THE FORMATION OF MASSIVE STARS

S. W. STAHLER
University of California at Berkeley

F. PALLA
Osservatorio Astrofisico di Arcetri

and

P. T. P. HO
Harvard-Smithsonian Center for Astrophysics

We review observational and theoretical evidence bearing on the origin of high-mass stars. The conventional theory of protostellar infall successfully accounts for objects between about 2 and 10 $M_\odot$. After formation, these stars undergo quasistatic contraction with radiative interiors. In this stage, they correspond to Herbig Ae and Be stars. More massive objects have no pre-main-sequence phase and disrupt their infalling envelopes through strong winds and radiation pressure. Nevertheless, there is spectroscopic evidence for ongoing collapse in a number of systems associated with compact H II regions. The striking concentration of massive stars toward the centers of young clusters provides an important clue. The indication is that stars well in excess of 10 $M_\odot$ arise from the coalescence of already existing cluster members of lower mass. Coalescence may occur after the merger of dense molecular cores that contain the lower-mass stars.

I. OBSERVATIONAL BACKGROUND

Our picture of stellar birth is most securely established for objects of solar and subsolar mass. Pre-main-sequence stars in this regime, the well-studied T Tauri class, exist in abundance within a few hundred parsecs. Conversely, one must cast one’s observational net out to the ends of the Galaxy to sample a comparable number of O and B-type stars. It is true that the high luminosity of a massive star renders it conspicuous even at such distances and beyond. On the other hand, the object’s prodigious radiative output, coupled with the momentum of its stellar wind, quickly and effectively disperses natal cloud matter. This circumstance has impeded progress in understanding the formation mechanism.

All theoretical considerations must begin with the facts at hand. In this limited review, we will focus exclusively on our own Galaxy. Hence,
we will be unable to cover such fascinating topics as the starburst phenomenon, whose proper understanding may well shed light on massive stars generally. In any case, it has long been known that most Galactic O and B stars reside within loose associations ranging from tens to hundreds of parsecs in diameter (Blaauw 1991). Their large internal velocities, typically about 4 km s\(^{-1}\), doom these systems to expansion and eventual dispersal. The very nomenclature of OB associations reflects the historical tendency to focus on their brightest and most massive components. On the other hand, the infrared array surveys undertaken over the past decade have revealed a multitude of low-mass members (e.g., McCaughrean and Stauffer 1994). It is now clear that the very luminous stars in a typical association represent the tail of a much broader distribution. Where sufficient data are available, the latter resembles the field-star initial mass function (Brown et al. 1994; Massey et al. 1995a). This agreement bolsters the long-held view that most stars in the Galaxy originate from OB associations (Roberts 1957; Miller and Scalo 1978).

Discerning the properties of such a stellar group requires first a measure of its distance. The bulk of association members lie on the main sequence, so one may utilize the venerable technique of spectroscopic parallax. Establishing a distance then allows one to assess the masses of the visible stars. Here, both spectroscopic and dynamic methods are available. In the former, one fits model atmospheres to the observed narrow-band spectrum. The free parameters in such models are the effective temperature \(T\text{_{eff}}\) and the surface gravity \(g = GM_*/R_\star^2\). In principle, one could also obtain \(T\text{_{eff}}\) directly from a broadband color. For massive stars, however, such color indexes as \(B - V\) change very little with spectral type and are unreliable as temperature indicators (Massey et al. 1995b). The model atmosphere calculation also gives the bolometric correction, so that the \(M_V\) value now yields an estimate for \(L_{\text{bol}}\). The star’s mass then follows from its position within the main sequence in the \(L_{\text{bol}} - T\text{_{eff}}\) plane. For post-main-sequence objects, one finds the evolutionary track that passes through the star’s position in the diagram (see, e.g., Schaller et al. 1992). Alternatively, one may determine \(R_\star\) through the blackbody relation between \(L_{\text{bol}}\) and \(T\text{_{eff}}\). The mass then follows from knowledge of the surface gravity: \(M_\star = gR_\star^2/G\) (Kudritzki and Hummer 1990; Maeder and Conti 1994).

For O stars that are members of binaries, one may also obtain the mass through orbital Doppler shifts. Burkholder et al. (1997) have recently studied a sample of early-type spectroscopic binaries for this purpose. They find good agreement between the dynamical mass determination and that obtained spectroscopically. Unfortunately, their sample only extends to stars up to about 15 \(\text{M}_\odot\). For objects of higher mass, the surface gravity technique generally yields smaller \(M_\star\) values than does placement on evolutionary tracks. The origin of this discrepancy, which can be up to a factor of 1.5, is not yet clear. With the possible exception of the Pistol Star in the Galactic center (Figer et al. 1995), the current upper mass limit is held
by stars in the clusters Trumpler 14 and 16. Kudritzki et al. (1992) used
the spectroscopic method to derive masses of 100 $M_\odot$ for several objects
classified as O3.

The techniques just outlined have allowed detailed examination of the
few dozen OB associations within the nearest few kpc, and rapidly improv-
ing surveys of more distant systems (Garmany 1994). It is hardly surpris-
ing that the best-studied group is also the closest. This is the complex of
associations strung out along the Orion molecular cloud, at a distance of
450 pc. The nucleus of each association contains a tight grouping of sev-
eral O stars (Brown et al. 1994). Within the Orion B cloud, the grouping
also coincides with a dense agglomeration of molecular gas, as evidenced
through CS emission (Lada et al. 1997). This gas contains numerous in-
frared sources that represent low-mass stars in the act of formation. The
most famous collection of high-mass stars is, of course, the Trapezium that
powers the Orion Nebula. Here, the brightest object is $\theta^1$ C, which has a
spectral type of O6 and a mass of 33 $M_\odot$. The glowing Nebula itself is
energized by the ultraviolet emission from this object and its companions.
Even more massive stars are located in the OB1b subgroup, with $\zeta$ Ori A
topping the list at 49 $M_\odot$.

The relative proximity of the Orion region has also allowed the most
complete census of the low-mass complement to the brightest stars. The
Trapezium itself sits near the center of a heavily populated and dense clus-
ter, whose full extent was first revealed in the near-infrared (McCaughey-
ren et al. 1991). About half the cluster members are optically visible, appar-
tently because the O stars have already cleared away much of the ambi-
ent gas. Hillenbrand (1997) has succeeded in placing over 900 of these
objects in the HR diagram. The stellar density near the cluster’s center
is extremely high, some $10^4$ pc$^{-3}$, but is probably not the highest in the
Galaxy. One rival is NGC 3603, an H II region located in the Carina spi-
ral arm, at a distance of 7 kpc. Here one finds six stars with mass about
50 $M_\odot$ in a volume less than 0.03 pc$^3$, and a total cluster luminosity of
order $10^7$ $L_\odot$ (Eisenhauer et al. 1998). By comparison, the luminosity of
the Orion Nebula cluster is less by two orders of magnitude.

The infrared surveys have uncovered other, less spectacular, aggre-
gates with a central concentration of massive stars. The heavy obscuration
in these regions is a hallmark of extreme youth. With the progressive loss
of gas through stellar winds and radiation, such systems may expand to
become the more familiar OB associations. However, it is important to
bear in mind that not all massive stars are found inside associations or
clusters. About 25 percent of the population are true field stars with large
galactic latitudes. The radial velocities of these objects exhibit significant
dispersion about the local value expected from galactic rotation alone. Sta-
tistical analysis of the velocities reveals that most objects are leaving the
plane rather than entering it from above or below (Gies 1987). Thus, the
stars were likely born in associations, but with speeds that were higher
than average.
An important subset of the field objects are the runaway OB stars. These have extremely high spatial velocities, typically from 50 to 150 km s\(^{-1}\). They are usually situated far from the galactic plane. The proper motions of these stars can usually be traced back to known OB associations. The real problem is their velocities, which indicate that the objects were once subjected to strong forces. Thus, each runaway could have been a member of a close binary pair, propelled to high speed when its companion detonated as a supernova. Alternatively, the stars might have been ejected from the central cluster of their parent associations as a result of close encounters with neighbors. We shall later return to consider both possibilities in more detail.

II. RESULTS FROM PROTOSTAR THEORY

A. Evolution at Intermediate Masses

This sketch of the observational landscape suggests that O and B stars may form in a cloud medium similar to that yielding low-mass objects, but somewhat enhanced in density. Unfortunately, the degree of enhancement is difficult to discern using the standard molecular line techniques. When a new cluster of O stars has been revealed, as in the case of the Trapezium, there is little gas left in the immediate vicinity. On the other hand, infrared sources of high luminosity have poorly defined stellar properties and often prove under closer examination to be collections of lower-mass objects. One promising strategy is to focus on the brightest, pointlike drivers of outflows and maser emission. Using this approach, Molinari et al. (1998) recently discovered IRAS 23385+6053, a deeply embedded object with a strong molecular outflow and an \( L_{\text{bol}} \) of order \( 10^4 \, L_\odot \). Although no measure of volume density is yet available, the spectral energy distribution indicates an \( A_V \) of 2000 mag.

Millimeter continuum studies also suggest an especially dense environment for the most luminous young stars (e.g., Mooney et al. 1995). Let us, therefore, assume provisionally that such objects originate from more massive and compact analogs to the cloud cores known to be responsible for T Tauri stars (Myers and Benson 1983). Suppose we speculate further that the ensuing collapse process is essentially the same. Then the properties of the growing protostar follow almost entirely from \( \dot{M} \), the rate of mass infall near the core’s center. This rate, in turn, is given by

\[
\dot{M} = \frac{a^3}{G}
\]  \hspace{1cm} (1)

where \( a \) is the sound speed, and where the equality holds to within an order of magnitude (see, e.g., Foster and Chevalier 1993).

A remarkable feature of equation (1) is that the predicted infall rate is independent of the parent cloud density. Even if our putative cores have a density three or four orders of magnitude larger than those normally
encountered, $\dot{M}$ depends solely on $a$ and therefore on the ambient temperature. This simplification results from our tacit assumption that the cloud configuration was in near force balance between gravity and thermal pressure prior to collapse. The infall process then spreads outward from the center at the sound speed $a$. A cloud of fixed mass $M$ with a higher density $\rho$ has a shorter free-fall time (proportional to $\rho^{-1/2}$) but brings in correspondingly less mass to the center during this interval. Magnetic support in the precollapse configuration alters this result in detail, but not quantitatively (Li 1998).

Within our adopted framework, then, the description of protostar evolution rests on a proper assessment of cloud temperature. Observers routinely measure kinetic temperatures of order 100 K in such environments as Orion BN-KL or W3, an H II region 2.4 kpc distant (Cesaroni et al. 1994; Hofner et al. 1996). The most reliable values of $T$ follow from a level-population analysis designed to reproduce observed line strengths. Similar temperatures are needed to drive chemical reaction networks (e.g., Caselli et al. 1993). All such environments, however, are critically affected by the presence of an existing massive star or stellar group. We are rather interested in the temperature just before star formation. Until our “denser cores” are discovered and analyzed in detail, this information is unavailable.

The simplest and most conservative assumption is that the precollapse temperature is independent of core density and mass and is similar to well-observed values in clouds producing low-mass stars. These values range from 10 to 20 K in nearby complexes and yield, through equation (1), $M$ values from about $10^{-6}$ to $10^{-5}$ M$_{\odot}$ yr$^{-1}$. A feasible and instructive theoretical program is to use such rates to build up protostars through intermediate and higher masses. This was the approach initiated by Palla and Stahler (1991, 1992). We may briefly summarize their main results, which have since been corroborated and extended by other workers (Beech and Mitalas 1994; Bernasconi and Maeder 1996).

The structure of a protostar reflects both the addition of successive layers of infalling matter and the heat injection from nuclear processes. The gas either impacts the star directly (after passing through an accretion shock) or first lands on a circumstellar disk and only later spirals to the surface. In either case, the deepening gravitational potential well ensures that the specific entropy of the newly added gas rises inexorably with time. This circumstance would normally force the protostar to be stable against thermal convection. Beginning at subsolar masses, however, deuterium that has been accreted along with the interstellar hydrogen ignites in the deep interior. The heat input from deuterium fusion renders low-mass protostars fully convective (Stahler et al. 1980).

Convection persists as long as fresh deuterium can be dragged down toward the hot center. When the star’s mass reaches a value between 2 and 3 M$_{\odot}$, its interior reverts to a state of radiative stability. The change occurs because of the decline in opacity at higher temperatures. In more
detail, convection disappears first in a shell roughly midway to the surface. This radiative barrier blocks the transport of deuterium to the center via turbulent diffusion. Nuclear fusion therefore shuts down. At this point, the protostar consists of a thermally inert central core exhausted of deuterium, along with a mantle of deuterium-rich gas extending to the surface.

Continued addition of mass drives the interior temperature ever higher. When the base of the mantle reaches a temperature of about $10^6$ K, deuterium again ignites. The burning, now confined to a thick shell, drives convection out to the subphotospheric surface layers. This energy input to the mantle dramatically swells the radius of the protostar. However, the rising interior temperature and falling opacity guarantee the eventual end of convection. Both the burning shell and the convective mantle gradually retreat to the surface, vanishing by the time the total mass has reached $4 \, M_\odot$.

Thus far the protostar’s major source of luminosity has been the kinetic energy released at the accretion shock. Note again that this shock also covers the disk surfaces if the infall has significant angular momentum (Cassen and Moosman 1981; Stahler et al. 1994). At about the same time that the burning shell disappears, the star begins to contract under the influence of self-gravity. Contraction releases additional energy. The protostar’s total luminosity is now

$$L* = \frac{G M* \dot{M}}{R*} + L_{\text{int}}$$

(2)

Here, the first term is the accretion luminosity contributed by infalling matter, while the second is the component generated by contraction. In a radiatively stable protostar, $L_{\text{int}}$ varies as $M_*^{-1/2} R_*^{-1/2}$ (Cox and Giuli 1968). Hence the second term in equation (2) rapidly overwhelms the first as $M_*$ increases. The contracting object is beginning to resemble a pre-main-sequence star.

The rise of $L_{\text{int}}$ means that the star’s thermal energy is escaping at an accelerated rate, so the contraction process itself speeds up. The protostar’s central temperature also increases sharply and soon exceeds $10^7$ K. At this point, ordinary hydrogen ignites via the CN cycle. The heat from this source stops further contraction, and the star effectively joins the zero-age main sequence (ZAMS). The precise mass for this event depends weakly on the assumed rate of infall. For $\dot{M} = 1 \times 10^{-5} \, M_\odot \, \text{yr}^{-1}$, the mass is $8 \, M_\odot$; the mass would be $14 \, M_\odot$ if $\dot{M}$ were larger by an order of magnitude. The calculations also show that the evolutionary sequence is essentially unaltered if the infalling mass traverses a disk instead of landing directly on the surface.

We have been imagining a situation in which the mass buildup persists in a steady manner. For any given star, however, $\dot{M}$ must vanish at some point, allowing the object to contract quasistatically. The manner by which infall ceases is still problematic for any stellar mass. It may be
that the powerful jets and outflows observed to emanate from embedded stars are responsible for this reversal (see below). Magnetic support of the parent cloud could also provide the ultimate barrier (Mouschovias 1976). Whatever the process, subsequent pre-main-sequence evolution is largely unaffected if $M$ dies off in a period brief compared to the time scale for contraction. Under this assumption, we may select various models from the sequence of protostars and follow their evolution at fixed mass (Palla and Stahler 1993).

Each protostar model, then, represents the initial state of a pre-main-sequence star of the same mass. Divested of its dusty infalling envelope, the star is now optically visible and occupies a well-defined position in the HR diagram. The entire sequence of protostars yields the birthline in the diagram, i.e., the locus from which all pre-main-sequence contraction begins (Stahler 1983; Palla and Stahler 1990). We have just seen that protostars with more than about 10 M$_\odot$ are burning hydrogen. In the HR diagram, the upper end of the birthline intersects the main sequence at just this point (see Fig. 1). More massive stars have no contracting pre-main-sequence phase. This theoretical prediction seems to accord well

Figure 1. Theoretical pre-main-sequence tracks in the HR diagram (Palla and Stahler 1999). Each track is labeled by the corresponding stellar mass, in units of M$_\odot$. All the tracks start at the birthline, indicated by the dashed curve, and end at the zero-age main sequence (ZAMS), which has been omitted for clarity.
with observations of Herbig Ae and Be stars, which indeed have a mass limit of about the right value (see, e.g., Hillenbrand et al. 1993).

Pre-main-sequence stars from 2 to 10 $M_\odot$ begin with radiative interiors, apart from their outer mantles of deuterium-rich gas. These mantles quickly disappear as the contraction proceeds, and they do not affect subsequent evolution. However, the contraction itself is quite different from that envisioned in the original studies of the pre-main-sequence phase (Ezer and Cameron 1967). For masses up to about 4 $M_\odot$, the star first appears below its classical track in the HR diagram. It then moves upward and to the left before joining the more horizontal path that leads to the ZAMS. This luminosity jump is the outward manifestation of an internal readjustment in entropy (Stahler 1989). That is, the structure built up through accretion is incompatible with uniform, homologous contraction. The star expels a thermal pulse as the pattern of contraction shifts.

Corroboration of the evolutionary calculations with the data on Herbig Ae and Be stars has only just begun. One interesting prediction is that the ages of these objects should be much less than previously thought. Here, we start the clock at the birthline, so that the total contraction age shrinks to zero for masses near 10 $M_\odot$. Now the record of star formation in a cluster may be read from the age distribution of its members. Application of the new evolutionary tracks to well-observed clusters may substantially alter our conception of their history and of cluster formation in general.

**B. Disruption of Infall**

We have seen that the rise of $L_{\text{int}}$ in equation (2) causes protostars of sufficient mass to contract even while they are still accreting. That is, the Kelvin-Helmholtz time characterizing the contraction, $t_{\text{KH}} = GM_\star^2/R_\star L_\star$, becomes shorter than the accretion timescale, $t_{\text{acc}} = M_\star/M$. The fact that this transition occurs at a fairly well-defined $M_\star$ value simply reflects the sensitivity of $L_\star$ to mass. Heavier protostars are burning hydrogen and have the structure of main-sequence objects. However, they are still cloaked in dusty envelopes. Once infall ends, these stars appear directly on the ZAMS in the HR diagram.

Our theoretical picture has thus far ignored the structure and dynamics of the diffuse envelope. In particular, we must consider the mechanical effect of the protostar’s own radiation on the infalling gas. The accretion shock at the stellar and disk surfaces produces photons at optical and ultraviolet wavelengths. These are absorbed by the envelope’s dust grains and reradiated into the infrared. Trapping of this infrared radiation creates thermal pressure that retards infall. The stellar luminosity also raises the envelope’s temperature until the grains sublimate at the dust destruction front (Stahler et al. 1980). The direct impact of short-wavelength photons on grains at this front provides an additional retarding force.

Radiation pressure begins to become significant in protostars of intermediate mass. Infall cannot be stopped, however, as long as the accretion
component dominates the total luminosity. Any retardation of the flow diminishes $M$, which in turn lowers $L_\ast$. Once the luminosity falls, the flow picks up again. Numerical simulations of collapsing clouds with central stars display this effect. For stars in the proper mass range, $M$ oscillates strongly without actually reversing (Yorke and Krügel 1977).

The feedback between radiation pressure and infall is cut once $L_{\text{int}}$ dominates $L_\ast$. Again, the transition occurs just where $f_{\text{KH}}$ falls below $t_{\text{acc}}$, i.e., for protostars that have joined the main sequence. The internal luminosity is now supplied by hydrogen fusion rather than contraction, but it still scales as $M_\ast^{1/2}R_\ast^{-1/2}$. Hence the retarding force on the envelope increases rapidly with mass. Within the context of spherical collapse, the result is inevitable. Wolfire and Cassinelli (1987) analyzed the dynamics of envelopes falling onto stars with $M_\ast$ exceeding 60 $M_\odot$. They found that infall was reversed unless $M$ were arbitrarily increased to at least $10^{-3} M_\odot$ yr$^{-1}$ and grains were depleted in abundance by an order of magnitude. These extreme conditions are unlikely to occur, so we are faced with a basic dilemma. (For earlier discussion of this issue, see Larson and Starrfield 1971, and Kahn 1974.)

There is little doubt that a massive star produces more than enough energy to expel its envelope. In fact, the inequality $f_{\text{KH}} < t_{\text{acc}}$ implies that the radiated luminosity can easily unbind the parent cloud before the star has gained appreciable mass. The real issue is the efficiency of energy transfer. Here, any departure of the infall from spherical symmetry is likely to play a major role (Nakano 1989).

Consider, for example, the collapse of a rotating cloud. As we have mentioned, fluid elements far from the central axis depart from radial trajectories and strike the equatorial disk. Let us suppose that the cloud’s rotation rate is sufficiently large that the disk edge lies beyond the dust destruction front. Then the outer portion of the flow never attains the temperature that would create high radiation pressure. Once an element has entered the disk, the relatively large optical thickness could shield it from further heating as it spirals inward. Note, however, that O stars with no protostellar envelopes, such as those in the Trapezium, destroy both their own disks and those of neighboring stars through photoevaporation (Hollenbach et al 1994; Johnstone et al. 1998). Hence, this accretion scenario can work only if the radiation color temperature is significantly lowered before the photons strike the disk.

That portion of the infall closer to the rotation axis must still bear the full brunt of the radiation pressure. In spherical symmetry, this gas brakes and reverses its velocity. The fluid within a two-dimensional flow can respond by veering off laterally. Calculating this effect self-consistently is a difficult (and still unsolved) problem in radiation hydrodynamics. Jijina and Adams (1996) have presented a simplified, but instructive solution in which they retain an isotropic radiation field. They find that incoming elements can penetrate only to a “turnaround” surface that grows with time.
The density in their steady-state flow formally diverges at this locus. More realistically, the strong deceleration would create a standing shock front. This would effectively replace the accretion shock in lower-mass protostars, but it would be spatially removed from the star and disk. We may picture incoming gas entering the shock obliquely, then turning toward the equator within a condensed shell of postshock matter. It remains to be seen whether such a flow pattern is stable or whether the shell tends to break up.

Stellar winds act as a further impediment to infall. Those from visible O and B stars are radiatively driven and have the highest velocities along the main sequence (Ignace et al. 1998). There is no apparent reason why massive protostars should not also produce such flows while still accreting. Theoretical work in this area has focused almost exclusively on the mechanical interaction of winds with static envelopes. Here we may broadly distinguish two cases. Wind material of relatively low speed (below about 300 km s\(^{-1}\)) quickly cools after impact to produce thin, dense shells. Traditionally, there has been much interest in these curved surfaces in relation to low-mass stellar jets (Canto 1980). The faster winds from O and B stars cannot radiate as efficiently after the shock, because of the ionization of metallic coolants. In the extreme case of strict energy conservation, these winds inflate hot bubbles that expand into the external medium (Königl 1982).

How a wind of any speed impacts a collapsing envelope is still poorly understood. In rotating collapse, the centrifugal repulsion lowers both the infall density and ram pressure within an expanding central column. The diminished density alone would cause an initially isotropic wind to lengthen in the polar direction (Mellema and Frank 1997). Wilkin and Stahler (1998) have recently calculated the steady-state interaction of a spherical wind and a rotating, collapsing flow. They assumed the existence of a narrow, shocked shell, so the calculation would need to be modified for massive stellar winds. Nevertheless, their general results are of relevance.

The two most important characteristics of rotating infall are that it is anisotropic and that it is time dependent. The second feature stems from the fact that the specific angular momentum in the parent cloud increases away from the central axis. Because the collapse process in a marginally unstable configuration spreads outward at the sound speed \(a\), the angular momentum being transported inward steadily rises. Concurrently, the region of lowered density and ram pressure also expands.

The infall is nearly spherical relatively early in the collapse. At this stage, the wind material cannot progress far before it falls back to the star and disk. In the calculation of Wilkin and Stahler, this fallback occurs through motion within the narrow shell. As the impinging ram pressure falls, the shell elongates, especially in the axial direction. Eventually, the tips of the configuration erupt in a bipolar fashion. The wind now flows out unimpeded within a solid angle that continually widens.
Returning to massive stars, the collapse must already be in a relatively advanced stage by the time a large $M_*$ has accumulated. Thus, we expect centrifugal distortion of the infall to be severe, and for the wind to have little difficulty in breaking out within the polar region. Calculations assuming a hot bubble rather than a thin shell arrive at a similar result, at least for static envelopes (Frank and Noriega-Crespo 1994). In reality, of course, both stellar winds and radiation pressure operate simultaneously. Our discussion indicates that the only portion of the infall with a chance to survive is that which impacts the outer disk. It remains to be seen whether this material can indeed spiral inward and reach the stellar surface.

### III. EVIDENCE FOR INFALL

#### A. Observational Strategy

The current difficulties in understanding how infall proceeds would be greatly alleviated by direct, empirical measurements of cloud collapse. Unfortunately, the problem is inherently difficult for any stellar mass. First, true dynamical infall occurs only at a radius $r$ that is relatively close to the central object. Second, the magnitude of the infall speed scales as $(M_*/r)^{1/2}$. For solar-type protostars, the envelope gas reaches a speed of 1 km s$^{-1}$ only at $r = 0.004$ pc, or about 9" at a distance of 100 pc. Direct imaging of the region is therefore infeasible, and one turns to spectroscopic techniques. However, the amount of matter actually participating in collapse is typically small compared with surrounding material in the telescope beam. Absorption by cooler foreground gas and confusion with outflows, rotation, or turbulence cast doubt on putative detections.

The search for infall around low-mass protostars has recently concentrated on the detection of asymmetric line profiles. This technique, resurrected by Zhou et al. (1993), utilizes the fact that any substantial radial motion separates in velocity the foreground and background halves of the envelope. As the line becomes optically thick, it emanates chiefly from the nearer parts of both halves. Suppose, in addition, that the protostar creates a substantial gradient in gas temperature. Then the detected portion of the foreground half will be relatively cool, as it lies farther from the illuminating source. Conversely, the background half will appear hotter. The resulting asymmetric line profile is a promising diagnostic for infall. Several groups have carried out surveys; these include observations at multiple lines designed to test the predicted radiative transfer effect (e.g., Mardones et al. 1997; see also the chapter by Myers et al., this volume). Interferometric studies are also addressing the problem of spatial resolution.

Observations of high-mass protostars have several distinct advantages. The greater depth of the gravitational potential implies that the magnitude of infall speed increases. We also expect a greater amount of mass involved in contraction and collapse. The observational benefit is an
enhanced optical depth in any molecular line. In addition, more species and higher-excitation transitions are amenable to study. The greater luminosity of the central star elevates the temperature of the ambient gas. In some cases, the relevant lines are seen in absorption against an H II region surrounding the source itself. The foreground gas is then separated kinematically from background material and can be examined more readily.

Unfortunately, there are also serious problems that are peculiar to the high-mass case. Because massive stars often form in clusters, the amount of diffuse gas being observed may be very large, from $10^2$ to $10^4$ $M_\odot$. The inevitable result is further confusion regarding multiple exciting sources, as well as a potentially complex morphology. There is also the fundamental constraint that massive stars are intrinsically rare, so that the distance to the star of interest may be several kpc. Thus, even when the collapsing volume is physically large, it subtends a small angle. Researchers have again turned to interferometers for detailed examination of infall around high-mass stars.

**B. Infall at Large and Small Scales**

While many molecular clouds associated with massive stars have now been observed (see the chapter by Kurtz et al., this volume), few show any indication for rapid, inward motion. Two exceptions are the complexes known as W49 and W51. Both contain a large number of ultracompact H II regions. The small sizes of such regions imply relatively short lifetimes. Thus, the simultaneous presence of many regions calls for some kind of global synchronization of the massive-star formation process. We shall return later to consider this issue.

Consider the case of W51. The first evidence for large-scale motion came from Berkeley-Illinois-Maryland Array (BIMA) observations of HCO$^+$ (Rudolph et al. 1990). Red shifted absorption features were seen; these appear to grow stronger at higher angular resolution. The indication was that the entire molecular cloud complex, containing some $4 \times 10^4$ $M_\odot$, could be undergoing global collapse. Here, the maximum infall speed actually observed was 20 km s$^{-1}$, over a sizescale of 1 pc.

This original interpretation suffered from a number of difficulties. First, the infall velocity did not appear to vary either with radial or transverse distance. Second, the amount of material presumed to be undergoing collapse seemed far too high, at least in comparison with the much smaller total mass of the existing stellar population. New interferometric observations, using a variety of molecular lines, have now shown that the dynamical effects are more localized than previously thought, and also more complicated. The molecules employed have been NH$_3$ (Zhang and Ho 1995; Ho and Young 1996), along with CS and CH$_3$CN (Zhang et al. 1998).

The more recent observations show that infall, if present, is confined to a few distinct, compact cores. At these sites, one finds broadening of the
Figure 2. Spectral line profiles toward W51e2 (left) and W51e8 (right). The various molecular transitions are indicated on the left and become more optically thick from top to bottom. Notice how the red wing goes into absorption in W51e2, where there is an embedded, bright H II region.

spectral lines, with this broadening always centered on the exciting star. When resolved with sufficient angular resolution (< 1″), position-velocity diagrams show the classic “C” or “O” shapes consistent with radial motion projected along the line of sight. Furthermore, the redshifted portion of the line is seen in absorption when there is an intervening H II region (see Fig. 2). Away from the H II region, both the blue- and redshifted portions are seen in emission and at equal intensity. This behavior suggests that we are observing the front and rear halves of a contracting or collapsing surface. Note finally that those lines exhibiting the broadening are more optically thick and elevated in temperature. This trend is also consistent with infall arising from the warmer and denser cloud interior.

From the resolved radial motion, the total gravitating mass at each site is of order 200 $M_\odot$, within a region of size 0.1 pc. Thus, it appears that both diffuse and stellar matter are important dynamically. Note that the original idea of global collapse was largely based on an absorption feature in HCO$^+$, redshifted by 8 km s$^{-1}$. This absorption apparently stems from a foreground screen that is both cool and tenuous. The more recent observations suggest instead that most of the complex is not undergoing
collapse. The localized motion appears instead to be infall onto individual stars, continuing after the initial formation epoch. Note that the associated ram pressure is large enough that it could, in principle, suppress expansion of each star’s H II region. However, a quantitative model of the stellar environment is still lacking.

Other isolated regions within W51 appear to be contracting without any central star. These regions could be at an earlier evolutionary phase. Similar line broadening, both with and without stars, is seen in the OB cluster G10.6–0.64 (Ho and Haschick 1986; Keto et al. 1987, 1988; Ho et al. 1994). Here again, we find redshifted absorption and blueshifted emission, all associated with denser and hotter gas. Further study of clouds at high angular resolution should reveal other examples in the near future.

In the case of low-mass stars, there are also signs of contraction over length scales as small as 100 AU. Studies of $^{13}$CO emission indicate both rotation and infall within flattened structures that may represent circumstellar disks (see the chapter by Myers et al., this volume). The difficulty in applying molecular-line techniques to massive stars stems from the angular resolution. A length of 100 AU at a distance of 1 kpc subtends only 0.1". At this resolution, the minimum brightness detectable in a line is a few times $10^3$ K. Probing the kinematics requires that we turn to nonthermal emission, such as masers.

Several recent experiments have succeeded in using Doppler shifts in the 22-GHz maser line of H$_2$O to probe motion within 100 AU of massive stars. Torrelles et al. (1996, 1997, 1998a,b) mapped the masers in W75N, NGC 2071, and Cepheus A, utilizing the VLA in its largest configuration. Here, an angular resolution of 0.08" was achieved. Each of these systems contains multiple O and B stars, radio continuum jets, and extensive molecular outflows. In each case, Torrelles et al. found a disklike distribution of masers surrounding a B star. The configurations are oriented perpendicular to the axes of the radio jets and centered on the jet sources. Detected motions of the masers indicate some combination of rotation and contraction. These results, while extremely promising, represent but a first step. Within the next few years, very long-baseline interferometry (VLBI) studies will be able to supply proper motions for the maser sources. Combined observations of transverse and radial velocities should greatly improve our understanding of the motion around massive stars.

IV. FORMATION WITHIN CLUSTERS

A. Mass Segregation

Our quantitative, theoretical efforts have thus far been a straightforward extension of those successfully applied to lower-mass objects. The disruptive effect of O and B stars is so great, however, that such an approach may be fundamentally flawed. It may simply be impossible for a star well in excess of 10 M$_\odot$ to continue building up in a steady manner. Could it
be that the final mass accumulates much more rapidly, so as to avoid the problems associated with winds and radiation pressure? One can begin to address this issue by looking once more at the environments where O and B stars actually form.

We have emphasized from the start that most massive objects are not born in isolation but within clusters and associations. We also noted that the latter are unbound, freely expanding entities. Careful measurement of the proper motions sometimes allows us to reconstruct the original stellar distribution (Blaauw 1991). For example, the Upper Scorpius association was a football-shaped swarm, some 45 pc in length, about $4.5 \times 10^6$ yr ago. This dimension is in accord with those of present-day molecular cloud complexes. Such findings reinforce the belief that the majority of O and B stars originate in such complexes. Moreover, the birth sites plausibly consist of relatively tight, deeply embedded clusters like those of Orion.

Most clusters contain only low-mass stars. Those few that remain gravitationally bound following dispersal of their gas may be examined in detail through the optical spectra of their stars. All such studies find a strong tendency toward mass segregation (e.g., Raboud and Mermilliod 1998); that is, the average stellar mass increases toward the system’s center, which is also the densest region. This trend is evident, for example, in the Pleiades. Although the entire cluster of some 800 stars has a radius of 10 pc, the brightest members all reside within a much smaller core. In this case, the system age is about $10^8$ yr, long enough for dynamical relaxation to have played a significant role. Gravitational scattering gradually leads to energy equipartition, so that heavier, slower-moving stars sink toward the deepest region of the potential well. Meanwhile, they impart higher velocities, and wider orbits, to stars of lesser mass.

The clusters spawning O and B stars are much younger than the Pleiades. Their embedded nature makes a complete stellar census impossible. Nevertheless, there is growing evidence that mass segregation still occurs. Consider, for example, the Monoceros Molecular Cloud. Within this complex lies the populous young cluster NGC 2264. Since the gas here has been partially cleared away, a large fraction of the cluster members are optically visible. Careful mapping of the stellar density reveals two concentrations: one surrounding the brightest O star, S Mon, and another associated with HD 47887, again the most massive object in its vicinity.

Several recent studies of Herbig Ae and Be stars have revealed significant trends in clustering and mass segregation. Hillenbrand (1995) first noted a correlation between the mass of such a star and the number of nearby, lower-mass objects. Her finding has been confirmed and extended by Testi et al. (1997, 1998a,b), who examined 44 fields around stars ranging in spectral type from A7 to O7. With increasing mass of the central star, the volume density of surrounding objects also rises. Thus, A-type stars are never accompanied by conspicuous groups. Conversely, rich clusters
with densities exceeding $10^3$ pc$^{-3}$ only appear around stars earlier than B5, corresponding to a mass near 6 $M_\odot$.

Could this trend represent an evolutionary effect? The typical Ae star in these surveys is at least several million years old, so the absence of clustering might conceivably reflect prior dispersal through dynamical relaxation. The recent studies, however, find no correlation between the age of a star and the presence of a cluster; only the mass seems to count. In addition, the clusters surrounding early-type stars all have similar sizes, regardless of their total membership. The typical cluster radius, about 0.2 pc, is reminiscent of molecular cloud cores.

Returning to the case of Orion, we note again that optical spectra are available for about half the stars surrounding the Trapezium. The partial clearing of gas within this Orion Nebula Cluster must be due to the O stars themselves. Placement of the visible members in the HR diagram indicates a cluster age of about $1 \times 10^6$ yr (Hillenbrand 1997). In a recent series of $N$-body simulations of the system, Bonnell and Davies (1998) find that dynamical relaxation is insignificant over this time. Yet mass segregation is undeniably present, as can be seen by the central location of the Trapezium stars themselves. More generally, Figure 6 of Hillenbrand and Hartmann (1998) shows how stars in excess of 5 $M_\odot$ crowd toward the cluster’s geometrical center.

The primary question, then, is not how massive stars find their way to the densest, inner region of a cluster. The problem is rather why they preferentially form in that location (Zinnecker et al. 1993). Addressing this issue theoretically will require broadening our focus from the mutual interactions of the member stars to the structure and dynamics of their parent cloud medium. Again, the Orion Nebula Cluster illustrates this important point. Here the distribution of stars is not spherically symmetric, as one would expect in a relaxed swarm of point masses. Instead, the contours of projected stellar surface density form concentric ellipses, with an aspect ratio of about 2:1 (Hillenbrand and Hartmann 1998, Fig. 3). The orientation of these ellipses coincides with the filamentary structure evident in CO maps of the nearby molecular gas (Bergin et al. 1996).

One should hardly be surprised, at least in hindsight, that the stellar distribution in a system this young retains a memory of the diffuse medium from which it arose. After all, a similar match holds in much looser systems of low-mass objects such as that of Taurus-Auriga (Gomez et al. 1993). It seems profitable, in any case, to view the birth of massive stars as an event associated with especially dense and heavily populated clusters. Molecular line studies of clouds harboring O and B stars have demonstrated vividly how coherent structures are shredded and dispersed within a few times $10^6$ yr (Leisawitz et al. 1989). On the other hand, low-mass stars form over an interval longer by an order of magnitude (Strom et al. 1993). It appears, then, that both the parent cloud and its emerging cluster slowly condense throughout the longer timescale. Only when the
central density crosses some threshold does the rapid formation of massive stars begin.

B. Binarity and Stellar Coalescence

If most high-mass stars indeed form in the middle of dense clusters, then such crowded regions might also be expected to produce an elevated number of binary and higher-order multiple systems. Now the origin of binaries in general is still a matter of uncertainty, so the reasoning here cannot be ironclad. Nevertheless, there are several intriguing lines of evidence bearing on this issue.

A major advance in our understanding of early stellar evolution has been the discovery that binarity is the rule, rather than the exception. This fact has long been known with regard to main-sequence objects. According to Duquennoy and Mayor (1991), 57 percent of G dwarfs in the solar neighborhood have at least one companion. The corresponding figure is estimated to be 42 percent for M stars (Fischer and Marcy 1992). Optical and infrared studies of nearby cloud complexes find a similar, or even higher, incidence of multiplicity among T Tauri stars (for a review, see Mathieu 1994). Hartigan et al. (1994) and Brandner and Zinnecker (1997) have examined the ages of several dozen wide T Tauri binaries by placing both components in the theoretical HR diagram. The estimated ages are sufficiently close that the typical binary must form in situ rather than through the accidental binding of two stars initially dispersed within the parent cluster.

There is no reason why the same, local formation mechanism should not apply to binaries containing O and B stars. What are the properties of such systems? Let us begin with $q$, the ratio of secondary to primary mass. In the better studied G-dwarf binaries, the typical $q$ value is well under unity. Indeed, the distribution of secondary masses is consistent with the field star initial mass function (IMF) (Duquennoy and Mayor 1991). For M stars the distribution is much flatter (Fischer and Marcy 1992). Recall that the primary in the latter case has a mass no greater than $0.5 \, M_\odot$, so the secondaries represent an even lower-mass population. The flattening of the $q$ distribution could thus reflect the similar feature displayed in the IMF for stars of a few times $0.1 \, M_\odot$ (Scalo 1986). In summary, the evidence is that both components of present-day binaries have masses drawn statistically from the same distribution.

Complete surveys of binarity among O stars are unavailable. Nevertheless, it is clear that the secondary masses are not distributed according to the IMF. If they were, there would be only a tiny probability for the companion to be an O- or even a B-type star, yet such doubly massive pairs are not hard to find. Eggleton et al. (1989) have confirmed and quantified the statistical anomaly of this situation. Close examination shows that, in many of the tighter pairs amenable to spectroscopic study, the components are exchanging mass. A notable exception is S Mon, the O7 star
in the NGC 2264 cluster. Radial velocity measurements and interferometry reveal that a detached companion is orbiting with a period of 24 yr. The system is both a spectroscopic and astrometric binary, so that one may obtain both masses, which are 35 and 24 $M_\odot$ (Gies et al. 1997).

Within associations, the fraction of O and B stars that are visible binaries appears similar to that of lower-mass systems (McAlister et al. 1993; Mason et al. 1998). Our initial expectation of a higher multiplicity may therefore be mistaken. However, we must bear in mind the severe difficulty of finding solar-type companions to massive objects, considering both the luminosity contrast with the primary and the greater distances involved. There is also a severe lack of data for systems that are too wide for spectroscopic detection, yet too close for the components to be spatially resolved.

It is interesting that the proportion of binaries declines markedly for the field O stars located well outside associations (Mason et al. 1998). Among runaway O and B stars, the fraction is essentially zero. We mentioned earlier the theory for these objects that posits an original close binary. One star leaves the system at high speed after its even more massive companion detonates as a supernova. Detailed evolutionary calculations show that the initially more massive star (the primary) transfers a great deal of mass to the secondary before the primary explodes (for a review, see van den Heuvel 1993). Usually the supernova does not disrupt the binary. Instead, the system’s center of mass accelerates from the recoil accompanying ejection of the supernova shell. The binary thereafter consists of the now-massive secondary along with a neutron star or black hole. Such, at least, is the prediction of theory. In fact, no such system has ever been detected.

We are left with the alternative hypothesis: that the runaways originate from clusters as a result of close, dynamical encounters. Such high-speed ejections are actually a well-known effect in stellar dynamics (Chapter 8 of Binney and Tremaine 1987). Most common are three-body encounters, in which a single star approaches a pre-existing binary pair. After a complicated interaction, either the incident star or one of the pair leaves at a speed comparable to the internal orbital velocities in the original binary. Whatever its composition, the pair that remains must tighten as a result of the encounter, as a result of energy conservation.

Leonard and Duncan (1990) have followed these dynamical ejections in $N$-body simulations of massive clusters. They found that binary-binary encounters play a significant role in this case. More intriguing is the fact that the ejected objects are generally the least massive of those involved in the encounter. This rule is also obeyed in clusters of ordinary stars, whenever a broad mass spectrum is present (Van Albada 1968). The implication for runaway O and B stars is that the relevant central region of the parent cluster is severely deficient in low-mass members (Clarke and Pringle 1992). In addition, the original binaries must have been biased toward unit mass ratio, just as the observations indicate.
The emerging picture is that the centers of evolving, dense clusters become efficient factories for the production of both single and multiple massive stars. If we are to reject the traditional notion that such objects form through the collapse of individual cloud cores, then their precursors must be stars already crowded into the region. Could it be that the highest-mass stars form through the coalescence of low- and intermediate-mass cluster members?

One potential difficulty here is the time required for stellar encounters. The radii of stars are so tiny that direct collisions are highly improbable, even in the densest cluster environments. In this regard, Larson (1990) has emphasized that gravitational focusing increases the effective cross section, as does the presence of circumstellar disks (see also Clarke and Pringle 1993). Bonnell et al. (1998) have recently advocated a picture in which the stars competitively sweep up ambient, diffuse gas. These studies indicate that in a region like the Orion Nebula Cluster, with a stellar density of order $10^3 \text{ pc}^{-3}$, one would expect roughly one collision every $10^6 \text{ yr}$. Thus, the timescale per se may not be an insurmountable problem.

In the absence of a detailed physical model, the actual outcome of such close encounters is unclear. It is plausible that the two stars first form a binary, whose orbit then decays. However, the $N$-body studies demonstrate that even a few tight binaries forming at a cluster’s center quickly and efficiently disperse the remaining stars (Terlevich 1987). Clarke and Pringle (1993) have pointed out that collisions between stars and their disks would alleviate this problem, since the excess energy would go into disruption of the disks rather than ejection of stars. For the disks to soak up enough energy, they must have masses comparable to those of their parent stars (McDonald and Clarke 1995). Such massive structures are not presently observed, but theory does posit their existence at a stage when the star is still forming out of its molecular cloud core (Stahler et al. 1994).

Our discussion thus indicates a link between the formation of massive stars and the general problem of binary formation. More importantly, these theoretical considerations stress once more the critical role of the cloud environment. If massive objects indeed form through coalescence, then the interacting units are unlikely to be bare stars, or even stars with disks. The relevant clusters are so young that most of the lower-mass objects near their centers are still deeply embedded, whether or not they are accreting protostars. The colliding units may therefore be dense cores containing single stars or binaries. Since the cross section for core interaction is much greater, the collision rate would be interestingly high even in the moderately crowded clusters forming Be stars. Price and Podsiałowski (1995) have considered the encounters of cloud fragments containing protostars, and hypothesize that one or both stars will be stripped entirely of its envelope. However, gauging the true outcome of a core-core interaction will require more detailed calculations that include the magnetic support of these structures.
Assuming that the embedded stars remain within the merged cores, theory must also confront the matter of how these objects lose enough energy and angular momentum to finally coalesce. An interesting early effort in this regard is that of Boss (1984), who followed the orbital evolution of a binary embedded within an idealized, uniform-density medium. Here, the point masses are subject to gravitational torquing from the nonaxisymmetric background gas. After taking into account the varying phase angle between the binary and its medium, Boss found interesting, quasiperiodic behavior in the orbital motion but no secular loss of angular momentum. The implication is that the phase angle does not vary; that is, the binary itself generates the disturbance that brakes its motion. Pursuing this line of reasoning, Gorti and Bhatt (1996) have invoked dynamical friction with an ambient cloud core. Both these calculations were directed toward the issue of close binary formation, but they also represent promising first steps for the problem at hand.

It is evident that a number of difficult theoretical problems must be confronted before coalescence becomes a fully quantitative model. The important point for now is that understanding massive stars requires a departure from traditional ideas. It is no longer profitable to think of these objects as arising from the same basic mechanism that produces their low-mass counterparts. Massive star formation appears instead to be a critical event in the life of a populous cluster, an event touched off when the central density, in both gas and stars, exceeds some threshold value. This new picture should help provide a conceptual framework as we continue to probe the origin of the still enigmatic stars of highest mass.

REFERENCES

THE FORMATION OF MASSIVE STARS


