained in a top-hat sphere of comoving radius R (e.g. σ_8 for $R = 8\,h^{-1}$ Mpc). In this case, the average accretion rate is given by equation (A5) with

$$s(M) = \tau M^\alpha, \quad \tau \equiv \frac{1}{\sigma_R}\left(\frac{2/\pi}{q^{2\alpha} - 1}\right)^{1/2}. \quad (A11)$$

Then,

$$F(M) = \frac{1}{\alpha\tau}\left(M^{-\alpha} - M_0^{-\alpha}\right), \quad (A12)$$

and the explicit solution for the average main-progenitor mass is

$$M(t|M_0) = \left[M_0^{-\alpha} + \alpha\tau(\omega - \omega_0)\right]^{-1/\alpha} \quad (A13)$$

(with $\omega = \delta_c/D(t)$, M in units of M_R and $q \simeq 2.2$).

Equation (A5) with $s(M)$ from equation (A11) approximates the accretion rate for a general power spectrum once α is the local slope $\alpha(M) \equiv -\mathrm{d}\ln\sigma/\mathrm{d}\ln M$. For $n = -2.1$, appropriate for $M \sim 10^{12}\,\mathrm{M}_\odot$ in the ΛCDM cosmology, we have $\alpha \simeq 0.15$. This is indeed similar to the power index of the mass dependence from N-body simulations (Wechsler et al. 2002, fig. 12).

The time dependence can be approximated in standard ΛCDM by equation (A29). Using a crude power-law approximation for the growth of the Press–Schechter mass out to $z \sim 2$, $M_*(z)/M_{*0} \simeq (1+z)^{-5}$, we end up with a practical approximation for the accretion rate into haloes in the vicinity of $M \sim 10^{12}\,\mathrm{M}_\odot$,

$$\frac{\dot{M}}{M} \simeq 0.04\,\mathrm{Gyr}^{-1}(1+z)^{2.25}\,M_{12}^{0.15}, \tag{A14}$$

which is accurate to better than 10 per cent over the range $0 \leqslant z \leqslant 2$.

A4 ΛCDM power spectrum

For the standard ΛCDM power spectrum we adopt $P(k) = k\,T^2(k)$ with the transfer function (Bardeen et al. 1986)

$$T(k) = \frac{\ln(1 + 2.34\,\kappa)}{2.34\,\kappa}$$
$$\times [1 + 3.89\,\kappa + (16.1\,\kappa)^2 + (5.46\,\kappa)^3 + (6.71\,\kappa)^4]^{-1/4}. \tag{A15}$$

Here $\kappa \equiv k/\Gamma$, with the wave number k in units of $(h^{-1}\,\mathrm{Mpc})^{-1}$, and with the shape parameter Γ (Sugiyama 1995),

$$\Gamma = \Omega_\mathrm{m}h\,\exp[-\Omega_\mathrm{b}(1 + \sqrt{2h}/\Omega_\mathrm{m})]. \tag{A16}$$

This power spectrum can be used to derive $\sigma(M)$ and the corresponding $s(M)$ with $q = 2.2$ in equation (2). This can then be substituted in equation (A5) for an explicit estimate of the average accretion rate on to haloes of mass M at t. It can then be used in equation (A8) for computing $F(M)$, evaluating its inverse function numerically and using it in equation (4) to obtain the average history of the main progenitor, $M(t\,|\,M_0)$.

Alternatively, following van den Bosch (2002), the rms fluctuation corresponding the ΛCDM power spectrum is approximated (to better 1 per cent in the range $10^6 < M < 10^{16}\,\mathrm{M}_\odot$) by

$$\sigma(M) = \sigma_8\,\frac{\Psi\left[(M/M_\mathrm{c})^{1/3}\right]}{\Psi[32\,\Gamma]}, \tag{A17}$$

$$M_\mathrm{c} = 9.73 \times 10^{11}\,\mathrm{M}_\odot\,h_{0.7}^{-1}\,(\Omega_\mathrm{m}/0.3)\,(\Gamma/0.2)^{-3}, \tag{A18}$$

and with the fitting function

$$\Psi(x) = 64.09\,(1 + 1.074\,x^{0.3} - 1.581\,x^{0.4}$$
$$+ 0.954\,x^{0.5} - 0.185\,x^{0.6})^{-10}. \tag{A19}$$

Substituting this analytic approximation in equation (A8), numerically inverting $F(M)$, and using it in equation (4), Neistein et al. (2006) derived the following approximation for the average main progenitor history:

$$\frac{M(t\,|\,M_0)}{M_\mathrm{c}} = F_q^{-1}\left[\frac{\Psi(32\Gamma)}{\sigma_8}(\omega - \omega_0) + F_q\left(\frac{M_0}{M_\mathrm{c}}\right)\right], \tag{A20}$$

with $q = 2.2$ and the fitting function

$$F_{2.2}(u) = -6.92 \times 10^{-5}(\ln u)^4 + 5.0 \times 10^{-3}(\ln u)^3$$
$$+ 8.64 \times 10^{-2}(\ln u)^2 - 12.66\,(\ln u). \tag{A21}$$

This fitting function is accurate to better than 1 per cent over the range $10^6 < M < 10^{15}\,\mathrm{M}_\odot$ for the standard ΛCDM cosmology.

A5 In cosmological *N*-body simulations

Wechsler et al. (2002) analysed an N-body simulation of ΛCDM for the average mass-growth history of the main progenitor of a halo of mass M_0 at t_0. They found a good fit with

$$M(t\,|\,M_0) = M_0\,\mathrm{e}^{-2a_\mathrm{c}z}, \tag{A22}$$

$$a_\mathrm{c} = 0.28\left(\frac{M_0}{10^{12}\,\mathrm{M}_\odot}\right)^{0.15}. \tag{A23}$$

This function is a good practical approximation to the EPS estimate specified above, as long as t_0 is the present time, $z = 0$.

A6 Useful relations for linear fluctuations

Following Mo & White (2002), in a flat universe

$$D(z) = \frac{g(z)}{g(0)}\frac{1}{(1+z)}, \tag{A24}$$

$$g(z) \simeq \frac{5}{2}\left(\Omega_\mathrm{m}^{-3/7}(z) + \frac{3}{2}\right)^{-1}, \tag{A25}$$

with a negligible error on the order of $\Omega_\lambda/70$ in $g(z)$. Thus

$$\frac{\dot{D}}{D} \simeq H(z)\left\{1 + \frac{18}{35}\,g(z)\,\Omega_\mathrm{m}^{-3/7}(z)\,[\Omega_\mathrm{m}(z) - 1]\right\}. \tag{A26}$$

Here

$$H(z) = H_0 E(z), \quad \Omega_\mathrm{m}(z) = \frac{\Omega_{\mathrm{m}0}(1+z)^3}{E^2(z)}, \tag{A27}$$

$$E^2(z) = 1 - \Omega_{\mathrm{m}0} + \Omega_{\mathrm{m}0}(1+z)^3. \tag{A28}$$

For ΛCDM with $\Omega_\mathrm{m} = 0.3$ and $\Omega_\Lambda = 0.7$, a practical approximation is

$$\frac{\dot{D}}{D} \simeq 0.0385(1+z)^{3/2}\,h_{0.7}\,\mathrm{Gyr}^{-1}, \tag{A29}$$

which is accurate to better than 1 per cent at $z \geqslant 1$ and better than 10 per cent at $z \geqslant 0$.

The fluctuation variance is derived in the general case by integrating the power spectrum,

$$\sigma^2(M) = \frac{1}{2\pi}\int_0^\infty \mathrm{d}k\,k^2 P(k)\tilde{W}^2(kR), \tag{A30}$$

with $M = (4\pi/3)\,\rho_0\,R^3$, and with the Fourier transform of the real-space top-hat window function or radius R:

$$\tilde{W}(x) = 3(\sin x - x\cos x)/x^3. \tag{A31}$$

APPENDIX B: RECIPES FOR SAMS

B1 Following halo evolution in SAM

Given the mass growth of each halo in the merger tree, $M(t)$, the SAMBA onset time t_1 (redshift z_1) is identified by $M(t_1) = M_1 = 10^{12}\,\mathrm{M}_\odot$. The SAMBA is classified as 'early type' or 'late type' according to whether $z_1 > 2.5$ or $z_1 \leqslant 2.5$, respectively. Then the beginning and the end of the burst, as read from Fig. 5, are

$$t_2 = \begin{cases} 1.80\,t_1, & z_1 > 2.5 \\ 1.36\,t_1, & z_1 \leqslant 2.5, \end{cases} \quad t_3 = 1.18\,t_2. \tag{B1}$$

This defines the four phases of gas cooling.

The growth rates of cold and hot gas mass during each phase can be approximated as follows.

(i) *Cold-flow* phase, $t \leqslant t_1$,

$$\dot{M}_d = f_b \dot{M}, \qquad \dot{M}_{hot} = 0. \tag{B2}$$

(ii) *Tentative-quenching* phase, $t_1 < t \leqslant t_2$,

$$\dot{M}_d = f_b \frac{\Delta M_{12}}{\Delta t_{12}} \begin{cases} 0.08, & z_1 > 2.5, \\ 0.3, & z_1 \leqslant 2.5, \end{cases} \tag{B3}$$

$$\dot{M}_{hot} = f_b \dot{M} - \dot{M}_d, \tag{B4}$$

where $\Delta M_{ij} = M(t_j) - M(t_i)$ and $\Delta t_{ij} = t_j - t_i$. The numerical factors for early and late SAMBAs, both for the tentative quenching and for the burst below, are read from straightforward fits to $\Delta M_d / (\Delta M f_b)$ in Fig. 7.

(iii) *Burst* phase, $t_2 < t \leqslant t_3$,

$$\dot{M}_d = f_b \frac{\Delta M_{23}}{\Delta t_{23}} \begin{cases} 0.5, & z_1 > 2.5, \\ 1.6, & z_1 \leqslant 2.5, \end{cases} \tag{B5}$$

$$\dot{M}_{hot} = -\frac{M_{hot2}}{\Delta t_{23}}, \tag{B6}$$

where M_{hot2} is the accumulated hot gas mass by t_2, $M_{hot2} = f_b \Delta M_{12} - \dot{M}_{disc,2} \Delta t_{12}$. The cold accretion is more efficient in the late bursts because of the tentative shock instability due to the lower virial accretion rate at late times.

(iv) *Post-burst* phase, $t > t_3$,

$$\dot{M}_d = f_b \dot{M} \begin{cases} 0.12, & z_1 > 2.5, \\ 0, & z_1 \leqslant 2.5, \end{cases} \tag{B7}$$

$$\dot{M}_{hot} = f_b \dot{M} - \dot{M}_d. \tag{B8}$$

The numerical factor for early SAMBAs is read from the fit $M/M_d \simeq 4/f_b$ at to the results shown in Fig. 7, combined with the fact seen in Fig. 4, that during the post-burst phase $\dot{M}_d / M_d \simeq \dot{M}/M$ and $M/M_d \simeq$ const. Recall that the early SAMBAs, ending up in $M \geqslant 10^{13} M_\odot$ haloes today, should be quenched after t_3 by another mechanism, such as clumpy accretion or AGN feedback. If such a process is not explicitly included in the SAM, one can simply set $\dot{M}_d = 0$ in the post-burst phase for all haloes (as in Cattaneo et al. 2006).

The disc mass growth according to the proposed SAMBA recipe for SAM is shown in Fig. 10, in comparison with the standard disc growth following the evolution of a 'cooling radius'.

B2 SFR as a function of halo mass: FHM

An alternative, quick and crude way to learn about the implications of SAMBAs is by providing a recipe for the SFR as a function of halo mass and redshift, to be incorporated in a given halo population in a SAM while bypassing the detailed baryonic processes of cooling, star formation and feedback (FHM; Croton & Dekel, in preparation). The fact that the SAMBA phenomenon is driven by halo mass allows us to provide a concrete recipe of this sort.

The SAMBA characteristic halo masses M_1, M_2 and M_3, corresponding to the characteristic times t_1, t_2 and t_3, can be approximated based on our simulations by
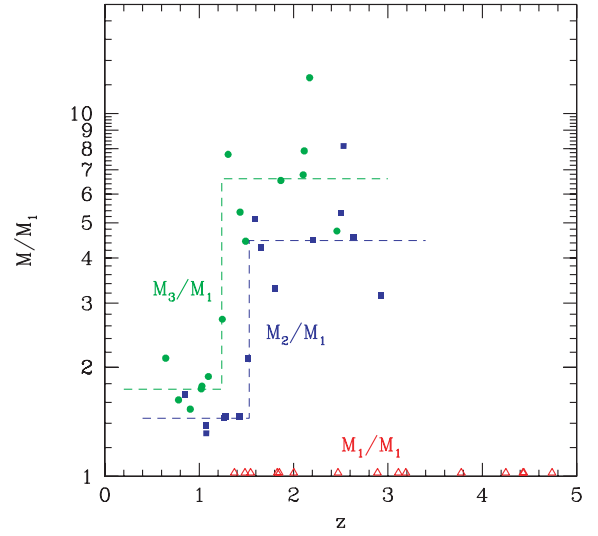
$$M_1 = 10^{12} M_\odot, \tag{B9}$$

Figure B1. The characteristic SAMBA masses relative to $M_1 \simeq 10^{12} M_\odot$. Open red triangles refer to z_1. Solid blue squares refer to M_2/M_1 at z_2. Solid green circles refer to M_3/M_1 at z_3. The lines mark the fits used for the FHM recipe.

$$M_2/M_1 = \begin{cases} 4.45, & z > 1.53, \\ 1.45, & z \leqslant 1.53, \end{cases} \tag{B10}$$

$$M_3/M_1 = \begin{cases} 6.61, & z > 1.24, \\ 1.74, & z \leqslant 1.24. \end{cases} \tag{B11}$$

The constant value of M_1 is as predicted by Dekel & Birnboim (2006) for a realistic gradual increase of metallicity with time. Then M_2 and M_3 are obtained from the fits shown in Fig. B1. Notice the interesting prediction that in the range $1.24 < z \leqslant 1.53 M_3$ is much larger than M_2, giving rise to a broad range of massive haloes with high SFR.

Given the virial accretion rate in the provided merger tree, $\dot{M}$, the disc accretion rate, which could be interpreted as an upper limit for the SFR, is approximated in the different mass ranges by the following.

(i) *Cold-flow* range, $M \leqslant M_1$,

$$\dot{M}_d = f_b \dot{M}. \tag{B12}$$

(ii) *Tentative-quenching* range, $M_1 < M \leqslant M_2$,

$$\dot{M}_d = f_b \dot{M} \begin{cases} 0.08, & z > 1.8, \\ 0.3, & z \leqslant 1.8. \end{cases} \tag{B13}$$

(iii) *Burst* range, $M_2 < M \leqslant M_3$,

$$\dot{M}_d = f_b \dot{M} \begin{cases} 0.5, & z > 1.4, \\ 1.6, & z \leqslant 1.4. \end{cases} \tag{B14}$$

The numerical values are read from crude fits to the simulation results for $\Delta M_d / (\Delta M f_b)$, as shown in Fig. 7.

(iv) *Post-burst* range, $M > M_3$,

$$\dot{M}_d = f_b \dot{M} \begin{cases} 0.12, & M > M_q(z), \\ 0, & M \leqslant M_q(z), \end{cases} \tag{B15}$$

$$M_q(z) \simeq 10^{13} M_\odot \exp(-0.75 z). \tag{B16}$$

The upper-limit for effective quenching, $M_q(z)$, is the mass of the intermediate SAMBA separating late from early SAMBAs, which eventually ends up as $M = 10^{13}\,M_\odot$ today. The above estimate is based on the fit to simulations, equation (A22). Note that once $z > 1.4$, $M_3 > M_q(z)$, so in this redshift range $M > M_3$ automatically implies $M > M_q$.

The numerical values in the quenching and burst phases are read from the crude fits to the simulation results shown in Fig. 7, in most of which the virial mass grows according to the average cosmological rate. When implementing the above recipe, one can either use the actual mass growth rate of the haloes drawn from merger-tree realizations, or the average virial accretion rate as provided in equation (3).

The main feature of the model is the quenching of SFR above a threshold halo mass, $M_{sh} \sim 10^{12}\,M_\odot$, as already implemented in SAM simulations (Cattaneo et al. 2006; Croton et al. 2006). The interesting new feature predicted by the model is the high SFR in a second range of masses above M_{sh}. At low redshifts, this is a slight effective increase of the threshold mass to $\sim 2 \times 10^{12}\,M_\odot$. At $z \gtrsim 1.3$, the high SFR is predicted to extend to haloes as large as $\sim 7 \times 10^{12}\,M_\odot$. This should allow the appearance of massive blue galaxies at high redshifts. In the special range $1.24 < z < 1.53$, the high SFR is expected over an especially broad range of masses. Given that $10^{12}\,M_\odot$ is the typical mass for haloes that form at this redshift range, this should lead to a peak in the history of cosmological SFR density (the 'Madau' plot).

B3 Fine-tuning of the SAMBA recipes

While the above recipes are our best guess based on the spherical model, with no free parameters, a better match to observations may require modest fine-tuning of model parameters, e.g. allowing for certain uncertainties in the assumptions made in the computations, as well as compensating for deviations from spherical symmetry and uniform accretion.

The uncertainty in the value of the critical mass for shock heating, M_{sh}, can easily accommodate a factor of 2 and some redshift dependence. As computed in Dekel & Birnboim (2006), it is a function of certain physical parameters, in particular redshift z and the average metallicity Z and gas fraction f_b in the inner halo, which also vary with z. The estimates of M_{sh} as a function of z are shown in their fig. 2. A practical approximation based on their equation (34) is

$$M_{sh} \simeq 10^{12}\,M_\odot \,\frac{Z_{0.2}^{0.52}\,f_{b\,0.05}^{3/4}}{(1+z)^{3/8}}\left[\frac{\ln(1+C_0)}{\ln(1+C)}\right]^{3/4}, \tag{B17}$$

where $Z_{0.2} \equiv Z/(0.2\,Z_\odot)$ and $f_{b\,0.05} \equiv f_b/0.05$. In this formula, following Dekel & Birnboim (2006), the variation of mean metallicity with redshift can be crudely approximated by

$$\log\frac{Z}{Z_0} = -0.17\,z. \tag{B18}$$

Following Bullock et al. (2001b), the average halo concentration near $M \sim 10^{12}\,M_\odot$ is evolving as

$$C = C_0\,(1+z)^{-1}, \quad C_0 \simeq 13. \tag{B19}$$

With this choice, the critical mass is $M_{sh} \sim 10^{12}\,M_\odot$ and is only a weak function of redshift. It becomes even weaker if f_b gets slightly higher at high z. One can see that a factor of 4 change in the unknown Z_0 results in a factor of ~ 2 in M_{sh}.

Another uncertainty, which is explicit in the FHM recipe, is in the actual SFR that is associated with the maximum possible rate provided by the disc accretion rate. This can be modelled as a free ratio, smaller than unity, between the two.

This paper has been typeset from a TeX/LaTeX file prepared by the author.