2.1. Gas Accretion

We begin by considering how gas flows inward through the CGM and accretes onto central galaxies. This includes the physics and properties of different modes of gas accretion (cold versus hot), some of their expected observational signatures, and possible effects on the evolution of galaxies.

Cold flows/cold streams: cold halo gas, typically at a temperature $T \sim 10^4$ K and filamentary in morphology; associated with "cold mode" accretion

Cooling flow: model for hot gas accretion in halos driven by radiative cooling and neglecting feedback; associated with "hot mode" accretion

2.1.1. Cold versus hot accretion. Before considering the full complexity of the CGM, it is useful to examine the cooling physics of gas in dark matter halos in the idealized approximation of spherical symmetry, neglecting feedback processes. Three timescales are important: the Hubble time $t_{\rm H} = 1/H$ (where H is the redshift-dependent Hubble parameter), the free-fall time in the gravitational potential $t_{\rm ff}$, and the cooling time of the gas $t_{\rm cool}$. As dark matter halos form from gravitational clustering, gas is dragged inward. In the lowest-mass halos the gas inflows remain subsonic owing to heating by photoionization by the cosmic ionizing background; in those small halos galaxy formation is suppressed (e.g., Efstathiou 1992, Noh & McQuinn 2014). In more massive halos, inflows reach supersonic velocities and are shock-heated to a temperature on the order of the virial temperature, $T_{\rm vir} = (\mu m_{\rm p}/2k)(GM_{\rm h}/R_{\rm vir})$, where μ is the mean molecular weight ($\approx$0.6 for an ionized cosmic plasma), and $m_{\rm p}$ is the proton mass. For a halo of mass $M_{\rm h} = 10^{12}$ $M_\odot$ at $z = 0$ (similar to the Milky Way), the virial radius $R_{\rm vir} \approx 260$ kpc, and the virial temperature $T_{\rm vir} \approx 6 \times 10^5$ K (Barkana & Loeb 2001). Because $R_{\rm vir} \propto M_{\rm h}^{1/3}$, the virial temperature $T_{\rm vir} \propto M_{\rm h}^{2/3}$; for $\gtrsim L_\star$ halos this gas emits in X-rays. The character of gas accretion onto the central galaxy (and of the CGM) depends on whether the cooling of the shocked gas is rapid or slow relative to the free-fall time.

2.1.1.1. Cold accretion. When $t_{\rm cool} < t_{\rm ff}$, the shocked gas rapidly cools and loses its thermal pressure support. The cold $T \sim 10^4$ K gas that results tends to fragment and clump, and can also form narrow filaments known as cold flows or cold streams (see Section 2.1.4 on cold streams and more in Section 3 about the small-scale properties of cold gas). If unimpeded, e.g., by feedback or angular momentum (AM), the cold gas can accrete onto the central galaxy in a free-fall time. Because the infall of the cold gas is highly supersonic (relative its internal sound speed), a strong shock can form on impact with the central galaxy.

2.1.1.2. Hot accretion. When $t_{\rm cool} > t_{\rm ff}$, gas cooling becomes a rate-limiting step. Shock-heated gas can be supported for an extended period of time $\sim t_{\rm cool}$ in the halo potential by thermal pressure. In the inner regions, within the cooling radius, where $t_{\rm cool} < t_{\rm H}$, there is sufficient time for the hot gas to cool and accrete smoothly onto the central galaxy. Absent feedback, these cooling regions tend to a steady-state cooling flow in which compressional heating in the inflowing gas balances radiative losses (e.g., Fabian et al. 1984), though in practice feedback processes can modify the flow.

The different limits corresponding to different regimes of $t_{\rm cool}/t_{\rm ff}$ are core ingredients of theories of galaxy formation, starting from influential analytic models from the 1970s (Binney 1977, Rees & Ostriker 1977, Silk 1977, White & Rees 1978). The implications of these limits for galaxy formation as well as the CGM have been the subject of extensive investigation ever since, using analytic and semianalytic techniques (e.g., White & Frenk 1991, Somerville & Primack 1999, Dekel & Birnboim 2006), idealized numerical simulations (e.g., Birnboim & Dekel 2003, Fielding et al. 2017b, Stern et al. 2020), and detailed cosmological simulations (e.g., Kereš et al. 2005, 2009b; Faucher-Giguère et al. 2011; van de Voort et al. 2011; Nelson et al. 2013). Some ideas are summarized in Section 2.1.6 though this is still an active area of research and (perhaps surprisingly) there is not yet agreement on the effects of cold versus hot accretion for galaxy formation and evolution.

In the above sketch, we have deliberately been ambiguous about where the cooling and free-fall times are evaluated. Modern hydrodynamic simulations as well as observations indicate that the CGM can be highly inhomogeneous and consist of multiple phases. Therefore, different $t_{\rm cool}/t_{\rm ff}$ limits can be realized in different regions. The physical picture is further complicated by outflows from stars and black holes (Section 2.3), as well as additional physics such as magnetic fields, thermal conduction, and CRs (Section 3), which imply there is in general much more to the CGM than just cooling and gravity.

2.1.2. Maximum hot gas accretion. To gain further insight into the different modes of gas accretion in halos, we consider some analytic results regarding the maximum rate of hot gas accretion. Our treatment here follows Stern et al. (2019, 2020), who analyzed the physics of cooling flows in galaxy-scale halos. Although real halos can be much more complex and dynamic than idealized cooling flows, this simplified setup allows us to develop analytic insights that apply in regions where the gas dynamics is dominated by gravity and cooling. These results build on and extend previous work on cooling flows in clusters of galaxies (Mathews & Bregman 1978, Fabian et al. 1984). In clusters it is well known that cooling flow models fail to explain the X-ray properties of the ICM. The jury is still out as to whether pure cooling flow models can adequately model the CGM of some lower-mass systems, because X-ray observations can currently only barely probe the hot gas in such halos.[1] We do not take a position on this here but simply use cooling flows as a useful baseline solution to gain insight into expected CGM properties before they are modified by feedback.

The setup is a spherically symmetric dark matter halo in which there is initially a pressure-supported, steady flow of gas near the virial temperature. The energy conservation equation is $v_r {\rm d}[v_r^2/2 + \gamma\epsilon + \Phi]/{\rm d}r = -q$, where r is the radius, v_r is the radial velocity, ϵ is the specific thermal energy, γ is the adiabatic index of the gas, Φ is the gravitational potential, and q is the cooling rate per unit mass. The sum in square brackets is the Bernoulli parameter, which is conserved along stream lines in a steady flow. To first approximation, the first two terms can be neglected for slow inflow and for potentials that are not too far from isothermal (such that the specific thermal energy gradient is small), so that ${\rm d}\Phi/{\rm d}r \approx -q/v_r = -n_{\rm H}^2\Lambda/\rho v_r$, where Λ is the cooling function. Because the mass accretion rate $\dot{M} = -4\pi r^2 \rho v_r$ (the accretion rate is positive when the radial velocity is negative), we have the following expression in terms of the gas cooling rate and the radial gradient of the potential:

$$\dot{M} \approx \frac{4\pi r^2 n_{\rm H}^2 \Lambda}{{\rm d}\Phi/{\rm d}r}. \qquad 1.$$

At any radius in the halo, there is a maximum steady accretion rate of hot gas, which is set by the requirement that the density must be low enough that $t_{\rm cool} \gtrsim t_{\rm ff}$. At higher densities, the rate of compressional heating in the cooling flow cannot balance the radiative cooling rate: The gas rapidly cools to $\ll T_{\rm vir}$. The maximum density can be evaluated using $t_{\rm ff} = \sqrt{2}r/v_{\rm c}$ (where $v_{\rm c}$ is the circular velocity in the potential) and $t_{\rm cool} = \epsilon/q = \rho\epsilon/n_{\rm H}^2\Lambda$:

$$n_{\rm H,max}(r) \approx \frac{m_{\rm p} v_{\rm c} \epsilon}{X\Lambda r} \approx \frac{m_{\rm p} v_{\rm c}^3}{X\Lambda r} \approx 0.007 \text{ cm}^{-3}\ v_{100}^3 r_{10}^{-1} \Lambda_{-22}^{-1}, \qquad 2.$$

[1]We note this is plausible because e.g., stellar feedback can in principle act very differently on galaxy scales than AGN feedback acts on cluster scales. Furthermore, outflows appear to be relatively weak around low-redshift $\sim L_\star$ galaxies such as the Milky Way, so their hot CGM may be reasonably well approximated by pure cooling physics (Stern et al. 2019).

Virialized gas: gas with thermal velocities set by the gravitational potential; in massive halos, this corresponds to a hot phase ($T_{\rm vir} \gtrsim 10^6$ K)

where $X = 0.75$ is the hydrogen mass fraction, $v_{100} = v_{\rm c}/(100\ {\rm km\ s^{-1}})$, $r_{10} = r/(10\ {\rm kpc})$, and $\Lambda_{-22} = \Lambda/(10^{-22}\ {\rm erg\ cm^3\ s^{-1}})$. The second equality follows from $\epsilon \approx v_{\rm c}^2$, which is equivalent to the statement that the gas is at the virial temperature. Using $d\Phi/dr = v_{\rm c}^2/r$, the maximum density corresponds to a maximum hot gas accretion rate,

$$\dot{M}_{\rm max}(r) \approx \frac{4\pi m_{\rm p}^2 v_{\rm c}^4 r}{X^2 \Lambda} \approx 3\ {\rm M_\odot\ year^{-1}}\ v_{100}^4 r_{10} \Lambda_{-22}^{-1}. \quad 3.$$

2.1.3. Virialization of the inner CGM and the threshold halo mass. Equations 2 and 3 imply that the maximum hot gas density and accretion rate depend on radius. For gas in cooling flows, the ratio of the cooling time to the free-fall time increases from the inside out (e.g., Stern et al. 2020)[2]. Therefore, the outer parts of the CGM can be hot and contain virialized gas ($t_{\rm cool}/t_{\rm ff} > 1$),[3] whereas the inner parts cool rapidly and tend toward free fall ($t_{\rm cool}/t_{\rm ff} < 1$). The fact that the inner CGM virializes last is important because this defines the time at which the boundary conditions of the central galaxy change.

Cooling flow solutions also reveal an important connection between cooling and whether the flow is subsonic or supersonic. In the hot part of the cooling flow, where the temperature $kT_{\rm vir} \sim m_{\rm p} v_{\rm c}^2$, the sound speed $c_{\rm s} \sim \sqrt{P/\rho} \sim v_{\rm c}$. Thus, the free-fall time $t_{\rm ff} \sim r/c_{\rm s}$ is on the order of a sound crossing time. In this region, the inflow rate is limited by cooling, so we have $t_{\rm cool} \sim r/|v_r|$. Combining these results and defining the Mach number $\mathcal{M} = |v_r|/c_{\rm s}$,

$$\frac{t_{\rm cool}}{t_{\rm ff}} \approx \mathcal{M}^{-1}. \quad 4.$$

In this expression, we have omitted a prefactor $\approx$1 whose exact value depends on the shape of the gravitational potential. It follows from Equation 4 that the radius where $t_{\rm cool}/t_{\rm ff} \approx 1$ coincides with the sonic radius $R_{\rm sonic}$, where $\mathcal{M} = 1$. The flow is subsonic outside $R_{\rm sonic}$ but supersonic inside that radius. The transition from a subsonic to a supersonic flow has important implications for both the physics and observational properties of the CGM. In particular, thermal instability is inhibited in the subsonic region of a standard cooling flow, whereas it can grow faster than the flow time in the supersonic region (Balbus & Soker 1989). In the supersonic region, large density and pressure fluctuations develop as a result of thermal instability (e.g., Stern et al. 2020), which may have important implications for observational signatures as well as how galaxies interact with the CGM via inflows and outflows (see Section 2.1.6). We discuss thermal instability further in Section 3.1, where we note that if feedback keeps the hot gas close to global hydrostatic and thermal equilibrium, local thermal instability can develop if $t_{\rm cool}/t_{\rm ff} \lesssim 10$.

We now address the question of which halos are expected to be virialized. Because a given halo can be virialized outside its sonic radius but not inside it, the question is made more precise by considering the point at which the CGM becomes entirely virialized, i.e., when the sonic radius becomes equal to the radius of the central galaxy. As stressed above, whether the CGM can sustain $t_{\rm cool}/t_{\rm ff} > 1$ depends on the complexities of the baryon cycle (see **Figure 2**) and feedback in particular, as it affects $t_{\rm cool}$ by heating up the gas and by changing its density and metallicity. However, a critical halo mass can be derived based on simplified assumptions.

[2] In other models for the structure of hot gas in halos, the $t_{\rm cool}/t_{\rm ff}$ ratio can be constant or decrease with radius (e.g., Sharma et al. 2012a, Voit et al. 2017, Faerman et al. 2020). However, these models assume that heating from galactic feedback balances cooling in the CGM.

[3] By virialized, we mean that a virial-temperature phase is long-lived. Such gas can be sustained for longer than either a cooling time or a free-fall time if there is a continuous supply of gas, e.g., through accretion from the IGM, because in the cooling flow that develops (if feedback is neglected), compressional heating in the accreting gas balances radiative cooling.

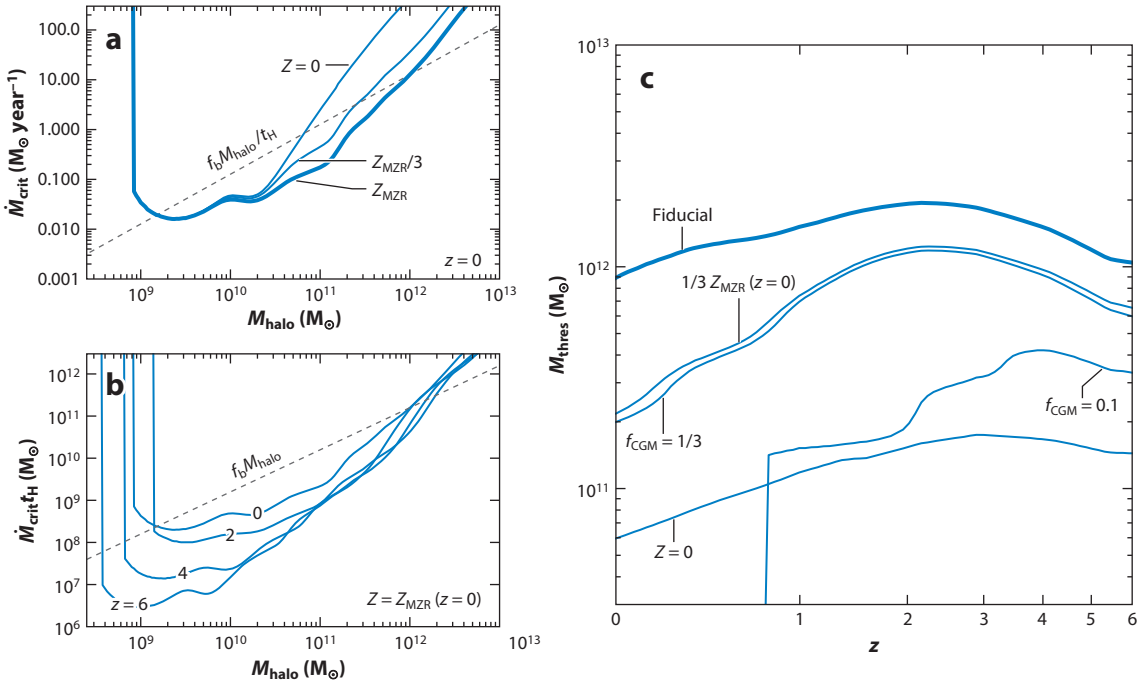

Figure 3

Critical gas accretion rate for hot gas and the threshold halo mass, indicating when halos complete virialization. (*a*) The critical accretion rate as a function of halo mass at $z = 0$ for different assumed metallicities Z for the gas (Z_{MZR} is the metallicity implied by the observed relationship between galaxy mass and ISM metallicity). (*b*) Similar to panel *a* but with the critical accretion rate converted into a total gas mass by multiplying by the Hubble time. The curves all assume that the gas metallicity is consistent with the $z = 0$ mass–metallicity relation but show the results for different redshifts. In this figure, $f_b = \Omega_b/\Omega_m$ is the cosmic baryon fraction, and f_{CGM} is the fraction of the halo baryonic mass that is in CGM gas. (*c*) The threshold halo mass above which the CGM is expected to be completely hot, M_{thres}, as a function of redshift for different assumptions. This corresponds to when the gas mass in the halo is equal to $\dot{M}_{crit} t_H$. The fiducial case corresponds to a baryon-complete CGM ($f_b = f_{CGM} = 1$) on the mass–metallicity relation, and the other curves show how the threshold mass is modified when either the gas mass in the CGM or its metallicity are reduced. The threshold halo mass also depends on the spin of the gas (via the circularization radius), but this dependence is not shown here for simplicity. Figure adapted with permission from Stern et al. (2020). Abbreviations: CGM, circumgalactic medium; ISM, interstellar medium.

The CGM can be considered to complete virialization when the accretion rate is below $\dot{M}_{max}$ all the way to the circularization radius $R_{circ} \approx \sqrt{2}\lambda R_{vir} \approx 0.05 R_{vir}$, where inflowing gas becomes supported by AM, which we use as a proxy for the inner boundary of the CGM (the spin parameter λ is defined and discussed further in Section 2.2). For any given halo, this defines a critical gas accretion rate $\dot{M}_{crit}$ equal to $\dot{M}_{max}(R_{circ})$. Approximating the cooling function as $\Lambda \propto T^{-0.7} Z^{0.9}$, valid for $T \sim 10^5$–10^7 K and metallicities $Z \gtrsim 0.3 Z_\odot$ (Wiersma et al. 2009), Stern et al. (2020) obtained

$$\dot{M}_{crit} \approx 0.7\ M_\odot\ \text{year}^{-1}\ v_{100}^{5.4} R_{10} Z_{0.3}^{-0.9}, \qquad 5.$$

where $v_{100} = v_c/(100\,\text{km s}^{-1})$ is the circular velocity, $R_{10} = R_{circ}/(10\,\text{kpc})$, and $Z_{0.3} = Z/(0.3\,Z_\odot)$. The circular velocity and gas metallicity are evaluated at R_{circ}. The value of $\dot{M}_{crit}$ as a function of halo mass is plotted in **Figure 3*a*** for different metallicities at $z = 0$. **Figure 3*b*** shows $\dot{M}_{crit} t_H$ as a function of halo mass for different redshifts, assuming a mass-dependent metallicity consistent with the observed mass–metallicity relation for galaxies.

The critical accretion rate can be translated into a threshold halo mass M_{thres} by setting the total gas mass in the halo $M_{gas} = f_{CGM} f_b M_h$ (where $f_b = \Omega_b/\Omega_m \approx 0.16$ is the cosmic baryon budget and

$f_{\rm CGM}$ is the fraction of this budget in CGM gas) to $\dot{M}_{\rm crit} t_{\rm H}$. The idea is that $\dot{M}_{\rm crit} t_{\rm H}$ is an estimate of the hot gas mass in the halo when virialization completes, so the CGM will be fully virialized only for $M_{\rm h} \geq M_{\rm thres}$. **Figure 3*c*** shows $M_{\rm thres}$ as a function of redshift. The solid curve in this panel shows the result for a baryon-complete CGM ($f_{\rm CGM} = 1$) and other fiducial assumptions. Interestingly, $M_{\rm thres}$ is roughly independent of redshift, staying in the range $\approx$(1–2) $\times$ 10^{12} $M_\odot$ from $z = 0$ to $z = 6$. To see why, note that at fixed metallicity, Equation 5 implies $\dot{M}_{\rm crit} \propto v_{\rm c}^{5.4} R_{\rm circ}$. For a matter-dominated Universe, at fixed $M_{\rm h}$, $R_{\rm circ} \propto R_{\rm vir} \propto 1/(1+z)$, $v_{\rm c} \sim v_{\rm vir} = \sqrt{GM_{\rm h}/R_{\rm vir}} \propto (1+z)^{1/2}$ ($v_{\rm vir}$ is the virial velocity), and $t_{\rm H} \propto (1+z)^{-3/2}$. Therefore, $\dot{M}_{\rm crit} t_{\rm H} \propto (1+z)^{0.2}$, which depends weakly on redshift.[4]

We note that halos can be virialized substantially below the fiducial threshold mass $M_{\rm h} \approx 10^{12}$ $M_\odot$, for example, if the gas metallicity is lower than assumed or if the CGM density is below that implied by the cosmic baryon budget. Strong stellar feedback may indeed deplete the CGM by large factors in low-mass halos (e.g., Hafen et al. 2019). In these limits, the entire CGM can potentially be virialized in halos of mass as low as $\sim 10^{11}$ $M_\odot$, or even less.

The threshold mass derived above based on cooling-flow arguments is similar to the threshold mass previously derived based on the stability of virial shocks (Birnboim & Dekel 2003, Dekel & Birnboim 2006). In these derivations, one considers cool gas accreting supersonically into halos and shocking as the central galaxy is approached in the inner regions. The shock is considered stable when the cooling time of the shocked gas is sufficiently long for its thermal pressure to drive outward expansion of the accretion shock. The threshold halo mass derived in this way roughly matches the one derived based on cooling flows because both follow from a comparison of similar cooling and flow timescales. The cooling-flow derivation has the advantage of highlighting the fact that the inner parts of the CGM stay hot and virialized last, which is opposite to the inside-out direction in which accretion shocks propagate. The key reason for this difference is that, once hot gas is created, whether it stays hot or rapidly cools in a given region of the CGM is a function of the local $t_{\rm cool}/t_{\rm ff}$ ratio, regardless of the directionality of the shock that originally heated the gas. On average this ratio increases from the inside out.

We stress that the cooling flow and virial shock stability treatments are two idealized models for gas virialization in halos. The two models provide complementary insights, but we do not expect either to perfectly describe the dynamics of the real CGM, which are more complex due to time-variable inflows and outflows as well as strong departures from spherical symmetry. In cosmological simulations including realistic feedback, such as the `FIRE` zoom-in simulation of a Milky Way–mass halo shown in **Figure 4**, it is found that the CGM is often first heated out to large radii by shocks due to star formation–driven galactic winds before the theory predicts that pure accretion-driven shocks should be stable.[5] As a result, the outer parts of low-mass halos can be hot well before cooling times in the inner CGM become long enough to sustain a virialized CGM throughout the halo. Although this fact has seldom been emphasized in the literature so far, other simulations also find that the outer CGM is typically heated to $\sim T_{\rm vir}$ before the inner CGM is able to virialize [for example, this is apparent in temperature profiles of halos from the `EAGLE` simulations analyzed by Correa et al. (2018) and Wijers et al. (2020)].

[4]**Figure 3*b*** shows that $\dot{M}_{\rm crit} t_{\rm H}$ depends more strongly on redshift for low-mass halos. This is because the weak redshift scaling depends on the $\Lambda \propto T^{-0.7}$ temperature scaling, which is only a valid approximation for $T \sim 10^5$–10^7 K, where metal lines dominate the cooling rate (at sufficiently high metallicities). For lower-mass halos, cooling by H and He is important, and Λ has a different temperature scaling.

[5]Note that there is evidence that star formation–driven outflows have typical velocities $\sim v_{\rm c}$ (see Section 2.3.1), so it can be difficult to distinguish gas that has been shocked-heated by gravitational versus feedback processes, especially after mixing.

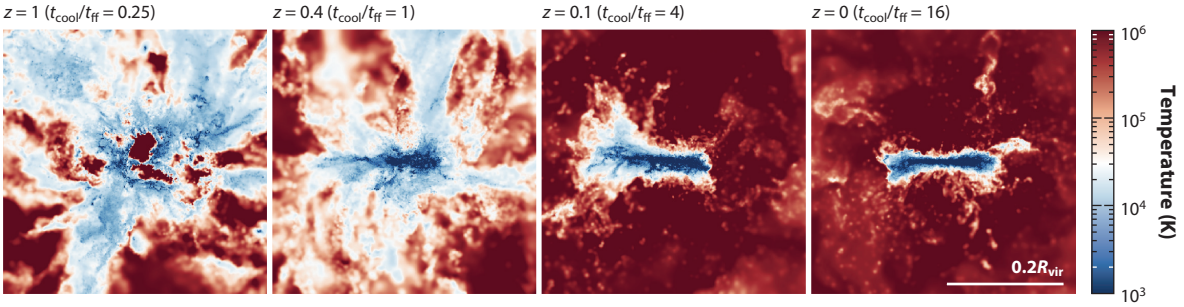

Figure 4

Time sequence, from $z = 1$ to $z = 0$, showing how the CGM completes virialization in a cosmological zoom-in simulation with realistic stellar feedback of a dark matter halo of final mass $M_h \sim 10^{12}$ $M_\odot$. Each panel shows gas temperature for a slice through the halo center and is labeled by the ratio of the cooling time of the hot, shocked gas to the free-fall time evaluated at 10% of the virial radius. The virialization of the CGM completes when $t_{cool}/t_{ff} \gtrsim 1$ in the inner regions but hot gas is present at larger radii earlier. Before the inner CGM virializes, the central galaxy is surrounded by a highly inhomogeneous mixture of cold and hot gas. Following virialization of the inner CGM, the halo is filled with hot, relatively uniform gas extending to the boundary of the central disk galaxy. Figure adapted with permission from Stern et al. (2021a). Abbreviation: CGM, circumgalactic medium.

Although it is beyond the scope of this review to discuss detailed observational predictions, we note here some possible observational implications of outside-in virialization: (*a*) The cool inner CGM should give rise to a high incidence of strong low-ionization absorbers, such as MgII, at small impact parameters. (*b*) Because most sight lines to background quasars intersect the outer CGM rather than the inner CGM (due to area weighting), the presence of hot gas at large radii may contribute to the prevalence of multiphase gas inferred in observations across a wide range of halo mass. These expectations appear broadly consistent with the existing quasar absorption line data (e.g., Tumlinson et al. 2017). Outside-in virialization would also be good news for observations that aim to detect low surface brightness rest-UV emission from the CGM, because emission is most sensitive to the inner CGM (luminosity $L \propto n^2$), and these wavelengths probe cool gas (Morrissey et al. 2018). These observations can potentially test models of CGM virialization, but the observational signatures in emission have yet to be worked out quantitatively.

2.1.4. Cold streams and where we expect them. Because the inflow is subsonic in virialized gas, the sound crossing time $t_s = r/c_s$ is short enough for pressure waves to smooth out density fluctuations. As a result, accretion of hot gas tends to proceed quasi-spherically. However, cold gas can clump into much finer-scale structures, as we discuss extensively in Section 3. Cosmological simulations predict that cold gas inflows often form filamentary structures known as cold streams or cold flows (e.g., Kereš et al. 2005, 2009b; Dekel et al. 2009a; van de Voort et al. 2011). **Figure 5*a*** shows examples of such cold filaments in $M_h \approx 10^{12}$ $M_\odot$ halos at $z = 2$ in cosmological zoom-in simulations evolved with the moving-mesh code `Arepo` (Nelson et al. 2016). The simulations shown in the figure neglect galactic winds and do not include cooling by metal lines, so these halos are significantly above the threshold mass for CGM virialization, which is lower than that for metal-free gas. In this regime, the narrow cold filaments are seen to coexist in the CGM with a volume-filling hot phase. Cold streams have been the subject of much attention because in some regimes they could be a primary mode of gas accretion for galaxies (e.g., Dekel et al. 2009a), although whether and when this is the case remains unclear as it depends on whether the cold gas survives all the way to the central galaxy during infall through the CGM, as well as the efficiency with which hot gas is accreted.

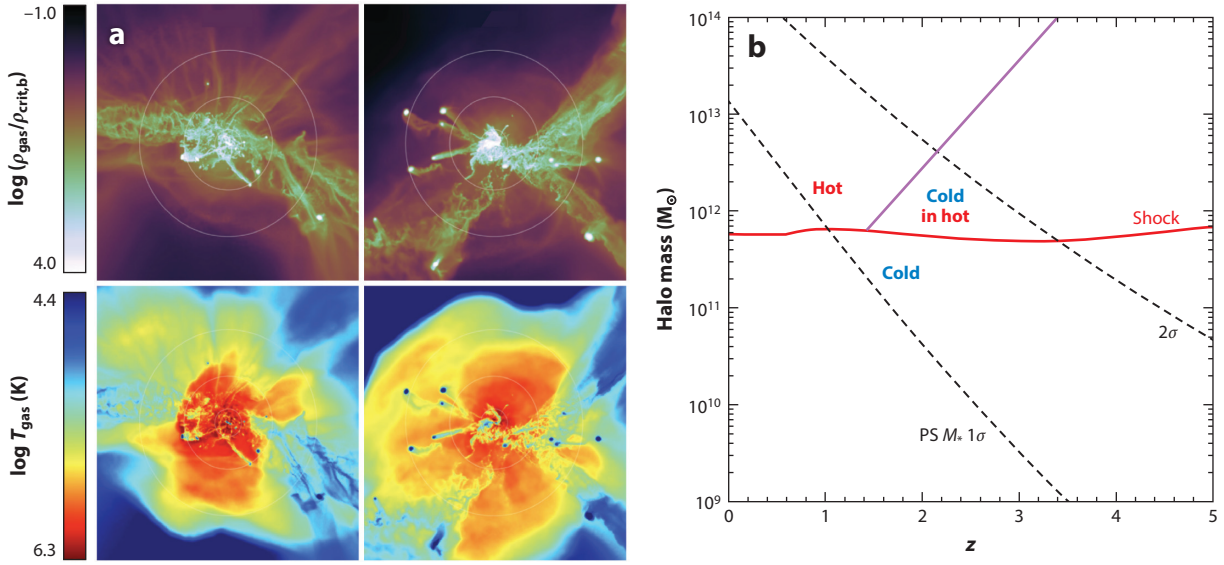

Figure 5

(*a*) Renderings of gas density and temperature in halos of total mass $\approx 10^{12}$ $M_\odot$ at $z = 2$ in cosmological moving-mesh, zoom-in simulations. These simulations neglect galactic winds, so the CGM structure is primarily set by the physics of gas accretion. In this regime, a hot virialized CGM fills most of the volume out to beyond the virial radius (*outermost circles*) but dense, clumpy cold streams penetrate deep into the inner halo. Panel adapted from Nelson et al. (2016). (*b*) Analytic theory for the threshold halo mass below which accretion onto galaxies is cold and above which it is hot, assuming a postshock gas metallicity of 0.1 $Z_\odot$ (*solid red*). This is similar to the M_{thres} mass shown on the right in **Figure 3** but is derived from a virial shock stability argument rather than cooling flow physics. The threshold halo mass is nearly constant with redshift. The inclined purple line shows a model for the maximum halo mass M_{stream} below which cold streams can persist in a hot CGM. The black dashed curves show the characteristic mass of newly forming halos versus redshift, corresponding to 1σ and 2σ fluctuations in PS theory. Figure adapted with permission from Dekel & Birnboim (2006). Abbreviations: CGM, circumgalactic medium; PS, Press–Schechter.

How can cold streams exist above $M_{thres} \sim 10^{12}$ $M_\odot$? Dekel & Birnboim (2006) proposed an explanation in terms of the geometry of the large-scale structure, which also provides insight into why cold streams in massive halos appear to be a high-redshift phenomenon (e.g., Kereš et al. 2005). The idea is that halos of different masses are, on average, located in different regions of the cosmic web. Although low-mass halos tend to be embedded in large-scale filaments whose cross sections are larger than halo radii, high-mass halos tend to reside at the nodes, where large-scale structure filaments meet. Therefore, whereas the environment of low-mass halos is roughly isotropic on the scale of the virial radius, high-mass halos are fed by collimated structures. What constitutes a high versus a low halo mass in this context is determined by the nonlinear clustering scale, M_{nl}, i.e., the halo mass corresponding to density peaks that become exponentially rare. The key point is that M_{nl} increases with time due to the growth of structure, so it is smaller at high redshift. **Figure 5*b*** shows the halo mass versus redshift corresponding to 1σ and 2σ peaks in Press–Schechter theory (Press & Schechter 1974). The panel also shows the threshold mass above which virial shocks are stable according to Dekel & Birnboim's (2006) analysis (similar to the M_{thres} based on the cooling flow argument outlined in Section 2.1.3). This plot shows that halos of mass M_{thres} are common ($<1\sigma$) and can be considered low mass at $z \lesssim 1$ but become increasingly rare ($>1\sigma$) above this redshift. Thus, above $z \sim 1$ halos more massive than M_{thres} are increasingly fed by collimated large-scale structure filaments. The higher densities in filaments, relative to the mean densities in halos, imply shorter cooling times. By contrast, the free-fall times are set

primarily by the global mass distribution in halos and are mostly unchanged. Their short cooling times enable gas filaments to remain cold as they fall into massive halos. The cooling times can be further shortened by compression of the cold streams by the volume-filling hot phase. The inclined line separating the hot and cold-in-hot regions in **Figure 5*b*** shows a simple analytic model from Dekel & Birnboim (2006) for the redshift-dependent maximum halo mass for which cold streams are expected in hot halos, $M_{\rm stream}$, based on a comparison of timescales taking into account the overdensities of filaments feeding massive halos. In this model, $M_{\rm stream} \approx (M_{\rm thres}/fM_{\rm nl})M_{\rm thres}$, where $f \approx 3$ is a dimensionless factor calibrated from numerical simulations.

Although this estimate for the maximum mass of halos expected to contain cold streams is a useful guide, it neglects a number of important questions regarding the survival of cold gas, especially as it interacts with a hot phase. Whether cold streams survive during infall into halos depends on processes, such as shocks and fluid mixing instabilities, that are not well resolved in cosmological simulations. We discuss the small-scale physics of cold gas survival in much more detail in Section 3.2.

Some early results on cold streams using cosmological simulations were questioned because they were obtained using traditional smoothed particle hydrodynamics (SPH) methods, which were shown to suppress fluid mixing instabilities and can lead to the artificial survival of cold gas (Agertz et al. 2007, Sijacki et al. 2012). Although the detailed properties of cold streams remain uncertain because of the relatively low resolutions in cosmological simulations, there is currently a broad consensus between different modeling methodologies that the existence of cold streams is a robust theoretical prediction. Cold streams are found not only in cosmological simulations evolved with modern SPH codes, which have been improved to more accurately capture mixing, but also in simulations using adaptive mesh refinement (AMR), moving mesh, and mesh-free codes (for a comparison including several of these methods, see Stewart et al. 2017). It is also noteworthy that cold streams are found in simulations that vary by orders of magnitude in resolution, ranging from large cosmological boxes to cosmological zoom-in simulations focusing on individual halos (e.g., Nelson et al. 2013, 2016). Nevertheless, it is important to keep in mind that cosmological simulations still fall short of capturing all the physics relevant to cold gas formation and survival, so the theory of cold streams could still evolve substantially. Approaches that incorporate insights from small-scale studies will play an important role going forward (see Section 4).

On large scales, interactions with galactic winds and with satellite galaxies can also modify the properties of cold streams. For example, galactic winds (including winds blown by dwarf galaxies embedded in cold streams) can puff up the cold gas distribution (Faucher-Giguère et al. 2015, Nelson et al. 2015). The increased cold gas cross section in halos due to winds and galaxy interactions (see also Section 2.4) has important implications for observables, such as the predicted cross section for Lyman limit absorption.

2.1.5. Absorption and Lyα emission from cold streams. Cold streams are of interest as observables in the CGM owing to their relatively high densities and their temperatures $T \sim 10^4$ K. In absorption, cold streams are predicted to manifest as HI absorbers with columns in the range of $N_{\rm HI} \sim 10^{16}$–10^{20} cm^{-2}, corresponding to Lyman limit systems (LLSs) and partial LLSs (e.g., Faucher-Giguère & Kereš 2011; Fumagalli et al. 2011b, 2014; Faucher-Giguère et al. 2015; Hafen et al. 2017). Cold streams may in fact dominate the incidence of these strong absorbers at most of the redshifts where they are observed (e.g., van de Voort et al. 2012), and metal-poor LLSs have been interpreted as detections of cold streams infalling from the IGM that have not yet been significantly enriched by feedback processes (e.g., Fumagalli et al. 2011a, Ribaudo et al. 2011).

In emission, cold streams may be important in explaining spatially extended structures known as Lyα halos (e.g., Steidel et al. 2011) or the more extreme Lyα blobs (e.g., Steidel et al. 2000,

Matsuda et al. 2004, Cantalupo et al. 2014). One possibility is that gravitational energy is released as Lyα cooling radiation during the infall of cold streams (e.g., Dijkstra & Loeb 2009). A simple estimate shows that cooling radiation could in principle be very important. Let $\dot{M}_{\rm gas}$ be the gas accretion rate in the halo and $\Delta\Phi$ the difference in gravitational potential as gas falls from the IGM down to the inner halo. Assuming a Navarro–Frenk–White potential (NFW; Navarro et al. 1997) with concentration $c = 5$ and a gas accretion rate $\dot{M}_{\rm gas} = f_{\rm b}\dot{M}_{\rm tot}$, where $\dot{M}_{\rm tot}$ is an average total mass accretion rate following Neistein & Dekel (2008), the cooling luminosity $L_\alpha^{\rm cool} \approx f_{\alpha,{\rm eff}}\dot{M}_{\rm gas}|\Delta\Phi| \approx 4 \times 10^{43}\ {\rm erg\ s^{-1}}\ f_{\alpha,{\rm eff}} M_{12}^{1.8}[(1+z)/4]^{3.5}$, where $f_{\alpha,{\rm eff}}$ is an efficiency factor quantifying how much of the gravitational energy is released in the Lyα line and $M_{12} = M_{\rm h}/(10^{12}\ {\rm M}_\odot)$ (see the appendix in Faucher-Giguère et al. 2010). This luminosity is comparable with observed Lyα halos.

However, the temperature of cold streams puts them on the exponential part of the Lyα emissivity function. Namely, the Lyα emissivity powered by collisions is $\epsilon_\alpha^{\rm coll} = C_\alpha(T) n_{\rm H\,I} n_{\rm e}$, where C_α is the collisional excitation coefficient, $n_{\rm H\,I}$ is the neutral hydrogen number density, and $n_{\rm e}$ is the free electron number density. The collisional excitation coefficient scales as $C_\alpha \propto T^{-1/2}\exp(-T_\alpha/T)$, where $T_\alpha \equiv h\nu_\alpha/k \approx 1.2 \times 10^5$ K, and ν_α is the Lyα frequency. This exponential dependence on temperature makes theoretical predictions for cooling radiation highly uncertain (Faucher-Giguère et al. 2010, Rosdahl & Blaizot 2012, Mandelker et al. 2020b). In simulations, the predictions are sensitive to the numerical methods used to model the hydrodynamics (because of the importance of fluid mixing instabilities and weak shocks) and radiation (because it alters the ionization structure, and photoionization also heats the gas). The structure of TMLs at the boundaries between cold and hot gas, discussed in Section 3.4, is relevant as TMLs may be where much of the energy dissipation occurs, but these layers are not resolved in cosmological simulations.

Alternatively, extended Lyα emission can be powered by recombinations following ionization by stars or AGNs. These recombinations can occur either in the ISM (H II regions) or, for ionizing radiation that escapes galaxies, in the CGM. In the case of Lyα photons produced within galaxies, diffuse halos can be formed by resonant scattering with neutral hydrogen in the CGM (e.g., Dijkstra et al. 2006, Gronke et al. 2015). For reference, the Lyα emission powered by stellar radiation in H II regions $L_\alpha^{\rm SF} \approx 10^{43}\ {\rm erg\ s^{-1}}\ f_{\alpha,{\rm esc}}{\rm SFR}_{10}$, where $f_{\alpha,{\rm esc}}$ is the fraction of Lyα photons that avoid destruction by dust and escape the medium and ${\rm SFR}_{10} = {\rm SFR}/(10\,{\rm M}_\odot\ {\rm year}^{-1})$ (e.g., Leitherer et al. 1999). This is comparable with the Lyα luminosity of cooling radiation, which in part explains why it has been difficult to unambiguously identify what powers observed sources (scattering can in principle be tested using polarization; Dijkstra & Loeb 2008). In the case of ionizing radiation that escapes galaxies, Lyα photons can be produced in the CGM via fluorescence, i.e., recombination emission powered by ionizing photons absorbed in the halo (Cantalupo et al. 2005, Kollmeier et al. 2010). The Lyα emissivity from recombinations $\epsilon_\alpha^{\rm rec} = f_{\alpha,{\rm rec}}\alpha_{\rm H\,I}(T) n_{\rm H\,II} n_{\rm e}$, where $f_{\alpha,{\rm rec}}$ is the average number of Lyα photons produced per recombination ($f_{\alpha,{\rm rec}} \approx 0.68$), $\alpha_{\rm H\,I}(T) \propto T^{-0.7}$ is the recombination coefficient, and $n_{\rm H\,II}$ is the ionized hydrogen number density. Although recombinations are not as sensitive to temperature as collisional excitation, the recombination emissivity is sensitive to gas clumping (the emissivity is proportional to the clumping factor $C = \langle n^2\rangle/\langle n\rangle^2$). This dependence on the clumping factor has been used to infer unexpected small-scale structure in the cold gas in the halos of some luminous Lyα blobs (Cantalupo et al. 2014, Hennawi et al. 2015). This has led to the proposal that the CGM could be filled with a fog or mist of tiny but high-density cold clouds; the physics of these tiny clouds is covered in Section 3.3.

Galactic winds can also power extended emission by depositing mechanical energy into the CGM, which can then be radiated away (Taniguchi & Shioya 2000, Sravan et al. 2016). Even if the ultimate energy source for extended emission (whether it be radiation or mechanical energy from stars and/or AGNs) originates from galaxies, cold streams may be important to explain Lyα

emission on halo scales. This is especially the case in massive halos exceeding the threshold mass $M_{\rm thres} \sim 10^{12}$ $M_\odot$, above which the volume-filling phase is expected to be hot. If all the halo gas were hot, most of the emission would be expected to come out in X-rays. Cold streams and other cold gas structures in halos, such as a possible cold fog (Section 3.3), can scatter Lyα photons that escape galaxies or otherwise ensure that a significant fraction of the energy deposited into the CGM is radiated in Lyα rather than in higher-energy bands.

2.1.6. Effects of CGM virialization and accretion mode on galaxies. Much of the interest in the CGM is rooted in the presumption that the physics of gaseous halos plays an important role in the formation of galaxies. In particular, there is broad but indirect observational evidence that CGM virialization is important for galaxy evolution. The characteristic luminosity of galaxies, $L_\star$ (above which the galaxy stellar mass function is exponentially suppressed), corresponds to a roughly constant halo mass $M_{\rm h} \sim 10^{12}$ $M_\odot$ (weakly dependent on redshift). This is also the halo mass scale above which the fraction of galaxies that are quiescent rises above ∼50% (e.g., Behroozi et al. 2019). In the past few years, observations of spatially resolved galaxy kinematics have suggested that the $L_\star$ mass scale is consistent with the emergence of large disk galaxies (e.g., Tiley et al. 2021). This mass scale, termed the "golden mass" by some authors (e.g., Dekel et al. 2019), is similar to the halo mass at which the CGM is theoretically expected to complete virialization (see **Figure 3**).

Despite the substantial evidence that CGM virialization correlates with major changes in galaxy properties, whether and how CGM physics affect galaxy evolution remains an active area of research, with basic questions still the subject of debate. We summarize below some ideas that have been proposed for how CGM processes could affect galaxy evolution for $L \sim L_\star$ galaxies, and which in our view deserve deeper investigation.

2.1.6.1. A quasi-isotropic, hot CGM is necessary for effective preventative feedback. There is a broad consensus that in order to explain the observed population of "red and dead" galaxies at the massive end, it is not sufficient for feedback to eject gas from galaxies. There must also be preventative feedback that prevents halo gas from cooling and raining onto galaxies at overly high rates (e.g., Bower et al. 2006, Croton et al. 2006). In the most massive halos, this feedback is often assumed to come from jets powered by AGNs, but wider-angle winds powered by either AGNs or supernovae (SNe) can play a role (Type Ia SNe can be energetically important in ellipticals with old stellar populations; e.g., Voit et al. 2015). An idea often discussed in this context is that preventative feedback only becomes important after most of the CGM has become hot and quasi-isotropic (e.g., Kereš et al. 2009a). This is because only in this limit can feedback keep the gas hot. In contrast, when there are massive inflows of clumpy or filamentary cool gas, the smaller geometric cross section of the inflows strongly reduces the efficiency with which feedback couples to accreting gas.

2.1.6.2. Pressure fluctuations change at the order-of-magnitude level at inner CGM virialization. Whether the CGM is virialized or not also changes the boundary conditions of the central galaxy. Both idealized simulations (Stern et al. 2019) and cosmological simulations (Stern et al. 2021a) show that when the inner CGM virializes, there is a change from order-of-magnitude thermal pressure fluctuations in the gas around the galaxy (prior to virialization) to a roughly uniform pressure (after virialization). Large pressure fluctuations in the inner CGM create paths of least resistance through which feedback can more easily expel gas from the galaxy. Thus, we may expect that large-scale galactic winds will be stronger and reach farther out before the CGM virializes. There is some evidence from galaxy-formation simulations with resolved ISM physics that star formation–driven outflows are suppressed when the inner CGM is virialized, such as around

Milky Way–like galaxies at $z \sim 0$ (e.g., Muratov et al. 2015, Stern et al. 2021a). Large pressure fluctuations in the inner CGM may also make it difficult for the ISM to reach a statistical steady state, which could result in highly time-variable (or bursty) SFRs (e.g., Gurvich et al. 2023).

2.1.6.3. CGM virialization changes the buoyancy of supernova-driven outflows. Keller et al. (2016) and Bower et al. (2017) proposed a related but different effect of CGM virialization on outflows. These authors suggested that SN-inflated superbubbles are buoyant in the CGM prior to virialization, so that outflows can be "lifted" in the CGM by buoyancy forces, but that the bubbles would cease being buoyant once a hot CGM develops. These authors argued that stellar feedback would therefore become ineffective at expelling gas once the CGM virializes. They furthermore hypothesized that this would lead to the accumulation of gas in galaxy centers, which would allow nuclear black holes to start growing more rapidly. If correct, this mechanism would represent another connection between CGM virialization and AGN feedback. Similar phenomenology regarding accelerated black hole feeding starting around L_*, found also in other simulations, has however been attributed by other authors to changes in star formation–driven outflows due to either confinement by gravity or the pressure fluctuations effect mentioned above (e.g., Dubois et al. 2015, Byrne et al. 2023).

2.1.6.4. Hot accretion promotes the formation of thin disks by making angular momentum coherent. Recently, Hafen et al. (2022) reported evidence from cosmological simulations that hot-mode accretion promotes the formation of galaxies with thin disks, such as that observed in low-redshift Milky Way–like galaxies. The basic idea is that gas from large-scale structure enters dark matter halos with a broad distribution of specific angular momentum (sAM). When the gas falls in toward the galaxy as cold clumps or filaments, spatially separated gas parcels are causally disconnected. In this regime, the cold gas reaches the halo center supersonically with a still-broad sAM distribution and tends to form stars in irregular or thick disk morphologies. In contrast, when the gas accretes onto the central galaxy in a smooth, subsonic cooling flow, the sAM distribution becomes coherent (i.e., narrow) before accretion onto the galaxy, and stars form in a thin disk configuration.[6]

The role of the gas accretion mode in determining the morphology of galaxies is an example of how there is not yet a consensus on the role of CGM physics in galaxy formation. Although the recent work mentioned in the previous paragraph highlights the role of hot mode accretion in the formation of thin disks, a substantial body of work has instead emphasized the role of cold streams in feeding massive disks at high redshift (e.g., Dekel et al. 2009b). These results are not necessarily inconsistent because the disks in the massive, high-redshift regime are highly turbulent and geometrically thick. More work on the role of the gas accretion mode on the formation of disk galaxies will be important, including special attention to how results vary as a function of halo mass and redshift.

Despite the plausible causal CGM mechanisms summarized above, we must stress it has proved challenging to disentangle whether CGM changes cause changes in galaxy properties or whether changes in CGM and galaxy properties simply correlate. For example, analytic arguments suggest that the mass scale of CGM virialization is similar to the mass scale in which SN-driven outflows become confined by gravity (e.g., Lapiner et al. 2021, Byrne et al. 2023), so it is possible that outflows are suppressed around the same time the CGM becomes virialized, but neither change

[6]It is likely that the net result depends not only on how the gas accretes but also on feedback, because absent feedback we would expect a thin gas disk to eventually form as a result of dissipation, even if the sAM distribution is not initially coherent (as in other astrophysical settings, e.g., protoplanetary disks that form in turbulent molecular clouds).