CRITICAL ASSESSMENT:
Gallium nitride based visible light emitting diodes

Rachel A. Oliver*
Department of Materials Science and Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS

Abstract
Solid state lighting based on light-emitting diodes (LEDs) is a technology with the potential to drastically reduce energy usage, made possible by the development of gallium nitride and its alloys. However, the nitride materials family exhibits high defect densities and, in the equilibrium wurtzite crystal phase, large piezo-electric and polarisation fields arising at polar interfaces. These unusual physical properties, coupled with a high degree of carrier localisation in devices emitting visible light, result in ongoing challenges in device development, such as efficiency "droop" (the reduction in efficiency of nitride LEDs with increasing drive current density), the "green gap" (the relatively low efficiency of green emitters in comparison to blue) and the challenge of driving down the cost of LED epitaxy.

Introduction to LEDs
Solid state lighting based on light emitting diodes (LEDs) represents a key energy saving technology in the fight to slow the usage of dwindling fossil fuel reserves and reduce greenhouse gas emissions. Electricity generation leads to one third of greenhouse gas emission in the United States\textsuperscript{1}, and approximately 20\% of this energy is used for lighting\textsuperscript{2}. LED lighting is expected to halve the energy cost of lighting on a 5 - 10 year timescale, with even greater reductions possible in the longer term\textsuperscript{2} if current technological challenges can be overcome. The LEDs used for lighting are based on Gallium Nitride (GaN) and their typical structure consists of an active region of InGaN quantum wells (QWs) separated by GaN barriers sandwiched between n-type and p-type layers (see Figure 1). Under forward bias, electrons are injected into the LED from the n-side and holes from the p-side. They become trapped together in the thin InGaN QWs (which have a lower bandgap than the surrounding GaN), and recombine, resulting in the emission of a photon, whose wavelength is determined by the energy levels in the QW, and hence by the QW's thickness and composition. (A very useful and much more extensive introduction to the physics of LED operation is given by Schubert\textsuperscript{3}.)

* To whom correspondence should be addressed: rao28@cam.ac.uk
Figure 1: Schematic of a basic GaN LED structure. Note that significantly more sophisticated layer structures are included in real devices to improve performance. For more details see Schubert.\(^3\)

In contrast to the traditional incandescent bulb, where light is only generated as a byproduct of heat generation, the emission of photons upon carrier recombination does not require any energy to be wasted in the form of heat, and this is the fundamental physical advantage of solid state lighting. The development of solid state lighting has been made possible by the rise of GaN: earlier efficient LEDs were based on materials with narrower bandgaps, producing lower energy red and infra-red photons. Up-conversion of these low energy photons to achieve the shorter wavelengths required for white light is both difficult and inefficient, whereas wide-bandgap materials such as GaN produce short wavelength photons which can either be easily down-converted or combined with longer wavelength LEDs to produce white light. The development of solid state lighting is thus intimately tied to the development of GaN as a material, which is outlined below.

A brief history of nitride LEDs

The first GaN-based LED was developed by Maruska et al.\(^4\) as early as 1973. These early devices, grown by halide vapour phase epitaxy (HVPE), were not p-n diodes of the type illustrated in Figure 1, but metal-insulator-semiconductor (MIS) diodes, with rather limited efficiency. The use of a MIS architecture was necessitated by the difficulty that early researchers found in growing p-type GaN. Early HVPE single crystalline GaN had a high concentration of impurity oxygen,\(^5\) and was thus n-type with electron densities\(^6\) in excess of \(10^{19} \text{ cm}^{-3}\). p-type doping was not achieved\(^7\) until the late 1980s when Amano et al.\(^8\) used metal-organic vapour phase epitaxy (MOVPE) to achieve greatly reduced levels of background doping. It was also necessary to irradiate the Mg-doped material with a low energy electron beam to achieve a hole density sufficient to make the first p-n junction LED.

Shortly afterwards, Nakamura et al.\(^9\) achieved p-type GaN by the simpler method of annealing an Mg-doped GaN layer in a hydrogen free atmosphere. Mg forms an electrically inactive complex with hydrogen in MOVPE-grown GaN (where hydrogen is ubiquitously present, either as a carrier gas or from the dissociation of ammonia to provide active nitrogen). Either low energy electron beam irradiation or annealing can dissociate these acceptor-hydrogen neutral complexes, resulting in enhanced p-conductivity\(^10,11\). (It should be noted however that Mg, even when not
complexed with hydrogen, is a deep acceptor in GaN with an ionisation energy of ca. 200 meV\textsuperscript{12}. Hence, even for dopant densities of the order of $10^{19}$ cm\textsuperscript{-3}, hole densities remain low, in the $10^{17}$ cm\textsuperscript{-3} regime. This is a challenge even for modern devices.

Nakamura and Mukai\textsuperscript{13} went on to develop methods for the growth of InGaN which allowed bright photoluminescence (PL) in the violet to blue spectral range, depending on the indium content of the InGaN layers. This enabled the growth of the first double heterostructure and QW LEDs in this materials system, similar to the device shown in Figure 1, with greatly enhanced output power compared to a simple p-n junction. This achievement paved the way for the rise of GaN as a key optoelectronic semiconductor, and eventually resulted in the Nobel prize for Nakamura, Amano and Akasaki.

To comprehend the scope of the achievement of these early pioneers, one must understand that GaN differs profoundly from other more conventional III-V semiconductors which had been used to make LEDs emitting at longer wavelengths such as InGaAs, and InGaP. In addition to the p-doping issue, a major question which requires consideration in the nitrides but not in the conventional III-Vs is the role of dislocations. The growth of large single crystals of GaAs by the liquid encapsulated Czochralski method is a well-established technology. However, growth of good quality bulk GaN single crystals requires extremely high pressures, and although it is now commercially viable to grow such crystals using ammonothermal methods\textsuperscript{14}, bulk substrates remain extremely expensive and the majority of LEDs are grown heteropitaxially, most commonly on sapphire substrates. This results in densities of threading dislocations greater than $10^8$ cm\textsuperscript{-2}, which in conventional III-Vs would prevent successful device operation due to non-radiative recombination at defects\textsuperscript{15}. The exact reasons for the robustness of GaN-based LEDs to such high defect densities have proved controversial\textsuperscript{15,16,17,18}, but there is a consensus that some feature of the QWs in which carriers are confined act to localise the carriers so that they cannot diffuse in the plane of the QW and are thus prevented from reaching the non-radiative sites at dislocation cores. Furthermore, the polar nature of the hexagonal wurtzite crystal structure of the III-nitrides, coupled with large piezoelectric constants\textsuperscript{19} results in significant internal electric fields arising in structures with heterointerfaces. Typical blue-emitting QWs in an unbiased GaN LED grown on the (0001) plane have internal electric fields of in excess of 1 MV/cm\textsuperscript{20}, which act to separate the electron and hole (see Figure 2), reducing the electron-hole wavefunction overlap and hence the radiative recombination rate. Overall, it is imperative that LED operation is understood within the context of a highly localised system beset by large internal electric fields.
Figure 2: Schematic diagram of the band structure of an InGaN/GaN QW, which is influenced by internal electric fields resulting in separation of the electron and hole wavefunctions along the [0001] direction.

Despite the peculiarities of the nitride materials system, blue-emitting diodes based on GaN are a highly successful commercial technology. However, for their key application in solid state lighting, emission of white light is required. This is typically achieved by converting some of the blue light into other colours using a phosphor, most commonly Y₃Al₅O₁₂:Ce³⁺, which emits in the yellow when excited in the blue. However, the down-conversion process wastes energy; it would notionally be more efficient to use separate highly efficient LEDs each emitting a different colour – red, green and blue. Unfortunately, current green emitters are less than half as efficient than their blue and red counterparts, making this approach ineffective. This “green gap” is one of the key challenges for the future development of nitride LEDs, along with so-called “efficiency droop” (a reduction in efficiency of the devices at the high drive currents typically used for solid state lighting applications) and a continuing imperative to drive down the cost of these devices. These three challenges will be discussed in the remainder of this article.

Droop

Efficiency droop – reduction of the light emission efficiency with the forward current driving the LED – may occur due to device self-heating, an issue which can be mitigated by appropriate heatsinking. Of greater concern are non-thermal mechanisms, which have proved difficult to understand and this lack of understanding has hampered attempts to design low-droop devices. A number of phenomena have been suggested to occur in the high current regime, giving rise to droop, including carrier leakage from the active region, non-radiative Auger recombination, and loss of carrier localisation.

These phenomena will be assessed in more detail below, but before doing so, it is helpful to parameterise the droop phenomenon using the so-called ABC model. The simplest form of the ABC model assumes that electron leakage from the active region may be ignored, so that the current flowing through the LED may be equated to the total rate of recombination. It is also assumed that the non-equilibrium concentration of electrons (n) and holes (p) in the active region are equal to one another. In this case, the operating current, \(I_{op}\), of the LED may be expressed as follows:

\[
I_{op} = qV_R(An + Bn^2 + Cn^3)
\]
where \( q \) is the electron charge, and \( V_R \) is the volume of the active region within which carrier recombination occurs. The term in \( n \) then relates to non-radiative Shockley-Read-Hall (SRH) recombination – i.e. recombination occurring via defect-related states, and \( A \) is the SRH parameter. The term in \( n^2 \) relates to radiative recombination, with \( B \) being the radiative coefficient and the term in \( n^3 \) relates to Auger recombination with \( C \) being the Auger coefficient. \( A, B \) and \( C \) are assumed to be approximately invariant with current density. The internal quantum efficiency, \( \eta_i \), is then given by the fraction of recombination events giving rise to a photon: i.e. the ratio of the rate of radiative recombination to the total recombination rate:

\[
\eta_i = \frac{Bn}{(A + Bn + Cn^2)}
\]

Experimentally, it is the external quantum efficiency that is usually measured, which also depends on the extraction efficiency, and furthermore the actual non-equilibrium carrier density is very difficult to determine so that approaches to fitting experimental data require the basic mathematics above to be expressed in terms of observable variables. Approaches to the experimental data are detailed in references 22 and 23, and a typical fit is shown in Figure 3. Such studies show that the model provides a reasonable fit to quite a wide range of LED efficiency versus current data, as long as care is taken to avoid artefacts from self-heating and current crowding\textsuperscript{22}, and thus provide a useful parameterisation of the droop phenomenon, allowing comparisons to be made between devices. However, it represents an over-simplification of the actual physical processes involved. The simple version expressed above ignores leakage currents, although these can be included by introducing a further relationship\textsuperscript{30}:

\[
I_{\text{leak}} = aI_{\text{QW}}^m
\]

such that

\[
I_{\text{op}} = I_{\text{QW}} + I_{\text{leak}} \quad \text{where} \quad I_{\text{QW}} = qV_R(An + Bn^2 + Cn^3).
\]

Unfortunately, the leakage parameter, \( m \), has been suggested to be device design dependent\textsuperscript{30}, such that it may be difficult from the form of the operating current versus efficiency data to determine whether current leakage or other factors are paramount in controlling the observed droop. Even with this added level of sophistication, and also consideration of additional effects such as band-filling, the model does not always fit the data perfectly\textsuperscript{31}, and a wide variety of combinations of Auger, radiative and defect-related recombination rates can all provide reasonable fits\textsuperscript{37}, making it difficult to use such fits to reliably distinguish between the relevant mechanisms\textsuperscript{32}. Perhaps more fundamentally, the model ignores the impact of carrier localisation, which will certainly affect the rate of SRH recombination substantially\textsuperscript{33}, and may also make it impossible to meaningfully estimate the local rather than the average carrier density in the QW.

**Figure 3:** A typical example of the experimentally-determined relationship between internal quantum efficiency (\( \eta \)) of an LED and the current density (\( J \)) (circles) and a fit of the data using the ABC model (dashed line). Reprinted with permission from A. David and M. J. Grundman: *Appl. Phys. Lett.*, 2010, 96, 103504 [34]. Copyright 2010, AIP Publishing LLC.
Despite these various limitations, the ABC model is very widely used in the literature, particularly in attempting to assess the relative contributions of the carrier leakage mechanism and Auger recombination. Proponents of the carrier leakage mechanism suggest that the polarisation fields in the multiple QW active region promote flow of electrons through and out of the active region, and that droop can be explained by recombination occurring outside the active region. In addition to showing that a modified version of the ABC model including leakage currents can provide a good fit to the experimental data even if Auger recombination is excluded (an approach incriminated by all the caveats outlined above), this hypothesis was originally supported by the claim that droop is only observable in electroluminescence (EL) data, where carriers are injected separately from the p and n sides of the device, and not in photoluminescence (PL) data where no forward bias is present to drive the carrier leakage process. However, later studies have shown that droop is in fact observable in PL at appropriate carrier densities, weakening this contention, despite a counter-example that the processes involved in PL and EL droop may be non-identical and have different carrier densities at their onset. The difficulty in determining the true local non-equilibrium carrier density in either the PL or the EL experiment makes it almost impossible to judge the veracity of this latter statement. The carrier leakage model has been further weakened recently by Piprek, one of its original proponents, who suggests that the observed temperature dependence of LED efficiency is inconsistent with electron leakage being the primary cause of droop. Nonetheless, carefully designed experiments have provided direct experimental evidence that electron leakage and recombination outside the active region can occur, so that whilst this leakage may not be the sole, over-riding cause of droop, it would certainly be imprudent to ignore the possibility entirely when developing new LED designs.

Similarly, direct evidence from electron emission spectroscopy exists for the occurrence of Auger recombination during LED operation at high currents, but nonetheless there are aspects of the Auger recombination picture of droop which require further scrutiny. Auger coefficients have been measured, largely based on fitting of the ABC model, and the resulting values are largely in the range from $1 \times 10^{-30}$ to $25 \times 10^{-30}$ cm$^3$/s. Given the phenomenological nature of the ABC model, it is necessary to compare these values against the expected theoretical Auger coefficients to determine whether identifying the $n^3$ component of the $I$ versus $n$ dependence as Auger-related is reasonable. Unfortunately, the microscopic nature of the Auger process in InGaN – for example whether it involves two holes and one electron or two electrons and one hole - is not yet understood, and direct and indirect Auger processes, alloy scattering, and phonon coupling all need to be considered, making the theoretical calculations very challenging. However, it should be noted that most calculated values are at least one, and often up to four orders of magnitude smaller than the typical experimental values. Furthermore, the temperature dependence of LED EL suggests that the $n^3$ non-radiative process exhibits a temperature dependence which is incompatible with the usual theoretical picture of Auger recombination. Overall, the experimental and theoretical picture in support of Auger-dominated droop remains incomplete.

The poor correspondence of the temperature dependence of the EL data with the expected physics of Auger recombination has led Hader et al. to develop an alternative model in which droop occurs because the carrier localisation which prevents non-radiative recombination at defects fails to remain effective at high currents. Hammersley et al. developed a similar model, citing saturation of the available localised states as carrier density increases. Whilst this picture has supporting evidence from both PL and EL, and good agreement between theoretical and experimental data has been achieved for a limited number of samples, it has been much less studied than either carrier leakage or Auger recombination and thus subjected to less critical scrutiny. Localisation has been widely ignored in the droop debate to date, despite the fact that the basic emission mechanisms of nitride LEDs have long been believed to be reliant on it. A complete picture of droop may actually involve a combination of carrier leakage, Auger recombination and saturation of localised states, and indeed changes to the local carrier density at localised states may affect the probability of Auger recombination or leakage. It should be noted, however, that within all of the
models droop is expect to increase with carrier density as opposed to current density so that device designs which distribute the same number of carriers over a greater volume\textsuperscript{29}, for example by increasing the number of QWs in the active region, would be expected to mitigate droop no matter what its mechanism.

**The Green Gap**

White LED bulbs made from blue LEDs coupled with phosphors are finding increasing traction in the lighting market, but do not represent the most energy efficient lighting solution. By combining efficient red, green and blue (RGB) LEDs it would be possible to reduce the energy cost of lighting still further whilst simultaneously offering tuneable light bulbs in which the intensity of the different LEDs could be dynamically controlled to adjust the colour temperature\textsuperscript{62} to the user's preference. Unfortunately, the efficiency of nitride LEDs which emit green light is typically less than half that of blue and red LEDs, and other candidate materials for green emission have even less impressive performance. This absence of efficient LEDs in the green spectral region is known as "the green gap".

Fundamentally, the low efficiency of green emitters implies that compared to the blue devices these LEDs have either a lower rate of radiative recombination or a higher rate of non-radiative recombination (or indeed both). Green emitters grown on the polar c-plane of GaN require either higher indium contents or larger well thicknesses than blue emitters. Either of these changes increases the separation between electrons and holes occasioned by the internal electric fields, and thus reduces the electron-hole wavefunction overlap and reduces the radiative rate, as has been observed both in PL\textsuperscript{45} and EL\textsuperscript{46}. However, this reduction in the radiative rate cannot by itself account for how low the observed efficiencies are\textsuperscript{47}, and hence we must also explore possible increases in the non-radiative rate.

Evidence for an increased rate of non-radiative recombination in green samples has been seen in time-resolved PL measurements\textsuperscript{45} and attributed to an increase in the defect-related (SRH) recombination rate. However, PL measurements cannot identify which specific defects are culpable for the low efficiency. Given the requirement to use higher indium contents and/or larger layer thicknesses to achieve green emission, the strain energy associated with the growth of QWs is increased providing a driving force for plastic relaxation - i.e. the introduction of misfit dislocations\textsuperscript{48}. The involvement of misfit dislocations in SRH recombination is supported by the work of Langer et al.\textsuperscript{49}, who have observed a correlation between an increase in non-radiative recombination and the onset of partial strain relaxation. If misfit dislocations were the key defect, one would expect it to be relatively straightforward to design LEDs with a lower total strain energy by, for example, limiting the total number of QWs. Whilst this might be problematic for the avoidance of droop in the high current regime, where large active volumes help to reduce carrier densities, it should allow good internal quantum efficiencies to be achieved at low currents. However, green emitters seem to show comparably low efficiencies across the board, and given the general robustness of nitride LEDs to the presence of dislocations, dislocations are unlikely to be the only culprit.

Another type of extended defect, apparently peculiar to the nitrides, whose density is known to increase drastically in higher indium content, green-emitting samples\textsuperscript{50} is the so-called "trench defect" which consists of a basal plane stacking fault in the plane of the QW, bounded by a stacking mismatch boundary, and enclosing QWs with altered properties\textsuperscript{51} (The structure of the defect is illustrated in Figure 4, which also shows the eponymous trench which arises at the surface of a multiple QW when the defect is present. These trenches are filled upon growth of the p-cap in an LED). Such defects have been shown to deteriorate the efficiency of blue-emitting LEDs\textsuperscript{50}, but in that case they can be eliminated by careful manipulation of the growth conditions\textsuperscript{52}. They may still play a role in green emitters where they are likely to be more difficult to eliminate.
Figure 4: (a) Schematic of the structure of a so-called "trench defect". The name arises from the trench that arises where such a defect intersects the surface of a multiple QW structure, as illustrated in the AFM image in (b). These trenches are no longer visible after growth of the p-cap to form an LED structure. Part (a) reprinted with permission from [51]. Copyright 2012, AIP Publishing LLC.

Extended defects can, as we have seen, often be controlled in a systematic fashion in device growth. Point defects on the other hand are much more difficult to either quantify or control. In InGaN QWs, nitrogen and indium vacancies may be the predominant point defects due to the weak In-N bond. Deep level optical spectroscopy of LEDs suggests an increased density of deep levels in the band gap associated with point defects at higher indium contents. Such deep levels provide a non-radiative recombination path. Recently, Hammersley et al. have shown that where samples are grown at varying temperatures but with the same composition a greater non-radiative recombination rate is seen in the samples grown at the lowest temperatures. Since the QW structure is unchanged between these samples, there is no extra driving force for dislocation formation, but the limited diffusion lengths associated with low temperature growth may provide a route to increased point-defect incorporation. Overall, point defects are a rather overlooked topic in the nitrides, but (particularly for green emitters) the substantial challenges of understanding, quantifying and controlling them must be overcome to achieve the highest possible efficiencies.

A diverse range of potential solutions to the green gap problem is being considered. In order to increase the radiative rate, one generally needs to reduce the charge carrier separation by the internal electric fields. Using the standard polar orientation, this can potentially be achieved by engineering the QW band profile by through-thickness grading of the indium content, or by
reducing the strain in the InGaN QWs by (for example) using thick, relaxed InGaN layers as templates either grown on conventional sapphire or on alternative substrates such as cubic (111) MgAl₂O₄ spinel. However, a more radical approach is to switch to growth on an alternative crystal facet. Whilst internal electric fields across the QW can be avoided entirely by growing on a non-polar facets such as (1-100), it has proved difficult to incorporate sufficient indium to achieve green emission on these facets. However, semi-polar facets, at an angle between the polar (0001) and the non-polar orientations can be used to grow green emitters with reduced internal electric fields and higher radiative rates, although at the cost of either more complex materials processing or a greatly increased defect density.

Considering the alternative approach of attempting to decrease the non-radiative rate, it is clearly desirable to reduce defect formation. Higher growth temperatures may reduce the incorporation of point defects and impurities, and engineering the QW thicknesses, composition and number may avoid strain relaxation. Some epitaxial structures which might be used to reduce the probability of strain relaxation, such as the use of very thick GaN barriers between the QWs, may cause other problems, such as increasing the resistance of the device, making electrical injection difficult. Hence more radical solutions have been suggested, including growing green QWs in a nanorod geometry where elastic strain relaxation in the plane of the QW is possible (this will also reduce the internal electric fields) or even optically pumping green QWs with a highly efficient blue LED. This latter approach effectively uses the green QWs as a down-converter for the efficient blue device, and so obviously cannot achieve an efficiency as high as that of the original blue device. Nonetheless, the use of phosphor down-converters to convert blue to green light is currently a pragmatic approach and can yield good quality white light, although it does add expense and complexity. Another radical potential solution is to use GaN LEDs doped with rare earth elements, since electron transitions within the 4f shell of such species can result in efficient luminescence within a wide bandgap host semiconductor. This approach has yet to be widely applied, but is illustrative of the wide range of possible solutions under consideration to try and address this particularly difficult problem.

Cost

Despite the challenges outlined above, LED bulbs are already widely available which have better energy efficiencies than equivalent incandescent or compact fluorescent units. According to the U.S. Department of energy the "primary deterrent" for uptake of these bulbs is cost. In fact, such LED bulbs will save the consumer money in the long term - replacing an incandescent lamp with a current LED unit is predicted to save the consumer over $150 over the lifetime of the lamp. Nonetheless the initial purchase price remains off-putting. The scope of this article in considering how to reduce costs is really the cost of the LED itself, consisting of a substrate and the epitaxial nitride layers grown on it. However, we must note that the LED substrate and epitaxy only make up about one third of the cost of the LED package: processing the epitaxy into a device, the phosphor to create white light and (most significantly) the packaging of the device represent the majority of the cost (see Figure 5). LED packages are then incorporated into a replacement bulb unit, where they make up only 23% of the cost, other costs including drive electronics (26% of the total) and heat sinking (18%). The substantial cost of heat sinking illustrates the fact that efficiency and cost cannot be addressed separately: more efficient LEDs produce less heat and thus require less heat-sinking driving down costs. Lowering the LED epitaxy cost at the expense of efficiency is thus not a route to a cheap and cheerful product, but to an expensive, inefficient one.

Within this context, we can consider what can usefully be done to drive down the costs of LED epitaxy. Most LEDs were originally grown on 2" diameter substrates, either sapphire or SiC. By moving to large area substrates the number of useable LEDs produced in each growth run can be increased for the same reactor size, partly because there is then less area close to the edge of the wafer where the growth is disrupted. For example, in a commercial AIXTRON 2800G4 HT reactor a
44% increase in throughput may be achieved by using 6" wafers instead of 2" wafers\(^6\). Furthermore, processing costs are lower for larger wafers, particularly where automated systems developed in the silicon industry can be used\(^6\). Whilst sapphire and SiC can be obtained in large wafer sizes, the cost of large area wafers is high. Larger diameter Si wafers of very high quality are, however, widely available at low cost. However, switching to growth on Si presents its own unique challenges. There is a large lattice mismatch between Si and GaN (17%). Whilst large lattice mismatches also exist for GaN and other common substrates, on Si the problem is exacerbated by an enormous thermal expansion mismatch (46%) between the substrate and the epitaxial layer\(^6\). This results in a tensile stress in the GaN upon cooling from growth temperature, which can result in cracking. Since the likelihood of cracking rises as the GaN layer thickness increases, standard routes to reduce the high density of defects generated by the large lattice mismatch cannot be readily applied to GaN/Si since they generally require the growth of thick layers.

Strain management strategies such as graded AlGaN buffer layers\(^6\), AlN insertions\(^6\) or AlGaN/GaN superlattices\(^6\) are required to avoid cracking. Figure 6 compares whole wafer optical images of GaN wafers with and without appropriate strain management. The same strategies also help to manage the bow of the wafer. Definitions of wafer bow vary\(^9\), but it can be considered to be the height difference between the wafer surface at the wafer centre and a reference plane which is defined by the height of an equivalent flat wafer. If this height deviation is too large, the wafer will not physically fit through standard Si processing lines\(^6\), and even where this issue is avoided, bow occurring during growth can lead to local variations in temperature, which affect the composition of key elements of an LED, particularly the QWs, leading to significant non-uniformity of emission wavelength and intensity across a wafer\(^6\), which is unacceptable in the commercial context. Strain management strategies must be designed around an imperative to initiate the growth with AlN rather than GaN, since the exposure of the Si surface to metallic Ga leads to melt-back etching via the formation of a Ga-Si eutectic\(^6,7\)

Despite all these challenges, significant progress in the growth of GaN LEDs on Si has been made and has recently been extensively reviewed\(^6,6\). Commercial devices on Si are now available. Here, rather than detailing the technical achievements which have made this possible, we consider future opportunities. Amongst these is the possibility of pursuing the large area growth strategy to an even bigger scale with the use of 8" or even 12" diameter Si wafers for even greater cost reductions. The challenges inherent in this ambition are large, since for a constant radius of curvature, the value of wafer bow increases with the square of the wafer diameter\(^6\). This extra bow increases the temperature uniformity which can lead to drastically non-uniform emission across a wafer. In the worst case, stresses induced by thermal gradients can causes plastic deformation in
Figure 6: Images from a GraphicEye whole wafer optical imaging system of GaN layers grown on Si (a) without and (b) with successful strain management employing a graded AlGaN buffer. Reprinted from [72]. (The red lines in the image are an artefact of the measurement system).

the Si wafer\(^{72}\). Shaped susceptors designed to fit the bowed wafer shape at elevated temperatures at critical junctures during the growth may form part of the solution to this problem, but the demands on the design of strain management will nonetheless become ever more stringent, particularly since the use of larger diameter Si wafers may necessitate the use of thicker substrates which alter the strain balance between substrate and epitaxy necessitating a complete redesign of the strain management strategies\(^{65}\).

Si substrates offer further opportunities for cost reduction if current Si-based LED and high electron mobility transistor (HEMT) technologies based on Si substrates can be combined to allow monolithic integration of the LED control mechanisms with the light emitter, obviating the need for expensive external control electronics. This has thus far only been achieved on sapphire\(^{73,74}\), which is a non-ideal substrate in context due to its low thermal conductivity. Other novel devices have been successfully realised on Si however, most notably micro-LEDs\(^{75}\), which provide the opportunity for fast switching (1 GBps\(^{76}\)) of the LEDs for use in visible light optical communications - sometimes called "LiFi" - which may play a role in ubiquitous data transfer within the "Internet of Things". Integrating additional function into LED lighting, in addition to reducing its cost, provides another route to make it more attractive to the consumer, improving uptake of the low energy technology. Furthermore, micro-LEDs on Si are a particularly attractive option since the silicon can be easily etched away allowing free-standing LEDs to be transferred to alternative substrates\(^{77}\), including flexible substrates\(^{78}\), opening up other new opportunities for LED technology including in bio-integrated devices for medical monitoring\(^{79}\).

Other strategies in the development of cheap LED bulbs link to our earlier discussion of the challenges of achieving high efficiencies in the green. By switching to a solution utilising separate red, green and blue LEDs we can eliminate the use of phosphors, reducing the cost of each LED package by 10\(^{64}\). However, this is only workable if green devices match the performance of their blue counterparts since otherwise it would be necessary to use extra LED packages, provide additional heat sinking and develop more complex drive electronics to manage the performance mismatch. It should be noted here that in considering the interplay between performance and cost, some authors claim that the best performance in LEDs can only be achieved by the use of bulk GaN substrates\(^{7}\), and that the resulting reduction in the cost of the remaining elements of the bulb unit may make this a cost-effective approach in the long term. Currently, however the very high cost of bulk substrates means that this idea is not part of the commercial mainstream.

**Future opportunities**

The cost considerations outlined above highlight the need for LED design and development to be performed in a systems context, taking into account all elements of the final bulb into which the LED will be incorporated and the ways in which it will be used. Currently, usage of LED bulbs remains relatively low, and a timely opportunity is available to rethink the lighting of our homes, transport systems, workplaces, manufacturing and agriculture so that uptake of LEDs can result not just in reduced energy consumption, but also in improvements in our overall quality of life, health
and productivity. To achieve this, we need to consider LEDs in the context of building design, with input from the architectural, systems and human physiological viewpoints. There is evidence\(^1\) that tuning the colour of light to our physiological needs can promote health, for example mitigating seasonal affective disorder, jet-lag and some types of insomnia, and may help maintain alertness and improve performance in the workplace. The required dynamic control of light colour necessitates, as mentioned above, the use of separate red, green and blue light emitters within a bulb, rather than the use of phosphor-converted blue LEDs, highlighting again the importance of bridging the green gap. The use of energy efficient lighting to perform extra functions, such as LiFi, can help solve additional problems, such as the limited bandwidth available to satiate our ever-growing demand for wireless data transfer. Overall, LED lighting offers a great deal more than just improved efficiency. It offers the consumer enhanced controllability, new functionality and novel design opportunities\(^6\). The future for LED lighting is bright!

\(^20\) P. Lefebvre, A. Morel, M. Gallart, T. Taliercio, J. Allègre, B. Gil, H. Mathieu, B. Damilano, N. Grandjean, and J.


