APPLIED PHYSICS REVIEW

High-κ gate dielectrics: Current status and materials properties considerations

G. D. Wilk[a)]
Agere Systems, Electronic Device Research Laboratory, Murray Hill, New Jersey 07974

R. M. Wallace[b)]
University of North Texas, Department of Materials Science, Denton, Texas 76203

J. M. Anthony
University of South Florida, Center for Microelectronics Research, Tampa, Florida 33620

(Received 9 November 2000; accepted for publication 19 January 2001)

Many materials systems are currently under consideration as potential replacements for SiO_2 as the gate dielectric material for sub-0.1 μm complementary metal–oxide–semiconductor (CMOS) technology. A systematic consideration of the required properties of gate dielectrics indicates that the key guidelines for selecting an alternative gate dielectric are (a) permittivity, band gap, and band alignment to silicon, (b) thermodynamic stability, (c) film morphology, (d) interface quality, (e) compatibility with the current or expected materials to be used in processing for CMOS devices, (f) process compatibility, and (g) reliability. Many dielectrics appear favorable in some of these areas, but very few materials are promising with respect to all of these guidelines. A review of current work and literature in the area of alternate gate dielectrics is given. Based on reported results and fundamental considerations, the pseudobinary materials systems offer large flexibility and show the most promise toward successful integration into the expected processing conditions for future CMOS technologies, especially due to their tendency to form at interfaces with Si (e.g. silicates). These pseudobinary systems also thereby enable the use of other high-κ materials by serving as an interfacial high-κ layer. While work is ongoing, much research is still required, as it is clear that any material which is to replace SiO_2 as the gate dielectric faces a formidable challenge. The requirements for process integration compatibility are remarkably demanding, and any serious candidates will emerge only through continued, intensive investigation. © *2001 American Institute of Physics.* [DOI: 10.1063/1.1361065]

I. INTRODUCTION

The rapid progress of complementary metal–oxide–semiconductor (CMOS) integrated circuit technology since the late 1980's has enabled the Si-based microelectronics industry to simultaneously meet several technological requirements to fuel market expansion. These requirements in-

[a)]G. D. Wilk is formerly of Bell Laboratories, Lucent Technologies; electronic mail: gwilk@agere.com
[b)]Electronic mail: rwallace@unt.edu

clude performance (speed), low static (off-state) power, and a wide range of power supply and output voltages.[1] This has been accomplished by developing the ability to perform a calculated reduction of the dimensions of the fundamental active device in the circuit: the field effect transistor (FET)—a practice termed ''scaling.''[2–4] The result has been a dramatic expansion in technology and communications markets including the market associated with high-performance microprocessors as well as low static-power applications, such as wireless systems.[5]

It can be argued that the key element enabling the scaling of the Si-based metal–oxide–semiconductor field effect transistor (MOSFET) is the materials (and resultant electrical) properties associated with the dielectric employed to isolate the transistor gate from the Si channel in CMOS devices for decades: silicon dioxide. The use of amorphous, thermally grown SiO_2 as a gate dielectric offers several key advantages in CMOS processing including a stable (thermodynamically and electrically), high-quality $Si–SiO_2$ interface as well as superior electrical isolation properties. In modern CMOS processing, defect charge densities are on the order of $10^{10}/cm^2$, midgap interface state densities are $\sim 10^{10}/cm^2\,eV$, and hard breakdown fields of 15 MV/cm are routinely obtained and are therefore expected regardless of the device dimensions. These outstanding electrical properties clearly present a significant challenge for any alternative gate dielectric candidate.

II. SCALING AND IMPROVED PERFORMANCE

The industry's demand for greater integrated circuit functionality and performance at lower cost requires an increased circuit density, which has translated into a higher density of transistors on a wafer.[3] This rapid shrinking of the transistor feature size has forced the channel length and gate dielectric thickness to also decrease rapidly. As will be discussed in the next few sections, the current CMOS gate dielectric SiO_2 thickness can scale to at least 13 Å, but there are several critical device parameters that must be balanced during this process.

The improved performance associated with the scaling of logic device dimensions can be seen by considering a simple model for the drive current associated with a FET.[1] The drive current can be written (using the gradual channel approximation) as

$$I_D=\frac{W}{L}\mu C_{\text{inv}}\left(V_G-V_T-\frac{V_D}{2}\right)V_D, \tag{1}$$

where W is the width of the transistor channel, L is the channel length, μ is the channel carrier mobility (assumed constant here), C_{inv} is the capacitance density associated with the gate dielectric when the underlying channel is in the inverted state, V_G and V_D are the voltages applied to the transistor gate and drain, respectively, and the threshold voltage is given by V_T. It can be seen that in this approximation the (V_D/L) along the channel direction. Initially, I_D increases linearly with V_D and then eventually saturates to a maximum when $V_{D,\text{sat}}=V_G-V_T$ to yield

$$I_{D,\text{sat}}=\frac{W}{L}\mu C_{\text{inv}}\frac{(V_G-V_T)^2}{2}. \tag{2}$$

The term (V_G-V_T) is limited in range due to reliability and room temperature operation constraints, since too large a V_G would create an undesirable, high electric field across the oxide. Furthermore, V_T cannot easily be reduced below about 200 mV, because $kT\sim 25$ mV at room temperature. Typical specification temperatures ($\leqslant 100$ °C) could therefore cause statistical fluctuations in thermal energy, which would adversely affect the desired the V_T value. Thus, even in this simplified approximation, a reduction in the channel length or an increase in the gate dielectric capacitance will result in an increased $I_{D,\text{sat}}$.

In the case of increasing the gate capacitance, consider a parallel plate capacitor (ignoring quantum mechanical and depletion effects from a Si substrate and gate)[6]

$$C=\frac{\kappa\epsilon_0 A}{t}, \tag{3}$$

where κ is the dielectric constant (also referred to as the relative permittivity in this article) of the material,[7] ϵ_0 is the permittivity of free space ($=8.85\times 10^{-3}$ fF/μm), A is the area of the capacitor, and t is the thickness of the dielectric. This expression for C can be rewritten in terms of t_{eq} (i.e., equivalent oxide thickness) and κ_{ox} ($=3.9$, dielectric constant of SiO_2) of the capacitor. The term t_{eq} represents the theoretical thickness of SiO_2 that would be required to achieve the same capacitance density as the dielectric (ignoring issues such as leakage current and reliability). For example, if the capacitor dielectric is SiO_2, $t_{\text{eq}}=3.9\epsilon_0(A/C)$, and a capacitance density of $C/A=34.5$ fF/μm^2 corresponds to $t_{\text{eq}}=10$ Å. Thus, the physical thickness of an alternative dielectric employed to achieve the equivalent capacitance density of $t_{\text{eq}}=10$ Å can be obtained from the expression

$$\frac{t_{\text{eq}}}{\kappa_{\text{ox}}}=\frac{t_{\text{high}-\kappa}}{\kappa_{\text{high}-\kappa}}$$

$$\text{or simply, } t_{\text{high}-\kappa}=\frac{\kappa_{\text{high}-\kappa}}{\kappa_{\text{ox}}}t_{\text{eq}}=\frac{\kappa_{\text{high}-\kappa}}{3.9}t_{\text{eq}}. \tag{4}$$

A dielectric with a relative permittivity of 16 therefore affords a physical thickness of ~40 Å to obtain $t_{\text{eq}}=10$ Å. (As noted above, actual performance of a CMOS gate stack does not scale directly with the dielectric due to possible quantum mechanical and depletion effects.)[6]

From a CMOS circuit performance point of view, a performance metric considers the dynamic response (i.e., charging and discharging) of the transistors, associated with a specific circuit element, and the supply voltage provided to the element at a representative (clock) frequency. A common element employed to examine such switching time effects is a CMOS inverter.[1] This circuit element is shown in Fig. 1 where the input signal is attached to the gates and the output

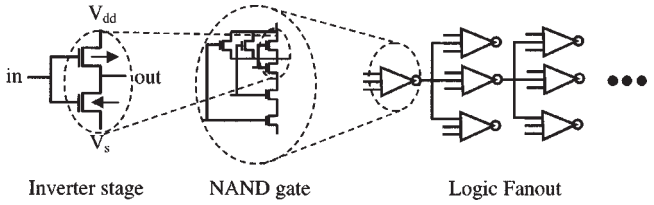


FIG. 1. Components used to test a CMOS FET technology. V_{DD} and V_S serve as the source and drain voltages, respectively, and are common to the NAND gates shown. Each NAND gate is connected to three others resulting in a fanout of 3.

stage. The switching time is limited by both the fall time required to discharge the load capacitance by the n-FET drive current and the rise time required to charge the load capacitance by the p-FET drive current. That is, the switching response times are given by[1]

$$\tau = \frac{C_{\mathrm{LOAD}} V_{DD}}{I_D}, \quad \text{where} \quad C_{\mathrm{LOAD}} = F C_{\mathrm{GATE}} + C_j + C_i, \tag{5}$$

and C_j and C_i are parasitic junction and local interconnection capacitances, respectively. The ''fan out'' for interconnected devices is given by the factor ''F.'' Ignoring delay in gate electrode response, as $\tau_{\mathrm{GATE}} \ll \tau_{n,p}$, the average switching time is therefore

$$\bar{\tau} = \frac{\tau_p + \tau_n}{2} = C_{\mathrm{LOAD}} V_{DD} \left\{ \frac{1}{I_D^n + I_D^p} \right\}. \tag{6}$$

The load capacitance in the case of a single CMOS inverter is simply the gate capacitance if one ignores parasitic contributions such as junction and interconnect capacitance. Hence, an increase in I_D is desirable to reduce switching speeds. For more realistic estimates of microprocessor performance, the load capacitance is connected (''fanned out'') to other inverter elements in a predetermined fashion. When coupled with other NMOS/PMOS transistor pairs in the configuration shown in Fig. 1, one can create a logic ''NAND'' gate which can be used to investigate the dynamic response of the transistors and thus examine their performance under such configurations. For example, in microprocessor estimates, a fan out of $F=3$ is often employed, as shown in Fig. 1.[1]

One can then characterize the performance of a circuit (based on a particular transistor structure) through this switching time. To do this, various ''figures of merit'' (FOM) have been proposed which incorporate parasitic capacitance as well as the influence of gate sheet resistance on the switching time.[8] For example, a common FOM employed is related to Eq. (6) simply by

$$\mathrm{FOM} \cong \frac{1}{\bar{\tau}} = \frac{2}{\tau_p + \tau_n}. \tag{7}$$

In the case where parasitics are ignored, it is easily seen then that an increase in the device drive current I_D results in a decrease in the switching time and an increase in the FOM value (performance). Even in this simple model, however,

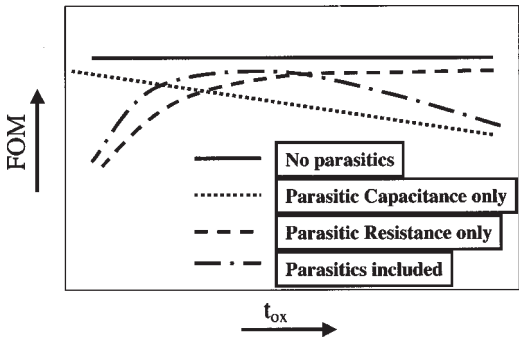


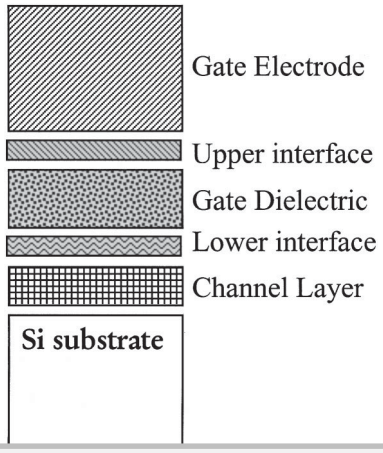
FIG. 2. FOM as a function of equivalent oxide thickness, t_{eq}. Parasitic capacitances and resistances result in transistor design tradeoffs to optimize performance.

dielectric capacitance. This can be seen in Fig. 2 where various FOM calculations are plotted as a function of an ''equivalent oxide thickness,'' t_{eq}, as described earlier.

Each FOM calculation shown in Fig. 2 corresponds to specific assumptions on the values of parasitic capacitance and gate sheet resistance, as indicated (gate length is kept constant in this analysis). Important aspects such as gate induced drain leakage and reliability are ignored in this simple model.[1] Nevertheless, the result of the FOM calculation shown in Fig. 1 indicates that tradeoffs on all aspects of the transistor design and scaling, including parasitics, must be carefully considered in order to increase the circuit performance.[8]

III. METAL–INSULATOR–SEMICONDUCTOR (MIS) GATE STACK STRUCTURES

Figure 3 provides the reader a schematic overview of the various regions associated with the gate stack of a CMOS FET (regions are separated simply to clarify the following discussion). The gate dielectric insulates the gate electrode (gate) from the Si substrate. Gate electrodes in modern CMOS technology are composed of polycrystalline Si (poly-Si) which can be highly doped (e.g. by ion implantation) and subsequently annealed in order to substantially increase conductivity. The selection of the dopant species and concentra-



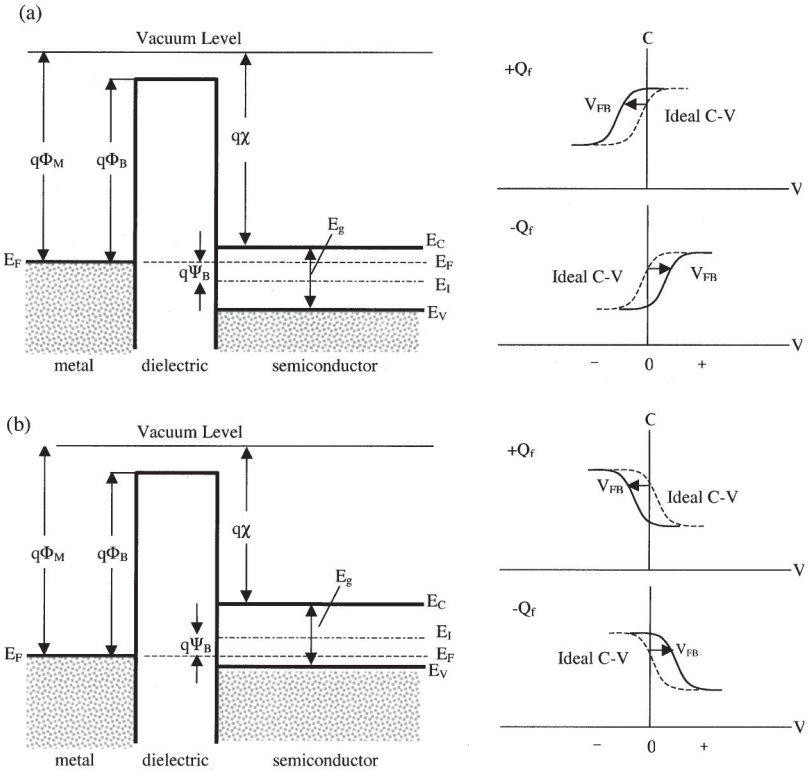


FIG. 4. Energy-band diagrams and associated high-frequency $C-V$ curves for ideal MIS diodes for (a) n-type and (b) p-type semiconductor substrates. For these ideal diodes, $V=0$ corresponds to a flatband condition. For dielectrics with positive $(+Q_f)$ or negative $(-Q_f)$ fixed charge, an applied voltage (V_{FB}) is required to obtain a flatband condition and the corresponding $C-V$ curve shifts in proportion to the fixed charge. (after Refs. 9 and 10).

tion permits the adjustment of the poly-Si Fermi level for either nMOS or pMOS FETs. Metals can also be used as the gate electrode, and, in fact, are commonly used for evaluation of capacitor structures. Work is underway to find suitable metal gates for CMOS (see Sec. VI E).

The interfaces with either the gate or the Si channel region are particularly important in regard to device performance. These regions, ~5 Å thick, serve as a transition between the atoms associated with the materials in the gate electrode, gate dielectric and Si channel. As will be discussed, these interface regions can alter the overall capacitance of the gate stack, particularly if they have a thickness which is substantial relative to the gate dielectric. Additionally, these interfacial regions can be exploited to obtain desirable properties. The upper interface, for example, can be engineered in order to block boron outdiffusion from the p^+ poly-Si gate. The lower interface, which is in direct contact with the CMOS channel region, must be engineered to permit low interface trap densities (e.g. dangling bonds) and minimize carrier scattering (maximize mobility) in order to obtain reliable, high performance.

It is instructive to consider the band diagrams for the MIS structures discussed in this review. Figure 4 shows the energy-band diagrams for ideal MIS diode structures using (a) n-type and (b) p-type semiconductor substrates.[9,10] For these ideal structures, at $V=0$ applied voltage on the metal gate, the work function difference between the metal and semiconductor, Φ_{MS}, is zero

$$\Phi_{\mathrm{MS}}=\Phi_{\mathrm{M}}-\left(\chi+\frac{E_g}{2q}+\Psi_B\right)=0; \quad p\text{-type}, \tag{8}$$

where Φ_{M} is the metal work function, χ is the semiconductor electron affinity, E_g is the semiconductor band gap, Φ_B is the potential barrier between the metal and dielectric, and ψ_B is the potential difference between the Fermi level E_F and the intrinsic Fermi level, E_I. Under these conditions, the energy bands are flat across the structure as shown in Fig. 4 and $V=V_{\mathrm{FB}}=0$, where V_{FB} is the flat band voltage (i.e., the voltage required to bring the Fermi levels into alignment). A more typical case is that the Fermi levels of the electrode and substrate are misaligned by an energy difference, and a voltage $(V_{\mathrm{FB}}\neq 0)$ must be applied to bring the Fermi levels into alignment.

Many dielectrics exhibit a fixed charge (Q_f), however, resulting in a required applied voltage $V=V_{\mathrm{FB}}\neq 0$ to achieve a flat band condition. The amount of fixed charge can be related to the measured V_{FB} value by the expression[9]

$$V_{\mathrm{FB}}=\Phi_{\mathrm{MS}}\pm Q_f/C_{\mathrm{acc}}, \tag{9}$$

where C_{acc} is the measured capacitance in accumulation. Thus, a value for fixed charge density Q_f can be determined from measured values of V_{FB}, Φ_{MS} and C_{acc}. The sign of the fixed charge is also important, as negative fixed charge correlates with the plus sign in Eq. (9), and positive fixed charge correlates with the minus sign. These expressions will be discussed further in Sec. V C 2.

The source of such fixed charge, often though *not always*

semiconductor interface. Several proposed explanations for the cause of the observed fixed charge will be discussed in Sec. V C 2. Figure 4 shows that for positive Q_f, a negative shift in the V_{FB} from ideal conditions (where $V=0$) is required for both *n*-type and *p*-type MIS structures. Similarly, a positive V_{FB} is required for negative Q_f.

Most of the alternate dielectric candidates examined to date appear to have a substantial amount of fixed charge, which could present significant issues for CMOS applications. Given the scaling limitations on applied voltages due to power consumption, shifts in the V_{FB} value are undesirable and must be minimized. In some applications, biasing the substrate to compensate for the fixed charge has been proposed.[5] Moreover, a reproducible V_{FB} (correspondingly V_T for transistors) value is also required for stable, reliable transistor operation. Thus, hysteretic changes in the V_{FB} from voltage cycling of less than 20 mV are often required.

Some dielectrics which incorporate aluminum, however, thus far suggest that a negative fixed charge is present. It has been recently proposed to combine Al ions with some alternate dielectric candidates in order to compensate positive and negative fixed charges to achieve a neutral state or, at least, minimize such fixed charge effects.[11] If fixed charge is determined to be large and difficult to minimize and control in high-κ dielectrics, it will be a significant issue for obtaining the desired device performance on both *n*MOS and *p*MOS transistors. The magnitude of measured V_{FB} shifts for many alternate dielectrics will be discussed later.

IV. SCALING LIMITS FOR CURRENT GATE DIELECTRICS

The previous sections outlined the need to scale oxide thicknesses to improve performance. The next two sections describe the present understanding in the field regarding the limits of scaling current gate dielectric materials, SiO_2 and Si-oxide-nitride variations, for CMOS. Issues include band offset, interfacial structure, boron penetration and reliability. Beyond this scaling limit, another material will be required as the gate dielectric to allow further CMOS scaling.

A. Ultrathin SiO_2 properties

Experiments and modeling have been done on ultrathin SiO_2 films on Si, as a way to determine how the SiO_2 band gap or band offsets to Si change with decreasing film thickness.[12–15] In the study by Muller *et al.*,[12] electron energy loss spectroscopy (EELS) was carried out on 7–15 Å SiO_2 layers on Si. It was found that the density of states (as measured by the oxygen *K*-edge in EELS, with a probe resolution <2 Å) transition from the substrate into the SiO_2 layer indicated that the full band gap of SiO_2 is obtained after only about two monolayers of SiO_2. This indicates that within two monolayers of the Si channel interface, oxygen atoms do not have the full arrangement of oxygen neighbors and therefore

An earlier *ab inito* model by Tang *et al.*[13] of extremely thin SiO_2, which was modeled as a modified beta-cristoballite phase, showed an important result, in that the band gap of SiO_2 did not begin to decrease until there were fewer than three monolayers of oxide. Moreover, estimates of the changes in the associated conduction and valance band offsets for these systems indicated that a minimum of 7 Å of SiO_2 is required to obtain bulk properties. The recent first principles study by Neaton *et al.*[14] determined that the local energy gap in SiO_2 is directly related to the number of O second nearest neighbors, for a given O atom. The *last row* of O atoms (next to the Si substrate) *by definition* cannot have the full six nearest neighbor O atoms. The second row of O atoms from the Si interface is thus the first layer of O atoms that have the required six second-nearest neighbor O atoms. The distance required to obtain the full band gap of SiO_2 at each interface is therefore given by 1.6 Å (the spacing of one Si–O bond length) +2.4 Å (the distance between neighboring O atoms is 2.7 Å, but this is variable because of Si–O bond bending. The distance is typically in the range ~2 to 2.4 Å). The thickness at each interface required for the full SiO_2 band gap is therefore ~3.5–4.0 Å. Counting both interfaces, the total thickness of 7–8 Å is required, in agreement with Tang *et al.*[13] and with the experiment.[12] These results set an *absolute* physical thickness limit of SiO_2 of 7 Å. Below this thickness, the Si-rich interfacial regions from the channel and polycrystalline Si gate interfaces used in MOSFETs overlap, causing an effective ''short'' through the dielectric, rendering it useless as an insulator.

The agreement between the experiment and simulation in these cases indicates that the inherent band gap of SiO_2 remains intact, even down to only a few monolayers of material. Other important properties of SiO_2 have been reported in the ultrathin, sub-20 Å regime, such as the conduction band offset ΔE_C to Si [using x-ray photoelectron spectroscopy (XPS)],[16] the tunneling electron effective mass m^* (from tunneling $I-V$ measurements),[17] and the photoelectron attenuation length.[18] These measurements have further demonstrated very little change in fundamental SiO_2 properties between bulk and ultrathin sub-20 Å films.

The apparent robust nature of SiO_2, coupled with industry's acquired knowledge of oxide process control, has helped the continued use of SiO_2 for the past several decades in CMOS technology. As experimental evidence of the excellent electrical properties of such ultrathin SiO_2 films, it has been demonstrated that transistors with gate oxides as thin as 13–15 Å continue to operate satisfactorily.[19–24] Although high leakage current densities of 1–10 A/cm^2 (at V_{DD}) are measured for such devices,[25] transistors intended for high-performance microprocessor applications can sustain these currents. As first reported by Timp *et al.*[20–22] scaling of CMOS structures with SiO_2 gate oxides thinner than about 10–12 Å results in no further gains in transistor drive current. This result has been subsequently and independently

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.