Chapter 2

A Fast-Transient Measurement System and Techniques

2.1 Necessity of Transient Measurement

Collection of reliable and complete information is essential for any subjects in scientific research. When it comes to the study of electrical reliability in semiconductor CMOS and non-volatile memory (NVM) devices, one usually concerns the variation of certain parameters (e.g. threshold voltage shift ΔV_t, drain current degradation ΔI_d, transconductance variation ΔG_m, …, etc.) as a function of stress time for voltages and/or temperatures of interest. As shown in Fig. 2.1(a), a widely-adopted methodology for evaluation of NVM endurance is to periodically interrupt the program/erase (P/E) operations and monitor the change in threshold voltage. The parameter analyzer (e.g. Agilent 4156) is an off-the-shelf instrument capable of providing measurement of reliable experimental data. However, the mechanical switch (e.g. Agilent 5250) used for transition from P/E operation to V_t measurement introduces a delay time up to seconds during switching. Similarly in Fig. 2.1(b), although a mechanical switch is not required in CMOS device characterization, the parameter analyzer still consumes time to change from settings for stress to settings for measurement. The missing information during the switching transient is illustrated in Fig. 2.2. The application of a positive gate voltage (V_g), for programming a memory cell or for stressing a transistor, injects electrons from the inversion channel toward the gate dielectric. Concomitant charge trapping into the dielectric traps

increases device V_t as indicated in Fig. 2.2(a). Upon relieving the stress/program V_g, the trapped electrons start to detrapp before the semiconductor analyzer completes measurement setting for V_t extraction, giving rise to underestimation of ΔV_t (Fig. 2.2(b)). The underestimation will be demonstrated to be significant in Section 2.3. In addition to the aforementioned delay, another drawback of the widely-used method is the way a parameter analyzer measures a signal (current/voltage): integrating over a period of time and averaging. The shortest rule-of-thumb integration time is on the order of 10ms. That is to say, the analyzer cannot capture any events with response time less than 10ms. In brief, a methodology with fast response and minimized switching delay is desired.

2.2 Existing Techniques

To retrieve the information ignored resulting from the slow response of the instruments, efforts have been made. Among them are usages of radio-frequency (RF) kits, of series resistance, and of operational amplifiers along with electronic switches.

2.2.1 RF Kits

The most aggressive method is to replace all the DC parameter analyzer and components (as shown in Fig. 2.1) with those for high-frequency applications as depicted in Fig. 2.3 [2.1][2.2]: a) transmission lines of fixed length and bias tees for minimization of signal delay and power loss through commonly-adopted bi-axial cables and adapters, b) pulse generators for short-period (~μs or even ~ns) input signals, c) a digital oscilloscope for

real-time output display, and d) special test patterns for RF characterizations. Success in extracting inversion charge density free of self-heating in SOI devices [2.1] as well as of charge-trapping in high-k devices [2.2] (for evaluating drain current and mobility) have been reported with the use of the system. Nonetheless, high cost of the RF kits and the effort for designing special test patterns hinder the method from being popular.

2.2.2 Series Resistance

Recently, a more economical characterization method has been proposed to investigate V_t-instability in high-k CMOSFETs [2.3]. Shown in Fig. 2.4 is the circuit developed by Kerber et al. A trapezoidal pulse is applied to the gate, and the voltage drop across the resistance R in series with the drain is measured to compute the drain current. However, two problems arise: a) *parasitic capacitance* [2.4][2.5] and b) *flexibility* [2.5]. a) The drain voltage is not constant due to changing drain current with changing V_g. The gate-to-drain capacitance (C_{gd}) and parasitic capacitance (C_p) need to be charged or discharged in response to varying V_g/V_d. C_p consists of the drain-to-bulk capacitance (C_{db}), the cable capacitance, and the input capacitance of the oscilloscope. The charging/discharging current distorts the measured drain current. The distortion is worsened with decreasing rise/fall time of the gate pulse. b) The setup lacks flexibility due to a fixed voltage at the resistor end (100mV in Fig. 2.4). To facilitate periodic V_t (or I_d) measurement during stress or during P/E cycles, voltages at the gate, drain, and source need to be changeable in accordance with the operations[1,2].

[1] In most cases, voltages for stress or P/E operation are higher than those for measurement.

2.2.3 Operational Amplifiers

To overcome the problems in Section 2.2.2, two resembling techniques both using operational amplifiers have been independently developed [2.4][2.5]. The setup proposed by Shen et al. takes the advantage of the virtual ground property of operational amplifiers to simultaneously hold the drain voltage constant while sensing the drain current, as indicated in Fig. 2.5 [2.4]. Thus, no charging/discharging current flows through C_p. To minimize the effect of C_{gd}, use of short channel length devices is suggested. Another technique is independently developed in this work and will be described in details in the following section.

2.3 A Computer-Automated Measurement System

2.3.1 The System

The setup developed in this work, in addition to an operational amplifier, incorporates electronic switches to maximize flexibility. The measurement setup shown in Fig. 2.6(a) consists of an operational amplifier (OPA655[3]), a digital oscilloscope[4], and high-speed analog switches (MAX333[5]). Computer programs are utilized to integrate the components into a system and to facilitate automatic control. Voltage sources for stress (or P/E cycle) and

---

For example, in NBTI experiments for pMOSFETs a typical stress condition is V_g= -2V with other terminals grounded, while the measurement bias is V_g swept from 0V to -1.2V and V_d= -0.1V.

[2] Complicated operation schemes may be required in flash memories. For example, in floating gate flash memories, a substrate bias is applied during programming in CHISEL operating scheme [2.6], or a pulse train is applied at the source in PASHEI [2.7]; in two-bit SONOS-type flash memories where stored charges are physically isolated in the nitride layer, the role of drain and source are interchanged in reverse-read scheme for NROM [2.8] or PHINES [2.9].

[3] The bandwidth for 1MΩ transimpedance is 1MHz.

[4] Tektronix Model TDS 5054.

[5] Quad single-pole-double-throw (SPDT) CMOS analog switches. Typical turn-on, turn-off, and break-before-make times are 50ns, 460ns, and 200ns respectively.

measurement can be independently controlled at all nodes[6]. The feedback resistance R is chosen to be 100KΩ. The waveforms are illustrated in Fig. 2.6(b). During stress phase, pre-set stress voltages are applied to the gate ($V_{g,stress}$) and the drain ($V_{d,stress}$) with all other nodes grounded. During measurement phase, voltages for measuring V_t or I_d are now applied to the gate ($V_{g,meas}$, usually much smaller than $V_{g,stress}$) and the drain ($V_{d,meas}$). For example, to perform PBTI experiments in high-k nMOSFETs, one chooses $V_{g,stress}$=2V, $V_{d,stress}$=0V, $V_{g,meas}$=1.2V, and $V_{d,meas}$=0.1V. The image of the measurement circuit is shown in Fig. 2.7.

2.3.2 System Capability

In addition to the parasitic delay from the cables and the probe station, the measurement and switching speed respectively limited by the frequency response of the operational amplifier and the electronic switches determine the capability of the system. Without optimization of the parasitics, the system is capable of capturing reliable signals 50μs right after stress voltages are relieved with a current resolution of ~5nA[7]. Fig. 2.8 shows the result of post-stress I_d recovery in a high-k nMOSFET. Electrons are injected into the high-k traps during stress, and right after the stress V_g is removed the trapped charges begin to escape. Therefore I_d increases. Detailed physics will be discussed in Chapter 3 through Chapter 5. As can be clearly seen from the figure, I_d change is significantly underestimated in a conventional measurement method where a long delay time is present.

---

[6] Although grounded in Fig. 2.6, it should be undoubted that a high-speed analog switch can be added to the substrate, too.

[7] Both the time response and the current resolution are strongly affected by the measurement environment, e.g. type of probe station, length and type of cables, etc.

In NVM characterization, in addition to retrieval of the missing transient information, another advantage is manifested in Fig. 2.9. Retention of a memory cell concerns the robustness of its V_t value against storage time. For a memory at program state, less charge loss during retention gives smaller V_t drop and thus longer time to keep correct data. Retention is evaluated by observing the V_t shift after programming charges into the memory cell. To obtain a reliable trend (V_t vs. time) for 10-year extrapolation or for understanding the underlying physics, one often needs to observe over a 4-decade period of time. For a conventional method, due to the delay time for mechanically switching from programming to retention, trustworthy data reads after 1s. Inevitably, a 4-decade period of time sums up to 10000s – time consuming for incomplete information. Because the proposed transient measurement system collects V_t data from 50μs after program, the efficiency and efficacy can both be greatly improved. Fig. 2.9 reads V_t from 50μs to 2000s. An 8-decade retention time is easily achieved with one-fifth the time a conventional method may consume. More importantly, V_t is almost unchanged until around milli-seconds. The conventional method with a long delay time will not be able to observe this corner feature.

2.4 Summary

In this chapter, the components, operation configurations, and capability of a fast-transient measurement system are described in details. The system features low-cost, high-efficiency (as well as efficacy), and flexibility. Acquisition of the experimental data for the rest parts of this thesis is completed using the system: single charge emission in Chapter 3, the

two-stage high-k PBTI degradation in Chapter 4, and the anomalous NBTI in high-k pMOS devices in Chapter 5.

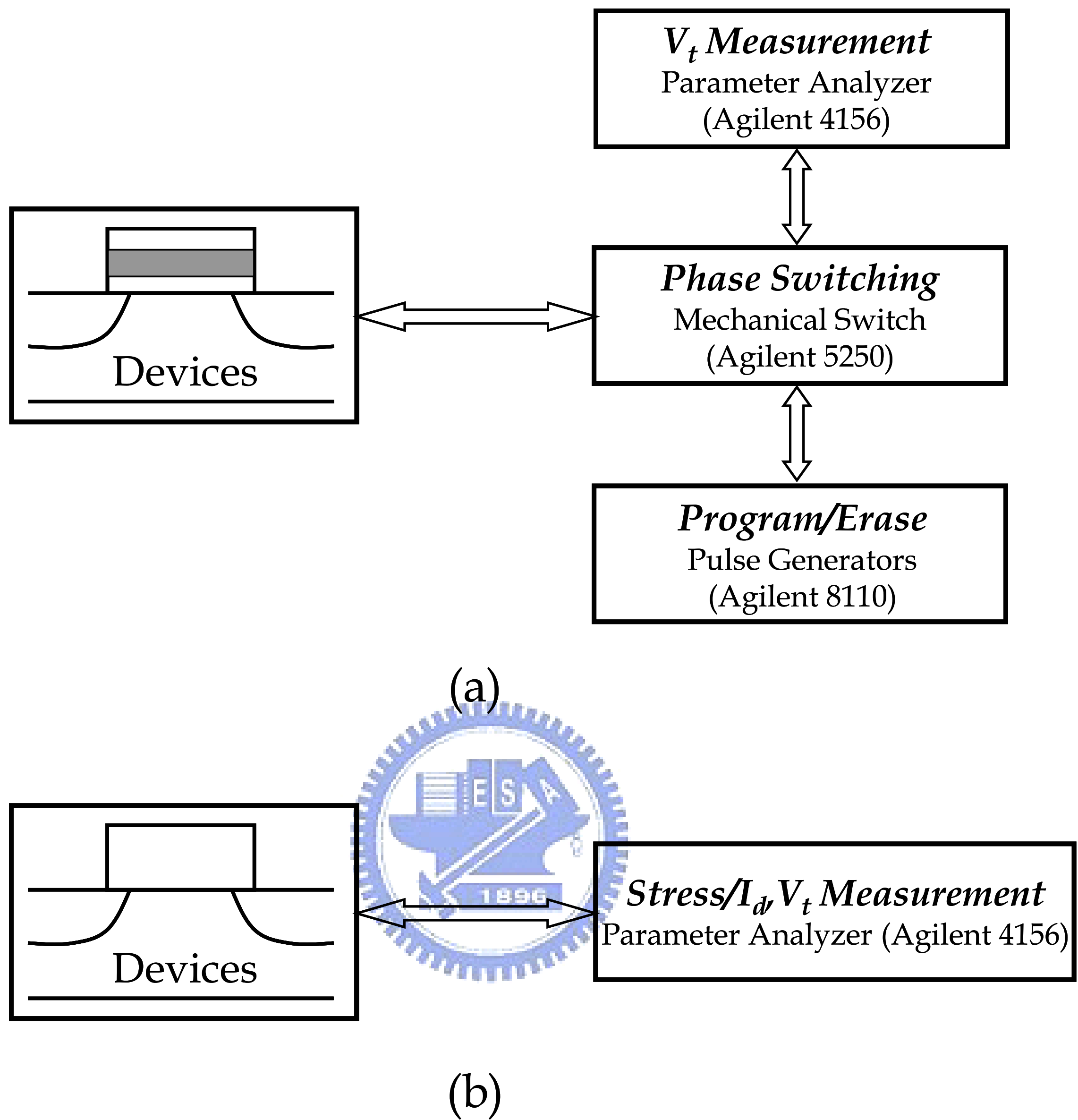
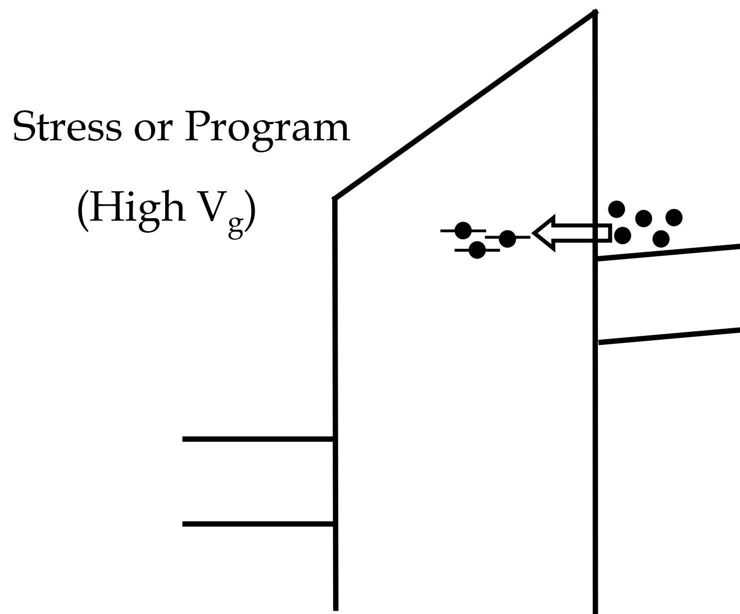
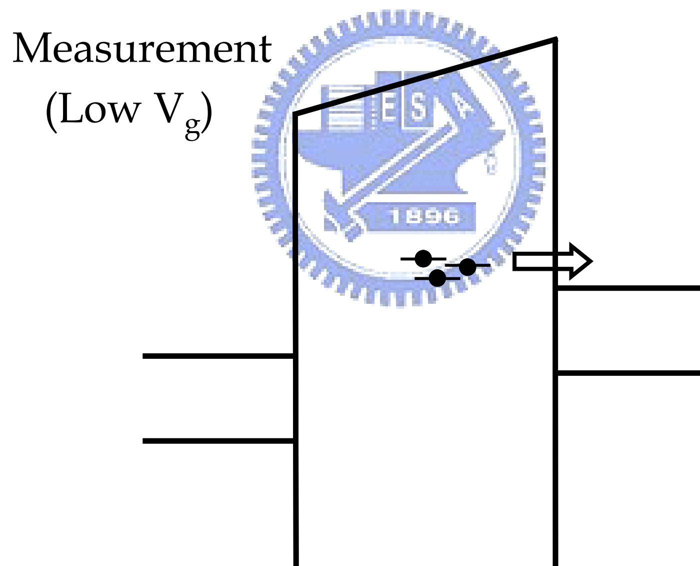


Fig. 2.1 The operating principles of a conventional measurement method in characterization of non-volatile memories (a) and of CMOS devices (b).



(a)



(b)

Fig. 2.2 Energy band diagrams during stressing or programming (a) and during measurement (b). Charge de-trapping may take place during measurement.

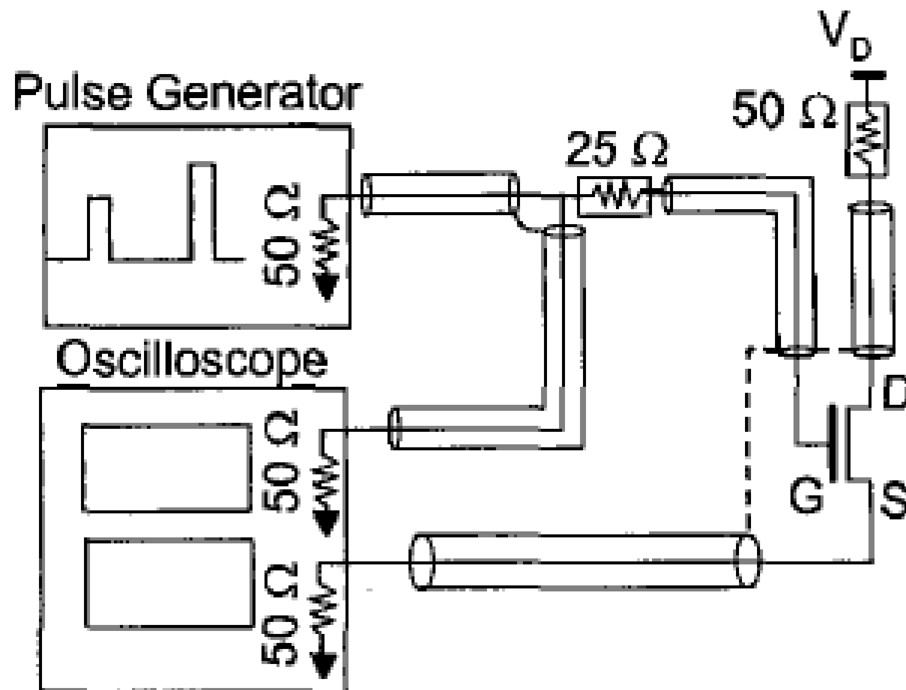


Fig. 2.3 The measurement setup including components with radio frequency response for self-heating free characterization in SOI devices proposed in [2.1].

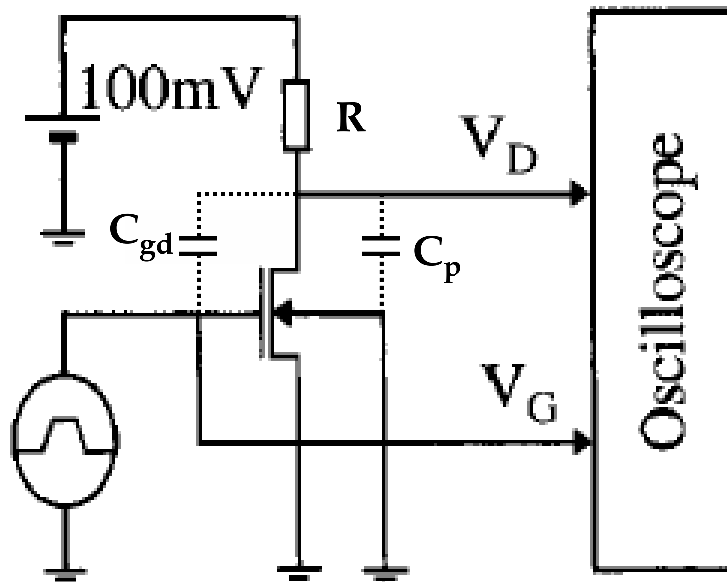
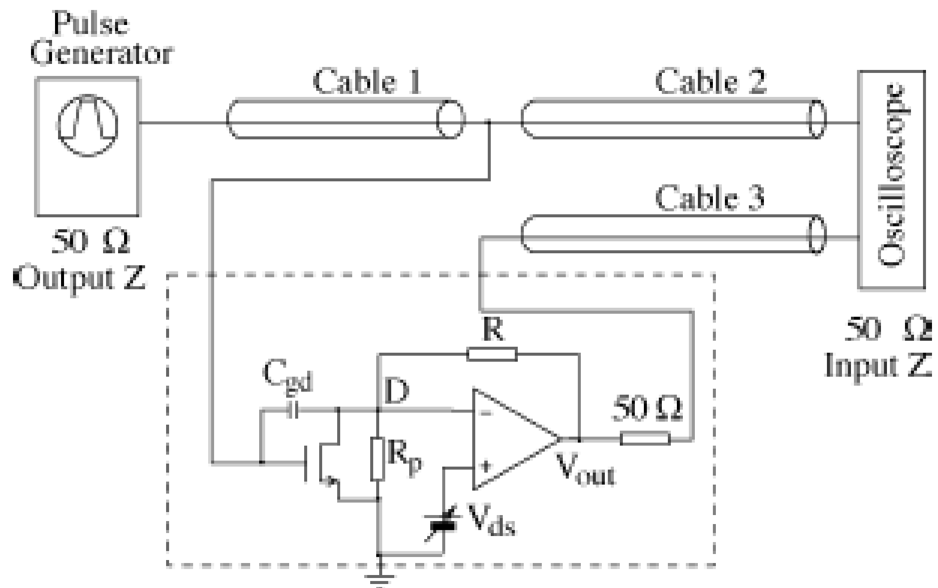


Fig. 2.4 The pulse-IV measurement setup incorporating a series resistance proposed in [2.3].



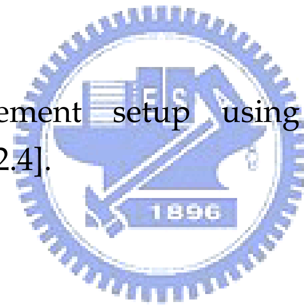
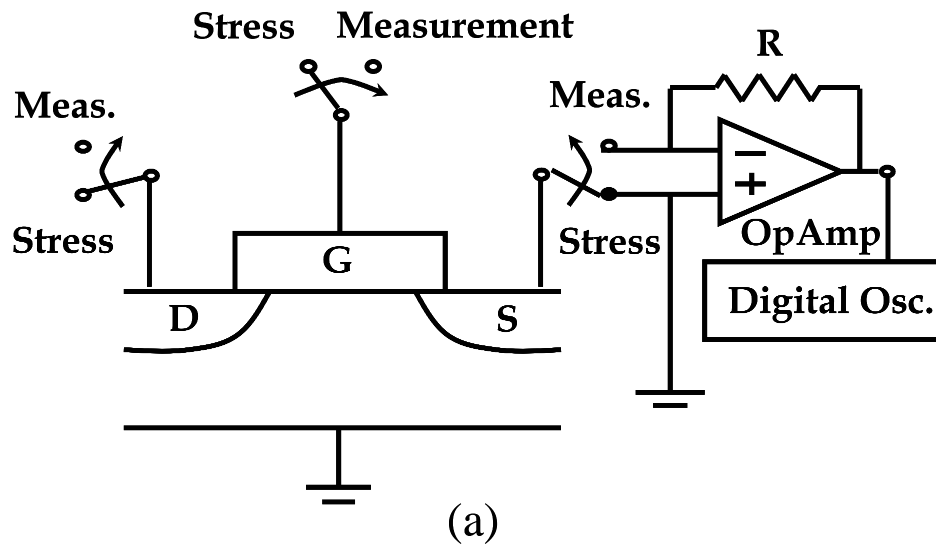


Fig. 2.5 The measurement setup using an operational amplifier proposed in [2.4].



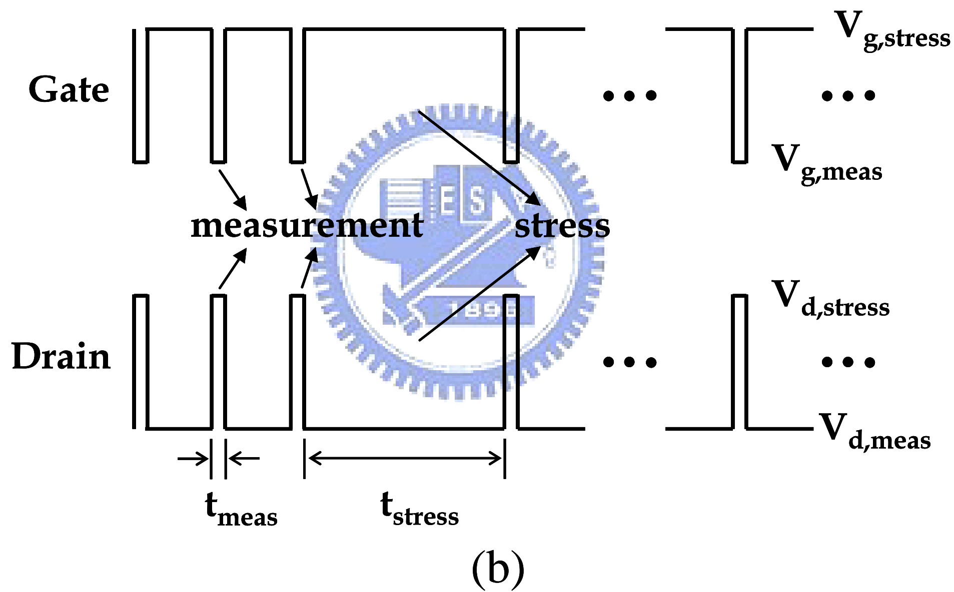


Fig. 2.6 (a) This work, in addition to the operational amplifier, takes advantage of high-speed analog switches to minimize the switching delay down to μs. (b) The waveforms for transient measurement.

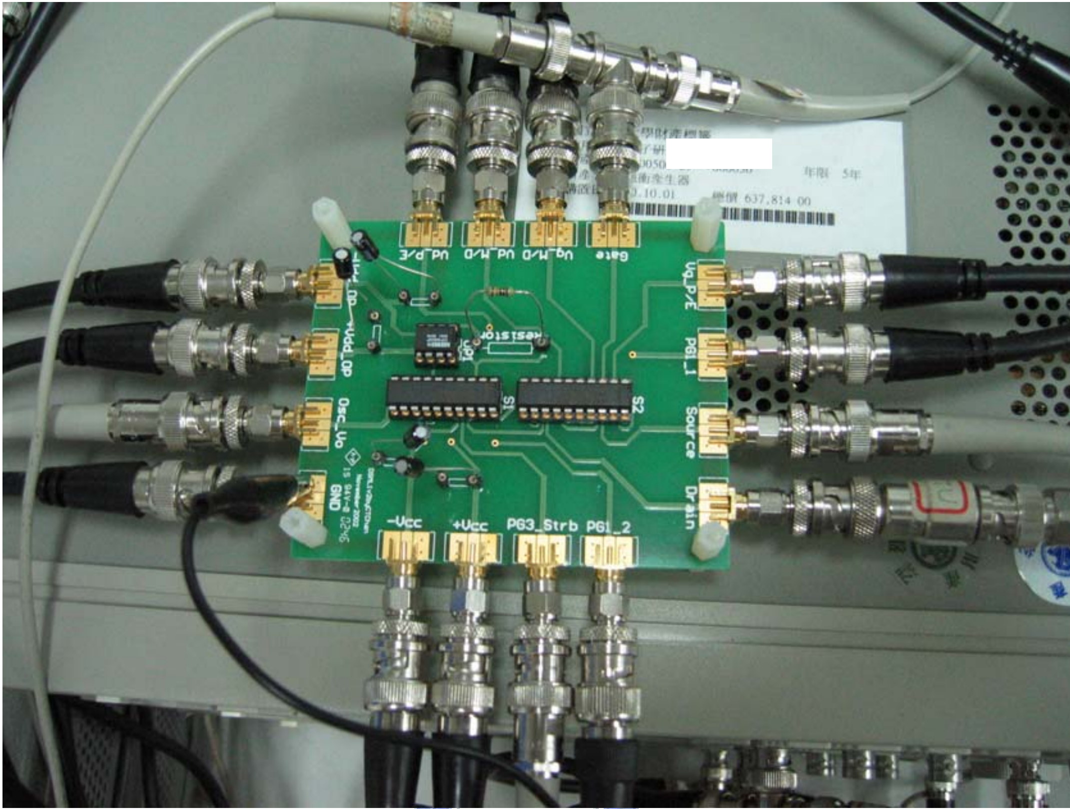


Fig. 2.7 The image of the transient measurement circuit.

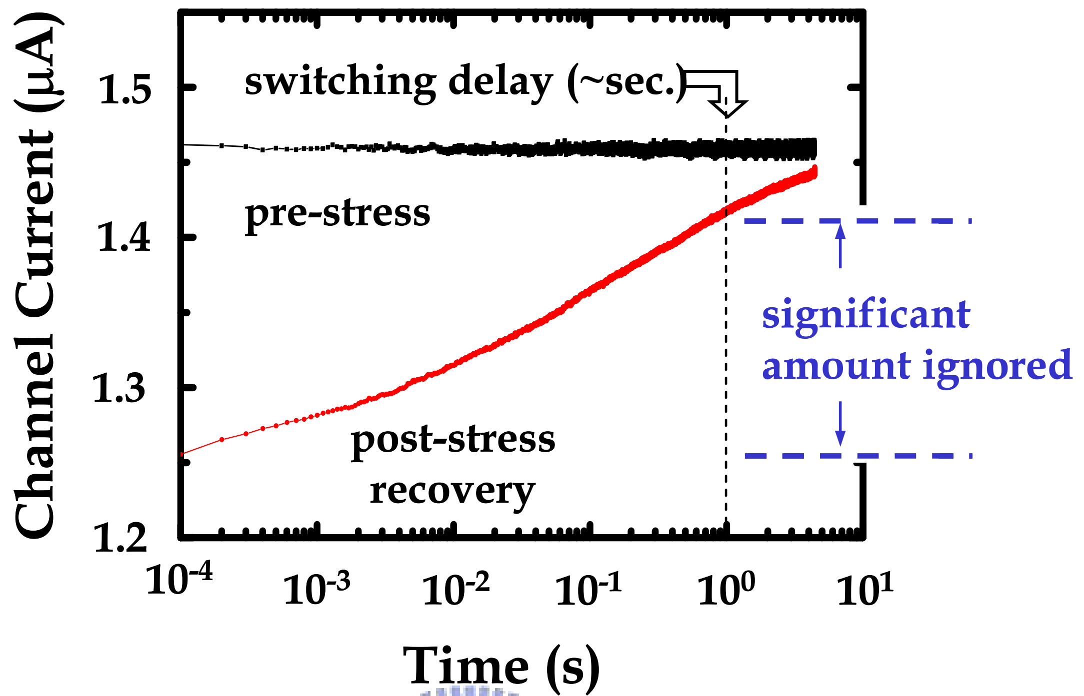


Fig. 2.8 Recovery transient in a high-k nMOS device. As clearly shown, significant charge de-trapping occurs in sub-second regime. Detailed physics will be discussed in the following chapters.

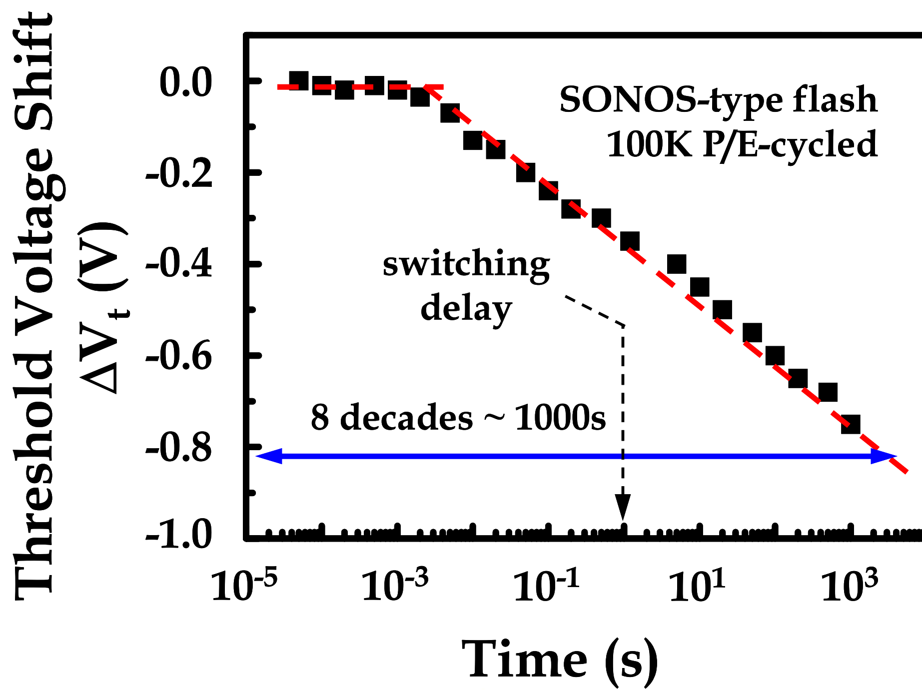


Fig. 2.9 Program state retention characteristics spanning eight decades of time.