The past, present, and future of fracture mechanics

B. Cotterell

Institute of Materials Research and Engineering, 3 Research Link, Singapore 117602, Singapore

Abstract

The science of fracture mechanics was born and came to maturity in the 20th century. Its literature is now vast. Perhaps the most successful application of fracture mechanics is to fatigue. However, this short paper is limited to a core topic of fracture, the initiation and propagation of fracture under monotonic loading at low strain rates.

As the author was invited to present this paper on the occasion of his 65th birthday, it is a somewhat personal view of the development and future of fracture mechanics, but it is hoped that it will interest many, especially the young researchers at the beginning of their careers. © 2002 Elsevier Science Ltd. All rights reserved.

1. Introduction

The start of new millennium is an appropriate time to stand back and look at the development of fracture mechanics. Fracture will always have a wide-ranging importance to man. An attempt to illustrate this importance is shown in Fig. 1. The first illustration is a stone hand-axe dating from the Palaeolithic era. However, man and his hominid ancestors had been fashioning stone for more than a million years when it was made. Once man built large expensive structures, he had to ensure that they did not fracture and collapse. Fracture mechanics is basically about scaling. Leonardo da Vinci (1452–1519) was the first to record an understanding the scaling of fracture and the second illustration in Fig. 1 comes from one of his notebooks [1] and illustrates his strength tests on iron wires. Galileo Galilei (1638) writing in his “Dialogues Concerning Two New Sciences” [2] was the first to give the correct scaling laws for bars under tension and bending. His illustration for a discussion of the fracture of beams is shown in Fig. 1. Size effect is very important in fracture and Galileo saw that this effect placed a limit on the size of structures, both man-made and natural, which makes it impossible to build ‘ships, palaces, or temples of enormous size in such a way that all their oars, yards, beams, iron bolts, etc. will hold together; nor can nature produce trees of extraordinary size because their branches would break down under their own weight’.

Iron, and from the 1860s, steel saw increasing structural use in the 19th century and fracture was a problem. David Kirkaldy opened his Testing and Experimental Works in 1865 and his testing mark is illustrated in Fig. 1. He published a comprehensive account of his experiments [3] and discussed some of the fracture problems with steel, the new structural material. However, the most prophetic pronouncement about the brittle fracture problems with steel that were to come, occurred in 1861 in a leading article of
"The Engineer" that is worth quoting: "Effects of percussion and frost upon iron... We need hardly say that this is one of the most important subjects that engineers of the present day are called upon to investigate. The lives of many persons, and the property of many more, will be saved if the truth of the matter be discovered—lost if it be not" [4].

Apart from the pioneering work, the understanding of fracture waited until the last century of the second Millennium. Even then it took until well into the second half of this century until a rational approach was used by engineers and taught in engineering schools. The author vividly remembers being taught the five theories of strength as an undergraduate in the early 1950s without any real explanation. A major spur to the development of fracture theory from 1940 onwards was the brittle fracture of large welded structures such as bridges, ships and oil storage containers. This period is illustrated in Fig. 1 by a photograph of the Schenectady that broke in two while lying in the fitting-out dock in January 1943. One of the most successful applications of linear elastic fracture mechanics (LEFM) is to the reliability assessment of aircraft. A picture of the ill-fated Comet is shown in Fig. 1. A fast fracture initiated from a fatigue crack that grew from an astro-navigation window and caused the loss of two aircraft in 1953 and 1954. One of the first applications of LEFM was an explanation of this fast fracture [5,6]. The next picture in Fig. 1 is the Boeing 737 that lost a top section of the fuselage over Hawaii in 1988 and yet it still managed to land safely. The cause of this fatigue failure was multisite fatigue, which is still only poorly understood. The final picture in Fig. 1 is a schematic illustration of a microelectronic package to indicate that many of current fracture interests are outside traditional engineering. What problems will the future bring? It does not take much imagination to realise that there will be more and more applications of fracture mechanics to tiny functional devices that need structural integrity. Structural problems will not go away, but most will be able to be solved with present techniques provided the problems are remembered. Alan Wells once remarked that the same fracture problem reappears every thirty years, the length of time for a new generation of engineers that have not experienced the problem.

In a short paper such as this it is possible to cover only a very small fraction of fracture mechanics. The paper will concentrate on quasi-static fracture at ambient temperatures under monotonic loading.
Unfortunately only a small fraction of those who have made major contributions to fracture can be mentioned. In Fig. 1, the names stop about thirty years ago since it becomes more difficult to know which names to omit. More names will be mentioned in the following sections, but space will prevent many others who have made important contributions. The author has tried to select those works that have had the most influence, but naturally the selection is personal. The author apologises to those whose names are omitted.

2. Development of fracture theory up until the Second World War

Until Rossmanith [7] rediscovered the work of Wieghardt [8], the first known works devoted to fracture were the two seminal papers of Griffith [9,10], which form the foundation of modern fracture theory. Griffith was motivated by the need to understand the effect of scratches on fatigue. It was originally thought that it should be possible to estimate the fatigue limit of a scratched component by using either the maximum principal stress criterion, favoured by Lamé and Rankine, or the maximum principal strain criterion, favoured by Ponclet and Saint-Venant. Griffith showed, using the results of Inglis [11], that scratches could increase the stress and strain level by a factor of between two and six. However, Griffith noted that the maximum stress or strain would be the same on a shaft 1 in. in diameter whether the scratches were one thousandth or one-hundredth of an inch deep provided that they were geometrically similar. These conclusions were in conflict with the fatigue results and led Griffith to reject the commonly held criteria of rupture. Wieghardt [8] had earlier rejected these strength criteria for a different reason. He was concerned with the paradox that the stresses at the tip of a sharp crack in an elastic body are infinite no matter how small is the applied stress. This fact led him to argue that rupture does not occur when the stress at a point exceeds some critical value, but only when the stress over a small portion of the body exceeds a critical value. The concept of a critical stress intensity factor only just slipped through Wieghardt’s fingers. Taylor [12] states that Griffith was also aware of the paradox. However, Griffith [9,10] does not mention this paradox in either of his classic papers. A possible reason was Griffith’s obvious concern that near the tip of a sharp crack the small strain assumption is violated and he was hesitant to discuss the stresses at the tip of a sharp crack.

Reasoning that a simple critical stress or strain criterion could not be used to predict fracture Griffith turned to energy concepts. He realised that a certain minimum work was necessary to produce a fracture, which for an ideal elastic material was the surface free energy. Such a system is conservative and he saw the fracture problem as just an extension of the elastic theory of minimum potential energy. All that had to be done was to consider the potential surface energy as well as the other potential energies of the system. Griffith’s global treatment of the energy balance for a cracked body was praised by Taylor [12] as ‘the first real advance in understanding the strength of materials’. The practical importance of Griffith’s work lies in his realisation that the critical stress depends on a length scale, the crack length.

Griffith performed his experiments on a model material glass. From his experiments [9], he estimated the theoretical strength of glass to be about 2 GPa. The observed tensile strength of glass was 170 MPa. Hence Griffith predicted there were flaws of the order of 5 μm. Griffith believed that the weakness of glass was due to internal flaws; and indeed believed that the surface layers might be of superior strength because flaws would be oriented parallel to the surface [9]. In his 1924 paper that Griffith [10] clearly stated that the ‘weakness (in pure silica) is due almost entirely to minute cracks in the surface, caused by various abrasive actions to which the material has been accidentally subjected after manufacture’. Griffith’s evidence was that if a strong silica rod was rubbed lightly with any other solid, it immediately lost its great strength. However, he did not state that the weakness in glass was due to surface flaws. During the 1920s Joffé [13] and others assumed that it was surface flaws that were responsible for the weakness in glass. Joffé [14]
presented his work inferring that the strength of rock salt was due to surface flaws because when the surface layer was dissolved in warm water the strength increased, at the same Delft conference as Griffith.

From the 1920s onwards there was a search for flaws in glass. The separation across a surface flaw in glass is of the order of 50 nm, only about one-tenth of the wavelength of light, and undetectable optically. It was not until 1933 that the experiments on mica by Orowan [15] proved conclusively that the reduction in strength was due to flaws. The usual tensile strength of mica is between 200–300 MPa, but Orowan obtained strengths of more than 3 GPa by stressing only the central strip of a sheet of mica using grips that were much narrower than the sheet. The small value of the usual tensile strength of mica is due to the presence of cracks at the edge of the sheet. The cleavage plane is near perfect. The first direct evidence for the existence of surface flaws in glass came by chance in 1935 during experiments by Andrade and Martindale [16] on the properties of thin films of metal, followed later a series of experiments on various glasses using sodium from a vapour to “decorate” the surface cracks [17].

The experiments demonstrating the reversibility of fracture of Obreimoff [18] in 1930 deserve a mention. The fracture of a Griffith crack is unstable so there is no possibility of reversibility. However, Obreimoff studied the fracture of mica using a stable geometry. Mica has a very pronounced cleavage plane and almost atomically perfect surfaces can be produced by cleavage. Obreimoff used a glass wedge to cleave thin lamellar of mica 0.1–0.2 mm thick from a block of mica. The fracture of such cantilever specimens under fixed deflection conditions is stable and can be analysed with the engineers’ theory of bending and is the start of a love affair between fracture mechanists and the double cantilever beam specimen. Since Obreimoff used a stable geometry, he found that a crack could grow under the combined effect of mechanical energy and moisture in the air. Obreimoff demonstrated the reversibility of fracture, the two mica surfaces re-adhering when the wedge was redrawn. Under atmospheric pressure the equilibrium measured surface energy of a healed crack was slightly less than for a virgin crack.

3. The Second World War and the brittle fracture problem

The transition temperature from ductile to brittle behaviour in structural steel of the time was around 20 °C. Problems had arisen from the brittle behaviour of steel with the introduction of high volume steel production by the Bessemer process in the 1860s. These problems were recognised by some engineers such as David Kirkaldy. However in riveted structures brittle fractures rarely caused catastrophes because a fracture was usually arrested at the edge of the plate in which it initiated. The earliest recorded case of brittle fracture in steel is that of a 75 m high by 5 m diameter, water standpipe at Gravesend, Long Island, NY in 1898 [19] a generation after the prophetic leader in “The Engineer” [4]. Electric arc welding for the construction of large steel structures was just introduced prior to World War II. Without proper fracture control, welding introduces the elements necessary for brittle fracture in steel: high residual stresses equal to the yield strength, a heat affected zone adjacent to a weld with a much higher transition temperature than the parent plate, and crack-like defects. Since a welded structure is continuous, unstable brittle fracture can easily run thorough a major part of its section and cause a catastrophe. The first brittle fracture in a large welded structure occurred just before World War II in the Vierendeel Truss Bridge in Hasselt, Belgium followed by failures in similar Belgium bridges during the war [19]. However, the Allies knew little about these fractures during the war and what started the investigation of brittle fracture in earnest were the widespread fractures in the welded Liberty Ships. There were 145 structural failures in Liberty ships where the vessel was either lost or the hull so weakened to be dangerous, a further 694 ships suffered major fractures requiring immediate repair [20]. A consequence of these failures was the setting up of the US Navy Ship Structure Committee and the Admiralty Ship Welding Committee. The British Committee was under the chairmanship of John Baker who assigned the metallurgical investigation to Constance Tipper.
Brittle fracture was initially seen as an almost purely metallurgical problem, if Griffith’s work was considered at all it was simply to point out the importance of a notch. The main aim up until the 1960s was to determine the transition temperature at which the fracture behaviour changed from ductile to cleavage. As early as 1909, Ludwik [21] explained the phenomenology of the transition from ductile to cleavage behaviour. He suggested that the cohesive strength was little affected by temperature, but there was a marked increase in the yield strength of low carbon steel as the temperature decreased so that a particular temperature cleavage fracture became easier than yielding. The effect of a notch on the transition temperature was seen to be primarily due to a constraint on yielding and Orowan [22], using Ludwik’s concept, showed how a notch would increase the transition temperature. The Charpy test [23], originally introduced to deal with the problem of temper brittleness, was one of the original and the most lasting of the small-scale notch bend tests to assess the transition temperature in steel.

The realisation that there was a size effect in the brittle fracture of steel took sometime to develop probably due to the limited size range of laboratory specimens, and as late as 1960, Biggs [20] could write: “The fundamental problems associated with size effect have received limited attention.” Size effect was recognised as early as 1932 by Docherty [24] but only slowly became widely appreciated. At the Navy Research Laboratories in Washington, Irwin [25] and Shearin, Ruark and Trimble [26] at the University of Carolina were examining size effect in the late 1940s. At the University of Illinois, Wilson, Hetchkman and Bruckner were using their huge 3,000,000 lb hydraulic testing machine to test plates 3/4 in. thick up to 72 in. wide [27].

It was Wells [28] who developed the first fracture test that fully simulated a welded plate structure. He designed a special simple 600-ton testing machine that was capable of testing 1 in. thick 36 × 36 in. butt-welded plates [29]; later models had capacities up to 4000 tons [30]. The Wells wide plate test consists of a butt-welded plate that has a fine saw cut made into the weld preparation. This saw cut is not fully buried by the weld. Since the plate is wide, full residual stresses can develop that are similar to those that would occur in normal construction. The welded plates could be tested either as-welded or after heat treatment and were welded into the test rig. The plates were cooled with dry-ice to the desired temperature before testing. The first results using the wide plate test were published in 1956 [28]. Typical results for low carbon steel are shown in Fig. 2 [31]. Those plates welded with rutile electrodes tended to have precracks at the saw cut. Fractures were often initiated in precracked specimens at low stress, or occurred spontaneously on cooling, these arrested at the edge of the tensile residual stress zone along the weld. The transition from high stress fractures at the yield strength to low stress fractures did not always occur at the Charpy transition temperature. The Wells wide plate test was widely accepted and various versions of the test were adopted around the world.

Though there are certainly size effects in the brittle fracture of steel, it is true that the problem was largely a metallurgical one. It was solved metallurgically by developing steels with lower transition temperatures. However, later metallurgical improvements to structural steel to increase their strength, brought problems with fast ductile fracture in welded oil and gas pipe lines. For comparatively thin walled pipe lines, this problem too in its turn was solved metallurgically, but in heavy sections such as found in offshore oil rigs or nuclear power plants, ductile fracture has to kept at bay by elasto-plastic fracture mechanics (EPFM).

4. The development of linear elastic fracture mechanics

Until the late 1940s, Griffith’s pioneering work [9,10] was not seen as having very much relevance to engineering. Griffith chose glass, one of the most brittle materials, for his model material. The size effect in

---

1 This belief was carried into the 1960s by some metallurgists. In the discussion of a paper by the author at an Australian conference in the 1960s, the main criticism from one metallurgist did not concern the substance of the paper, but that its author was not a metallurgist!
glass was accurately predicted by Griffith's theory. However, size effects were not so obvious in metal specimens. It was the work of Orowan [22] that led to the generalisation of Griffith's work to less brittle materials. Orowan [22] studied the depth of plastic strain beneath cleavage facets in low carbon steel using X-ray scattering. Irwin [25] noted that the energy expended in this plastic straining could be estimated from Orowan’s result. This fracture energy, $\gamma_p$, for low carbon steel around 0°C turns out to be roughly two thousand times the surface energy, $\gamma_s$. Irwin concluded that Griffith’s theory could be used if the plastic work were substituted for the surface energy [25]. Orowan presented the same idea a little later [32]. However, it was Irwin who grasped the engineering significance of the extension of Griffith’s work and went on to develop LEFM. One very interesting paper on the direct measurement of $\gamma_p$ is that of Wells [33] who used a thermocouple to measure the plane temperature wave emanating from a fast propagating fracture from which the heat source and $\gamma_p$ could be calculated.

At first Irwin’s development of LEFM was in terms of energy. He defined the elastic energy released for a unit increase in crack area, the crack extension force, $G$, [5]. A fracture would initiate when $G$ reached a critical value $G_c = 2\gamma_p$. Work on hot stretching of PMMA [34] led Kies, a collaborator of Irwin at NRL, to observe that the critical stress for a given crack size depended only on $G_cE$, where $E$ is the elastic modulus. The response of the Boeing engineers, who had initiated the work on hot stretching, was to use $(G_cE)^{1/2}$, which they termed the fracture toughness, $K_c$, in recognition of Kies, as their fracture parameter [35]. Irwin [36,37], using Westergaard’s paper [38], related $G$ to the stress field at the crack tip and introduced the stress intensity factor $K = (GE)^{1/2}$ which was also named in honour of Kies. For any symmetrical geometry, $K$ can be expressed as

\[ K = (GE)^{1/2} \]

---

2 Irwin used the term crack extension force in analogy to a force on a dislocation and chose a Gothic letter $G$ in honour of Griffith which caused some difficulties in reproducing, now fortunately a Roman $G$ is acceptable.
\[ K = \sigma \sqrt{\pi a} F(a/W) \]  

where \( \sigma \) is a representative stress, \( a \) is the crack length, and \( F(a/W) \) is a function of the geometry.

LEFM is one of the most successful concepts of continuum mechanics and has gone on the to be applied to:

1. fatigue crack growth,
2. stress-corrosion cracking,
3. dynamic fracture mechanics,
4. creep- and visco-elastic fracture.

However, there is not space for a discussion of these topics in this paper.

Apart from a few notable exceptions, such as Wells whom regularly visited Irwin in America, the development of fracture mechanics in Europe and America during the 40s and 50s was quite separate. In the main researchers in Europe were interested in low strength steels where the problem was more of transition from ductile to brittle behaviour whereas in America there was more interest in high strength steels, which were used in rocket motor cases. Griffith’s theory of fracture had little direct quantitative application to low strength steels whereas it had application to high strength steels and it is not surprising that the general development of LEFM should have occurred first in America. However, there some isolated developments in LEFM in Europe whose general significance was not seen at the time. One such paper is that by Rivlin and Thomas [39] on the rupture of rubber. As in the case of high strength metals, the deformation of rubber is essentially elastic except for a small region at the crack tip. However, the extension of rubber is very large and the deformation is non-linear. Rivlin and Thomas [39] independently proposed an application of a generalised Griffith theory of fracture to rubber. They proposed that the critical condition for catastrophic rupture would be when the energy released by crack propagation became equal to a tearing energy, \( T \), which is equivalent to Irwin’s \( G_c \), but since the deformation is non-linear, Griffith’s equations could not be used. Rubber is almost invariably tested under fixed grip conditions so that the energy released comes only from the strain energy stored. Rivlin and Thomas [39] used a graphical method of determining the strain energy release rate and showed that rupture occurred when this reached a critical value equal to the rupture energy, \( T \).

It is interesting to note that the first application of LEFM in a structural code occurred in the Australian Timber Engineering Code AS CA65 of 1972 [40] as a result of the work of Leicester in the CSIRO Division of Building Research [41].

5. The development of elasto-plastic fracture mechanics

LEFM predicts infinite stress at the crack tip, so that obviously there must be an inner core where the elastic solution breaks down. This factor was early recognised by Irwin [38] who, by simple equilibrium arguments, estimated the size of the plastic zone, \( d_p \), at a crack tip in a material with yield strength of \( \sigma_Y \) to be

\[ d_p = \frac{1}{n\pi} \left( \frac{K}{\sigma_Y} \right)^2 \]  

where \( n = 1 \) for plane stress and \( n = 3 \) for plane strain. Provided the stress field outside of the plastic zone is dominated by the \( K \)-field, then LEFM can be applied, which means that \( d_p \) must be small compared with the dimensions of the specimen. There is a regime where the plastic zone does have a significant effect on the outside stress field, but so not so great that it completely destroys the \( K \)-field. In this regime, Irwin [38]
showed that LEFM could still be used provided that an effective crack length equal to \(d_p/2\) was added to the actual crack length. However, as the plastic zone becomes large LEFM breaks down and plastic deformation has to be considered in detail. The beauty of LEFM is that it is a one-parameter model that only depends upon the state of stress and not its history. Plastic deformation is path dependent, and if that path dependence is modelled then it is impossible to have a simple one-parameter model. Fortunately up to the initiation of fracture the loading path is almost proportional and a deformational theory of plasticity is little different to a true incremental theory. A deformational theory of plasticity is identical to non-linear elasticity, but cannot model unloading.

Wells introduced the crack opening displacement concept, later called the crack tip opening displacement (CTOD), to model fracture under conditions of large plastic deformation [42]. Wells assumed that for fracture to occur there must be a critical crack tip opening. For complete yielding of a deep notch bend specimen, the CTOD could be obtained from slip line theory and the rotation of the arms since, according to slip line theory, the deformation is essentially that of rigid arms rotating about a rigid circular core. For small scale yielding Wells made an approximation based on the plastic particle of Irwin so that the CTOD, \(\delta\), is given by

\[
\delta = \frac{4}{\pi} \frac{K^2}{E\sigma_Y}
\]  

which is very close to the more exact expression obtained by Burdekin and Stone [43] from the analysis of Dugdale [44].

The \(J\)-integral, the EPFM equivalent to the crack extension force, \(G\), in LEFM, is the energy that is extracted through the crack tip singularity. Path independent integrals [45–47] had been discussed earlier, but it was Rice [48] who introduced and developed them for the fracture community. The path independence of the \(J\)-integral arises because the plastic deformation is treated as if it were non-linear elastic. For a power hardening material, the plastic stresses at the tip of a crack have a characteristic form (the HRR-field) with the strength of the singularity dependent on the strain-hardening exponent [49,50]. The \(J\)-integral, another one parameter model, has been very successfully applied to the initiation of fracture in the presence of significant plastic deformation. Its application to propagation is more problematical. The standard method adopted by ASTM [51] calculates the propagation value, \(J_R\), by assuming that the material is non-linear elastic using the work of Ernst [52].

Hutchinson and Paris [53] have shown that outside of a core of non-proportional loading the deformation is nearly proportional. Provided the region of non-proportional loading is well contained within the region dominated by the \(J\)-singularity there will exist an annular region where the HRR [49,50] field holds. If a specimen’s uncracked ligament is sufficiently large compared with the inner core of non-proportional loading and the \(J\)-stress field dominates the crack extension, the crack growth will be controlled by the \(J\)-integral [54]. In this regime, \(J_R\), is the specific essential work performed within the actual fracture process zone (FPZ) at the crack tip. However, outside of this regime, \(J_R\), contains a fraction of the plastic work. It would not be important that a fraction of the plastic work were included in \(J_R\) if it was a constant fraction, but it is a variable fraction. Different specimen sizes give different \(J_R\)-curves making the concept unreliable for other than small crack growths.

A further limitation on the application of the \(J\)-integral concept is the size of the FPZ where the material is undergoing strain softening not accounted for by normal plastic theory. Provided the FPZ is fully embedded within the characteristic \(J\)-field, then fracture occurs at a critical value of \(J\) which at initiation can be

---

\(^3\) This was the first fracture conference that the author attended and because of the absence of Wells, the author had the honour of delivering this first exposition of the concept.
identified with the specific essential work within the FPZ. The initiation of fracture is then determined by the critical value of $J$ since it is $J$ that determines the deformation within the FPZ. The $J$-integral can be extended somewhat to situations where the FPZ is outside of the $J$-field dominance. O’Dowd and Shih [55–57] have given a two parameter ($J$–$Q$) approximation of the stress field ahead of the crack tip and can be written as

$$\frac{\sigma_{ij}}{\sigma_0} = \left( \frac{J}{2\varepsilon_0 \sigma_0 I_x} \right)^{1/(n+1)} \tilde{\sigma}_{ij}(\theta, n) + Q \delta_{ij}$$

where $\sigma_0$ is the yield strength, $n$ the strain hardening exponent, and $\delta_{ij}$ is the Kronecker delta. The first term is the HRR field and the $Q$ term is essential a hydrostatic stress term. This two-parameter model has been successful in predicting the effects of constraint on cleavage fractures that initiate only after significant plastic deformation. Attempts have been made to use this two-parameter model to explain constraint effects on ductile fracture initiation, but with less success [58].

6. Modelling the fracture process zone

Both of the classic one parameter models of elastic and plastic fracture have been very successful, but they have two major limitations:

1. They can only be used if there is an initial crack, crack-like defect or notch.
2. The FPZ must be small compared with the dimensions of the specimen.

Classic elasto-plastic models have a third limitation:
3. Only small crack growth can be modelled.

Two-parameter models can ameliorate the second limitation somewhat, but can do nothing for the first. If the FPZ is modelled all three of these limitations can be overcome at the price of simplicity. The beauty of the classic one-parameter models, especially LEFM, is the simplicity. For LEFM, analytic solutions and handbooks of stress intensity factors make fracture analysis as easy as any other simple elastic problem that has a handbook solution. Many problems will always be able to be tackled in this simple manner, but there are many more that need more complex analysis using FPZ models. In the author’s view the FPZ should be defined as the largest identifiable regime where under steady state propagation, the specific dissipative work is a constant. With this definition, the crack tip plastic zone can be defined as the FPZ in classic LEFM. In cementitious materials, the FPZ is the region where microcracking, causing strain softening, takes place. In ductile fracture under large scale yielding, the region of void initiation growth and coalescence is the FPZ. In part the definition of the FPZ depends upon the range of specimen size considered. For example if in ductile fracture, the yielding is small scale over the whole size range, then the plastic zone can be defined as the FPZ, but if the scale of yielding changes from small scale in the large specimens to large scale in the small specimens, then the region of void initiation, growth and coalescence must be taken as the FPZ. The range in possible FPZ size varies enormously. In Fig. 3 the range in maximum size of the FPZ in bulk specimens large enough for the FPZ to develop fully is indicated. For specimens too small for the FPZ to fully develop, the FPZ size depends upon the specimen. In thin multilayers such as occur in microelectronics on-chip, the FPZ can be very tiny. In experiments with a typical multilayer where the weak interface

---

4 The two terms differ significantly and subtly from the first two terms in the series expansion.
is between SiO₂ and TiN the length of the yielded zone in an adjacent Al–Cu layer a micron thick is much less than a micron [59].

Barenblatt [60,61] was the first to formulate fracture mechanics in terms of a FPZ. Considering an elastic-brittle material he proposed that there was an inner zone to the crack where the atomic cohesive forces, dependent on the opening, were important. At equilibrium these forces produce an equal and opposite stress intensity factor at the crack tip so that the crack faces close smoothly to form a cusp. Barenblatt argued that if the FPZ is small compared with the crack length, the crack tip opening was steady state during crack propagation and the specific separation work a constant which essentially is the assumption in LEFM. Dugdale [44] used the concept of a simple FPZ, where the closing stress is assumed to be constant, to model the plane stress plastic zone at a tip of a crack as a fictitious extension to the real crack using the condition of smooth closure to find the extent of the plastic zone. The difference between Barenblatt and Dugdale was that the former was modelling the interatomic forces at the crack tip and the latter using the concept of a FPZ to model plastic deformation.

It was the cementitious fracture community that first made extensive use of the FPZ model. Although concrete is a brittle material, the fully developed FPZ in concrete is very large (of the order of 1 m). LEFM simply cannot be used on laboratory-sized specimens, but can still be used on very large structures such as dams. There are two main versions of the FPZ model:

1. The fictitious crack model pioneered by Hillerborg [62].
2. The crack band model proposed by Bazant [63].

If the FPZ is a region of strain softening, then it is only the microstructure that prevents the FPZ from collapsing into a line. Thus strain softening FPZs have a narrow structure, which can be approximated by a fictitious crack model. In the case of modelling an interfacial fracture, the fictitious crack model is an obvious choice, but in other situations there is little to choose between them.

6.1. The fictitious crack model

The fictitious crack model is similar to Barenblatt’s approach. The fictitious crack is assumed to extend to the tip of the FPZ, but closing stresses exist in the FPZ, which are a function of the crack opening (see Fig. 4). If there is an initial sharp crack or notch, the FPZ will start to grow at the smallest load. In a small specimen, the fracture may become unstable before the FPZ is fully established. The shape of the softening relationship is comparatively unimportant. The two most important parameters are the specific essential work of fracture, \( \Gamma_0 \), the integral under the stress–displacement curve in Fig. 4, and the maximum stress that can be sustained, \( \sigma \). The relative importance of these two parameters depends on the size of the FPZ compared with the crack length. If the crack length is small, or there is no significant defect at all, the most
important parameter is $\sigma$. As the crack length increases, the importance of $\Gamma_0$ increases until the crack length becomes large compared with the FPZ when the maximum stress loses its importance completely. In the case of an elastic-brittle material such as a cementitious material, the limiting case, of course, is simple classic LEFM. Thus in modelling the strain softening behaviour of say mortar, a simple linear relationship for strain softening will enable the maximum load in a notched bend test to be calculated quite accurately. Only if details of the load–deflection curve, such as the deep belly after maximum load, are required is it necessary to have a more detailed strain-softening relationship. For mortar and other cementitious materials, a bilinear curve enables the load–deflection curve to be modelled accurately. Finite element modelling can always be used outside the FPZ [64], but for an elastic-brittle material, it is also possible to use standard LEFM expressions for the stress intensity factor with the condition that the total $K$ at the tip of the fictitious crack must be zero [65]. Fig. 5 shows the modelling of the load–deflection curves for two mortar notch bend specimens $700 \times 150 \times 150$ mm. The bilinear stress–displacement curves for the
FPZ that give the best fits to the two load–deflection curves are shown. Once the stress–displacement curve for a particular material has been established, it can then be used to predict the behaviour of other structures. Using the bilinear curves obtained from the large specimens, the load–deflection curves for smaller beams 450×100×100 mm have been calculated, these theoretical predictions are compared with the experimental curves in Fig. 5.

Interfacial fracture is usually mixed-mode and is an especially appropriate application of the fictitious crack model. Tvergaard and Hutchinson [66] have suggested that the strong mode-mixity effect in interfacial fracture is more due to the deformation outside of the FPZ and have given the normal and tangential traction in terms of a traction potential. To avoid having to locate the tip of the FPZ, they suggest that the whole of the interface be modelled as a FPZ and include an elastic loading part to the curve (see Fig. 6). Liu et al. [67] have used this model to examine the interfacial delamination of a thin metal film in a four-point mixed-mode bending test (see Fig. 6). For small ($\tilde{\sigma}/\sigma_Y$) ratios, yielding does not penetrate the metal film completely. As the ratio ($\tilde{\sigma}/\sigma_Y$) increases so does the specific work of fracture, $\Gamma$ that includes the plastic work in the thin metal film.

6.2. The crack band model

The tendency for a FPZ to collapse to a line causes some problems with the crack band model. Either the thickness of the FPZ must be taken as a material property or a non-local definition of the strain.

---

5 If the finite element method is used to model the FPZ, it can only be one, two if there is symmetry, element wide or the solution is mesh size dependent.
must be used to prevent this instability that, in a real material, is prevented by the microstructure. Bažant has been a pioneer in the development of the theory of non-local strain [68]. Here, the crack band model using a fixed thickness for the FPZ is illustrated by the analysis of ductile fracture [69,70] using a Gurson material [71] for the FPZ. The parameters used for the Gurson material in the FPZ are: the cell size, initial porosity and critical porosity. Using the results from a three-point-bend (TPB) A533B steel specimen, Xia and Shih [72] found the parameters that gave the best fit to the experimental load–deflection curve, these parameters were then used to compare the predictions for other geometries with experiment (see Fig. 7).

7. Continuum deformation theories with a length scale

In classic continuum deformation theories, stress at a point depends upon the strain at that point. These theories are perfect unless the strain gradient is extremely large. Fleck et al. [72] made tension and torsion
tests on copper wires of different diameters (see Fig. 8). There is little difference in the stress–strain curves for the tension tests, but a very significant increase in strength in the torsion tests once the diameter is less than about 30 μm. Gradient effects in the plastic range are related to dislocation cell microstructures, which develop with deformation fields of submicron wavelengths. Strain gradient plasticity theories have been developed to explain the phenomenological behaviour—Fleck and Hutchinson have given an excellent review [73]. In ductile fracture, the FPZ where void growth occurs is of the order of a hundred microns and classic deformation theories are adequate. However, it is an issue when the fracture process is one of atomic separation. In this case classic plastic theories cannot explain how stresses high enough to cause atomic decohesion can be sustained. Strain gradient plasticity is necessary in these cases.

8. The bottom–up molecular dynamics approach

Hutchinson and Evans [74] have referred to all the models of fracture discussed above as the top–down approach. They opine that the bottom–up approach from fundamental mechanics and physics to link the atomic scale to macroscopic aspects are unlikely to be developed with adequate accuracy in the near future. However, considerable strides have been made in the bottom–up approach. Abraham et al.
[75] point out the sharp increase in the total number of atoms that can be simulated by molecular dynamics (MD) in the 1990s by the use of large-scale parallel computers. Already teraflop computing has taken the number to $10^9$; much larger numbers can be expected in the near future. There are effects that are revealed by MD that cannot be explained by a top–down approach. Thus the MD bottom–up approach can be expected to have an ever-increasing influence on fracture mechanics. Here a just a flavour of what can be learned is given.

The first example is an examination of dynamic fracture using a 2-D solid rare gas model material [75]. The specimen is a wafer with 1000 atoms on each side that can simulate the first stages of dynamic fracture. Although it might be objected that a 2-D model material is not very realistic, fast running dynamic fractures show less material dependence than almost any other class of fractures. The fracture is initiated at a notch midway along one side of the wafer. Some of the results of this simulation are shown in Figs. 9 and 10. In Fig. 9 the time evolution of a crack is shown where a grey scale is used to represent the velocity. Inverted $V_R$ are seen which are the wakes created by dislocations propagating away from the

---

6 Abraham et al. [75] makes the comment that Griffith too was criticised for using a model material.
crack tip every time the crack changes direction. The first picture in Fig. 9 is typical of the propagation when the crack speed is less than 0.32$C_R$ ($C_R$ is the velocity of Rayleigh waves). At this stage, the acceleration of the crack is smooth. However at velocities greater than 0.32$C_R$, the instantaneous crack speed becomes erratic and the average acceleration is less. The fracture surface is shown in Fig. 10. Here the familiar mirror-mist hackle development with crack velocity is seen. All these features have been recognised at the macroscale previously, but here they are seen at the nanoscale demonstrating a fractal nature of dynamic crack propagation.

The second example is also taken from Abraham et al. [75] and is of a 3-D model of a solid fcc rare gas and contains $10^8$ atoms. In this case the MD results show some unexpected results. Conventional wisdom says that fcc cubic crystals are ductile, yet in Fig. 11 cleavage occurs on a (1 1 0) plane. Again we expect the most likely cleavage plane in a crystal to be the one with the lowest surface energy, yet in Fig. 11 the fracture on a (1 1 1) surface is ductile while its surface energy is less than that on the (1 1 0) face. This second example shows that there may be fracture features to be learned from a MD simulation on only a comparatively small number of atoms.

9. Mixed atomistic and continuum methods

MD is teraflop hungry. Every atom is modelled the same, but away from defects one atom behaves very much like its neighbour. Shenoy et al. [76] have suggested a scheme whereby, away from regions of high gradient, a small subset of atoms is used to represent the energetics of all so that the total energy of the reduced subset is given by

$$E_{\text{reduced}} = \sum_{i=1}^{n} n_i E_i$$

$$E_i = \frac{1}{2} \sum_{j} \phi(r_{ij}) + U(\rho_j)$$  (5)
where $n_i$ is a weighting factor, $r_{ij}$ is the distance of the $i$th atom to the $j$th, $\varphi$ is a pairwise interaction, $U(\rho_i)$ is the embedding energy which is a function of the electron density. The reduced atomic description is stored on a finite element mesh. Fig. 12 illustrates the concept around a Lomer dislocation in Al. The atomistic and continuum regions are not tied together caused by the reduction in the number of degrees of freedom. Tadmor et al. [77] use the method to study plane strain nanoindentation of a slab 200 nm wide by 100 nm thick. Interaction of cracks and grain boundaries is promised in later papers.
10. Discussion and conclusions

What is needed from fracture mechanics? The answer depends upon the user; here an answer is attempted for a structural (in its widest sense) and materials engineer. There seem to be two, possibly three classes of need. The early use of fracture mechanics was to predict fracture, or to determine the cause of a fracture. Another application that comes under this first class is the determination of reliability and the setting inspection intervals and standards. The second class is newer, the prediction of a material’s fracture properties so that candidates for suitable new materials can be obtained by modelling and simulation to avoid long and costly testing. Also included in this class are methods of mechanically shielding the crack tip as in the transformation toughening of ceramics. These two classes of user are primarily concerned with avoiding fracture; there is a possible third class of user who is interested in exploiting fracture. The early man was exclusively interested in this third class and there are still many specialised areas where fracture is needed such as rock fracture in mining and machining.

The first class of user will obtain his fracture data from laboratory tests and, increasingly, from data banks. In this paper because of space limitations, it has been fracture in an inert environment under quasi-static monotonic loading that has been discussed. There are of course many possible fracture modes and all should be considered in a prediction or an assessment of fracture. In the past problems have arisen because while guarding against one mode of fracture, failure has been induced in another. Because of the difficulty in ensuring that all possible modes of failure have been considered, it is believed that there will always be a place for experimental testing no matter how sophisticated modelling and simulation becomes. The first class of user is unlikely to be interested in ab initio modelling except possibly in the nanoelectronics. The scale of the application will primarily determine the fracture mechanics necessary for this class, and for the third class, of user.

There are two ways in which new materials can be developed. There are composite materials, which use new combinations of existing materials with well known properties and then there are completely new materials such as new polymers or polymer blends. For composite materials, data banks will be of high importance. For completely new materials the necessary fracture mechanics will depend upon the dominant scale. Multiphase materials dominated by interfacial segregating trace elements or impurities may need MD even if their application scale is large. Films on the micron scale can often still be treated by conventional continuum mechanics, but may need continuum deformation mechanics that have a length scale or quasi-continuum mechanics. The thinnest films at the nanoscale will need MD. In Figs. 13 and 14 an attempt has been made to give the regimes of importance based on absolute FPZ size \((d)\) and on the ratio of a characteristic dimension \((W)\) to the FPZ size \((d)\).

What will be the future fracture mechanics? LEFM is almost certainly going to be taught and used in a hundred years time. The future of EPFM is less certain. The \(J\)-integral and the CTOD displacement have a place in characterising the initiation of fracture, but more work is needed to try to simplify the effects of constraint. For propagation, there is no simple approach. Modifications to remove some of using a
non-linear elastic $J$-integral will probably be developed in the short term. However, the more realistic computational methods based on modelling the FPZ are likely to become more important and available as part of standard finite element packages. So here it is believed that computational methods will eventually take over. MD will be able to model larger and larger numbers of atoms for longer times, but Hutchinson and Evans [73] are probably right in writing that the bottom–up approach is likely to have limited application. Modelling microstructure on the microscale is likely to be more important. Here the strain gradient and quasi-continuum type methods are likely to become increasingly important.

It is said that the best professions are in law, medicine, and the funeral industry because they are always needed, but fracture mechanics should be added. While it is difficult to predict the course of fracture mechanics over the next hundred years and impossible to predict what it will look like at the beginning of the next millennium, one thing is certain: there will be fracture mechanics as long as the human race survives.

Acknowledgements

The author would like to acknowledge all in the fracture community with whom he has worked, met, corresponded, or been influenced by. Some of these people are mentioned in this article, but because of space limitation, most are not. The reviewers are thanked for their comments which have much improved this paper.

References
