RC24391 (99867) October 29, 2007
Physics

IBM Research Report

Component Variability as a Limit in Digital Electronics

Robert W. Keyes
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Component variability as a limit in digital electronics

Robert W. Keyes
IBM Research Division

Abstract

Commercially successful electronic computers have been built with vacuum tubes and with transistors as active devices. Attempts to build computers with circuits based on other kinds of solid state devices have failed in spite of being the focus of large, well-funded efforts. The tunnel diode was invented shortly after the transistor and was the earliest of these. Josephson junctions and resonant tunneling devices also attracted massive investments as possible but unsuccessful alternatives to transistor based logic.
.
What happened? Why did large development efforts devoted to these novel technologies fail? The answer is found in their inability to deal with the variability inevitably found in nominally identical parts. Transistors and vacuum tubes act as switches that form logic signals from standards that are distributed and recognizable throughout a system, signals that do not depend on the switching device that produced them. High gain is needed to emulate switches and is only obtained by using the attractive force between positive and negative charges as a gate to control current.

Introduction

The discovery of transistor action in 1946 inaugurated a revolution in electronics.[1,2] Transistors quickly replaced vacuum tubes in most applications that used electronic amplification. The promise of low power demand, small size, and high reliability compared with vacuum tubes had particular impact on digital computers, where large numbers of devices must be combined into a single compact system, and transistors soon replaced tubes within the computer industry. The subsequent invention of the integrated circuit drastically reduced the cost of manufacturing and interconnecting large numbers of transistors and opened the path to ever-more affordable and powerful computing systems, creating an industry that continues to thrive and grow today.

A decade after the advent of the transistor another semiconductor device, the tunnel diode, surprised the world [3] and attracted comparable interest. The ability of the tunnel diode to carry current at very high density compelled attention. Tiny tunnel diodes had a small capacitance and exhibited a negative resistance that allowed circuits with two stable states to be constructed. The belief that very fast switching between the two stable states enabled by the high current density equaled very fast digital circuits attracted sizable efforts to build digital systems around the tunnel diode.[4] However, in spite of intensive efforts devoted to the tunnel diode it failed to find a place in switching circuits for digital logic.

What happened? An electronic computing industry was initially built around vacuum tubes, but transistors have dominated electronics since the late 1950s. Why did one semiconductor device succeed so spectacularly as a way to replace the vacuum tube in digital logic while another quickly failed in the same application? In what way did the tunnel diode differ from the relays, vacuum tubes and transistors that have been successfully used to build large electrical computing systems? Experimental scientists and engineers continue attempts to replace transistors with other solid state devices without effect. Here we seek the element that is missing from the various devices that failed as the basis for fast logic circuits but allowed large powerful digital electronic computers to be built with relays, vacuum tubes, and transistors.

Devices

The first microprocessor chip, a simple computer, contained 2000 transistors, a number that has rapidly increased with the passage of time at a rate sanctified as "Moore's Law" until today microprocessor chips contain millions of transistors. The implied low cost per device is achieved by making transistors and other components on large wafers that can contain more than a billion devices and that are divided into a few hundred chips.   The wafer manufacturing processes cannot control the exact locations of material defects, impurities, and dopants all over a large wafer. It is hoped that their effects will be determined by a uniform average, but as miniaturization progresses, averages involve fewer and fewer defects and dopant atoms, and non-uniformity grows.[5]

More significant is that perfect homogeneity of all parameters during processing of a large wafer is unachievable. Some uncertainty in temperature, in reagent mixing, in illumination, in almost any physical quantity all over a wafer is unavoidable. Nonuniformity of processing parameters during the long series of operations means that there is a significant amount of device-to-device variability in the finished product.

Further, devices change with age and use. The irreversible phenomena known by such names as creep, diffusion, bleaching, corrosion, electromigration, and thermomigration cannot be avoided and change the properties of device structures during use. A system must tolerate such changes in devices.

In addition to variability in the physical structure of devices, they are operated under varying conditions. Temperature is probably the most important environmental variable. Practically all physical properties of materials depend on temperature, and such things as Fermi levels, charge carrier mobility, and energy gaps are very directly reflected in device characteristics in the presence of the temperature gradients that may be found in large systems..

This litany of undesired effects means that circuitry must deal with a significant amount of uncertainty in the properties of its components.[6-8]

Logic for Computers

Computers are built from individual logic circuits, or gates, that execute very elementary functions. A computer contains a great many such logic elements and their interconnections, and much supporting equipment in the form of power supplies, memory, and interfaces to other devices. It is a large, complex system that can attack a variety of tasks.

Great depth in the handling of information may be involved, meaning that the result of an operation is used in a succeeding operation, the result is used again, and so on, through thousands of steps. For example, the outputs of a body of circuitry are recycled to become inputs many times in simulating the evolution of a system through time or in integrating a differential equation. Information must not be allowed to deteriorate through the cumulative effects of small errors during long series of operations. The remedy for the both the accumulation of errors from step to step and the device-to-device variability is found in digital representation of information. A digit can be restored to its nominal value at each step, thereby preventing the propagation of error and erasing memory of the device that produced it. The use of information in digital form is not new, of course, counting is a manifestation of digital information and the abacus is an application to computing.   Binary information is universally used in computers because of the simplicity of on-off devices. Practically arbitrary accuracy is attainable by using enough digits. Electrical potentials distributed throughout a system supply standard digital values in electrical computers. Even if the representation of a digit is not perfect, if it can be recognized it can be restored to its intended value.

An information processing system must have a way to send digits from one component to another. Wires carry information from place to place, and components must be able to transmit signals over the wires. Logic gates must be able to receive inputs from several sources (fan-in) and the output of a logic gate must be able to provide fan-out, the furnishing of inputs to a multiplicity of terminals. Also, the outputs must be isolated from the inputs so that the results of an operation are determined by the input signals and are not influenced by following stages. No information as to whether the result of a logic operation is a 0 or a 1 should be reflected back to the inputs.

A complete set of logic functions must include inversion, that is, the conversion of a one to a zero and vice versa. All Boolean operations can be implemented with combinations of NOR gates.

Transistor logic

Figure 1 shows how digital circuitry functions in the presence of variability and noisy signals.[9] The circuit performs the NOR function with field-effect transistors. The threshold voltages of the FETs that are turned on and off may be quite variable, the input signal swings must be large enough to encompass them.   In this example a positive gate voltage turns the n transistors on and the p transistors off while a zero or ground potential on the gate turns the n transistors off and the p transistors on. Let a positive input represent a ONE and ground potential a ZERO. Then if either or both of the inputs are ONE the p transistor is not conductive and the outputs are connected to ground through the n transistors, the outputs are ZERO. Only if both inputs are ZERO is the output connected to the positive voltage source and represents a ONE.

Figure 2 shows the response of the circuit in Figure 1 to an input and how it resembles the action of a mechanical relay, an ideal device for establishing an electrical connection.

The relay has only to close a circuit, details of it's construction are irrelevant. An input can vary significantly from one of its intended digital values and still produce an output obtained from a power supply that is available everywhere within the computer. This is what binary digital means: there are two standard signal levels in the system, one of which will be recognized as ZERO and the other as ONE. Even if the input signal has been degraded by 10% or 20% the output is restored to its intended value.

A small change in input voltage in the response shown in Figure 2 produces a large change in output voltage, a property known as gain. High gain that can turn connections on and off, emulating the switching action of a relay, is essential to maintenance of the digital values of signals. This is why the one constant through forty years of information processing with solid state electronics has been the transistor. The successful predecessor of the transistor in electronic processing of information, the vacuum tube, provided the same high gain. The dotted lines in figure 2 show that the intended switching action may occur over a range of threshold values; considerable variability of the transistor thresholds and other parameters can be tolerated.

The three terminals of transistors allow the required isolation of input from output. Charge can be amplified to provide fan-out by allowing current to flow through a transistor for a long enough time to charge or discharge a large number of following stages. The circuit of Fig. 1 shows how fan-in is accepted. The circuit can switch in either direction in comparable amounts of time, no separate resetting operation is needed. And the inversion that is necessary for a complete logic system is available.

Negative resistance

Negative resistance devices offer a way to form an electrical circuit that exhibits two stable states. Figure 3 shows the circuit, which consists of the negative resistance diode in series with a resistor connected to a source of power, and also includes an example of the quite variable current-voltage characteristic of the diode. The use of such negative resistance bistability to represent the zeros and ones of binary information was proposed shortly after the invention of the tunnel diode and the introduction of transistors into computers. Early experiments used the recently discovered tunnel diode as the negative resistance element in bistable circuits.[4] The voltage across the tunnel diode plus that across the resistance must sum to the applied voltage, *V*, a condition that is satisfied at the intersections of the dashed "load line" with the diode characteristic. A and C are stable states of the circuit, state B is unstable. To perform a logic operation the circuit of Fig. 3 is initially placed in state A. A pulse of current applied to input X adds to the current at A, increasing the current that must be supported by the diode to a value greater than the peak current and driving the circuit to C, the other stable state. The output, a pulse of current that depends on the diode characteristics, indicating that the circuit has switched, is passed on to the next logic stage and the diode is left in the state C. The output current pulse is shaped by the diode characteristics and will be quite variable rather than digital. An OR function can be formed with fan-in to terminal X. In principle, choosing current inputs to X such that one input would not cause the diode current to exceed the peak but two would allow a two-input AND to be realized.

If the switching current injected at X is smaller than the resulting change of output current the circuit may be said to have current gain. Obtaining gain in this way, as the difference between peak current and peak current plus input, depends on accurate knowledge of the peak. An attempt to make suitable devices during the waning of interest in tunnel diodes for logic found a variability of 2:3 between diodes in the peak currents illustrated in Fig. 3.[10]

Other difficulties accompany the use of this kind of two-terminal bistability. The input is not well-isolated from the output; the change in voltage upon switching can propagate back through input X. Circuits using additional components to improve isolation were invented. Transformers were sometimes introduced to produce gain. There is no inversion operation; the circuit cannot be switched in the opposite direction, from C to A, a separate operation that resets the circuit to A must be provided. Nevertheless, interest in the application of the tunnel diode in computer logic continued well into the 1960's.

Other negative resistance devices can be used in the same way to represent digital information. Recent attention has focused on resonant tunneling devices (RTD).[11] An RTD can replace the tunnel diode in the circuit of Fig. 3 but will encounter the same limitations. In spite of much effort, no negative resistance device has had any impact on computer logic. All suffer from similar problems: inability to tolerate variability of the devices, low gain, lack of input-output isolation, and difficult resetting.

Josephson Computers

Many laboratories attempted to replace transistors in electronic computers with superconducting circuits and Josephson devices around 1980, [12-14] efforts that failed in spite of massive investments. The motivation for interest in Josephson devices was the low voltage involved as compared to semiconductor devices. The currents can be comparable to those known in other technologies, permitting fast switching of small capacitive charges. As with negative resistance circuits, Josephson circuits did not provide standard values for digital signals. Circuits divided a two-dimensional current space into two regions and switching meant moving the state of a circuit from one region to another.[14] Because device variability appeared as uncertainty in the locations of the boundaries between regions, just as in the case of negative resistance devices, large signals were needed to be sure of switching and the outgoing signal depended on the devices that produced it. Circuits suffered from difficulties similar to those that plagued those based on negative resistance devices. Interest in Josephson logic declined rapidly after 1983.

Gain

Both positive and negative charges, electrons and holes, participate in transistor action. The first transistors were bipolar transistors. Charge carriers of one type, holes in an npn transistor, are held in the base by potential barriers where they attract carriers of opposite sign, electrons, that then move through the base from emitter to collector. The current is controlled by the number of holes inserted into the base through a base contact. The field-effect transistor was subsequently developed and gradually became the most common type used in computers. In a field-effect transistor a positive charge placed on a gate electrode attracts electrons to a surface separated from the gate by a thin insulating layer in an *n*-type FET. A current of electrons passing from a source electrode to a drain can be controlled by the amount of positive charge on the gate.

The same physical essence, the attraction between positive and negative charges that makes transistor action possible is the basis of the gain of vacuum tubes, in which voltage applied to a wire grid electrode controls the motion of electrons from an electron-emitting cathode through the grid to a positively charged plate. This attraction between positive and negative charges has been has been the only physical phenomenon supporting electronic amplification since DeForest's invention of the vacuum triode inaugurated the age of electronics a century ago.

Summary

A half-century of electronic computing has seen many fruitless attempts to find novel device technologies for computing. Devices that could be demonstrated in laboratory experiments were unable to cope with the requirements of large systems.[15]

The computer environment is unforgiving; devices are packed close together to minimize the time taken for signals to propagate between them and to take advantage of the economies of mass fabrication. The performance of devices is influenced by heat produced by their neighbors. Devices must endure high current densities and temperature cycling as host system is turned on and off. Besides the uncertainty inherent in devices, signals traversing wires in a computer are subject to attenuation and distortion during transmission and may be contaminated with crosstalk induced by nearby wires. In spite of these imperfections in devices and signals the logic circuits must make reliable binary decisions concerning the meaning of signals received from other devices. They must influence one another, forwarding information without feedback, and to do so with fan-out and fan-in. Success in preventing signal deterioration and loss in such an environment has only been achieved with true digital technology, the establishment of standard signal levels throughout a system and the resetting of signals to the correct digital value at each step. Resetting is accomplished by connecting to a standard with trnasistors, three terminal devices with high gain that can act as switches.

The prominent defect in novel devices that have been seriously considered as alternatives is a lack of the gain needed to make the connections that restore signals to standard values.   No solid state rival of the transistor has appeared in the four decades since transistors replaced the relay and the vacuum tube.   The transistor is unique, and is essential to solid state computational electronics.

REFERENCES

1. M. Riordan and L. Hoddeson, *Crystal Fire: The Birth of the Information Age* (W. W. Norton, New York, (1997)
2. W. F. Brinkman, D. E. Haggan, and W. W. Troutman, “A History of the Invention of the transistor” IEEE J. Solid-State Ckts, 32 pp.1858-1868 (1997)
3. L. Esaki “New phenomenon in narrow p-n junctions” Phys. Rev. 109 pp. 603-604 (1958)
4. S. P. Gentile *Theory and Application of Tunnel Diodes* (Princeton, NJ: van Nostrand, 1962), Chapter III.8.
5. R. W. Keyes “The Effect of Randomness in the Distribution of Impurity Atoms on FET Thresholds,” Applied Physics 8, pp. 251-279 (1975).
6. K. Bernstein et al “High-performance CMOS variability in the 65 nm regime,” IBM Journal Res. and Dev. 50 pp. 433-450 (2006)
7. I. V. Vernik, Q. P. Herr, K. Gaj, and M. J. Feldman, “Local timing parameter variations in RSFQ circuits,” IEEE Trans. Appl. Superconductivity 9, pp. 4341-44 (1999]
8. M. Orshansky et al, “Impact of Gate Length Variability on Performance of Digital Circuits,” IEEE Trans. Computer Aided Design 21, pp. 544-53 (2000).
9. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, (Cambridge: Cambridge University Press, (1998)
10. D. P. Holmes and P. L. Baynton, “A monolithic gallium arsenide tunnel diode construction,” Proceedings 1966 Symposium on Gallium Arsenide, pp. 236-240.
11. J. F. Whitaker et al “Picosecond switching time of resonant tunneling diode” Appl. Phys. Letters 53 pp. 385-387 (1988)
12. A. L. Robinson, “Superconducting Electronics: Toward an Ultrafast Computer,” Science 201 pp. 602-605 (1978)
13. S. Hasuo, “Towards the Josephson computer,” Physics World, pp. 37-40, (May 1990); “Towards the realization of a Josephson computer” Science 255 pp. 301-305 (1992)
14. T. R. Gheewala, “Josephson-Logic Devices and Circuits,” IEEE Trans. Electr. Dev. 10, pp. 1857-1863 (1980)
15. R. W. Keyes: “The cloudy crystal ball: electronic devices for logic,” Phil. Mag. B, 81, pp. 1315-30, (2001)

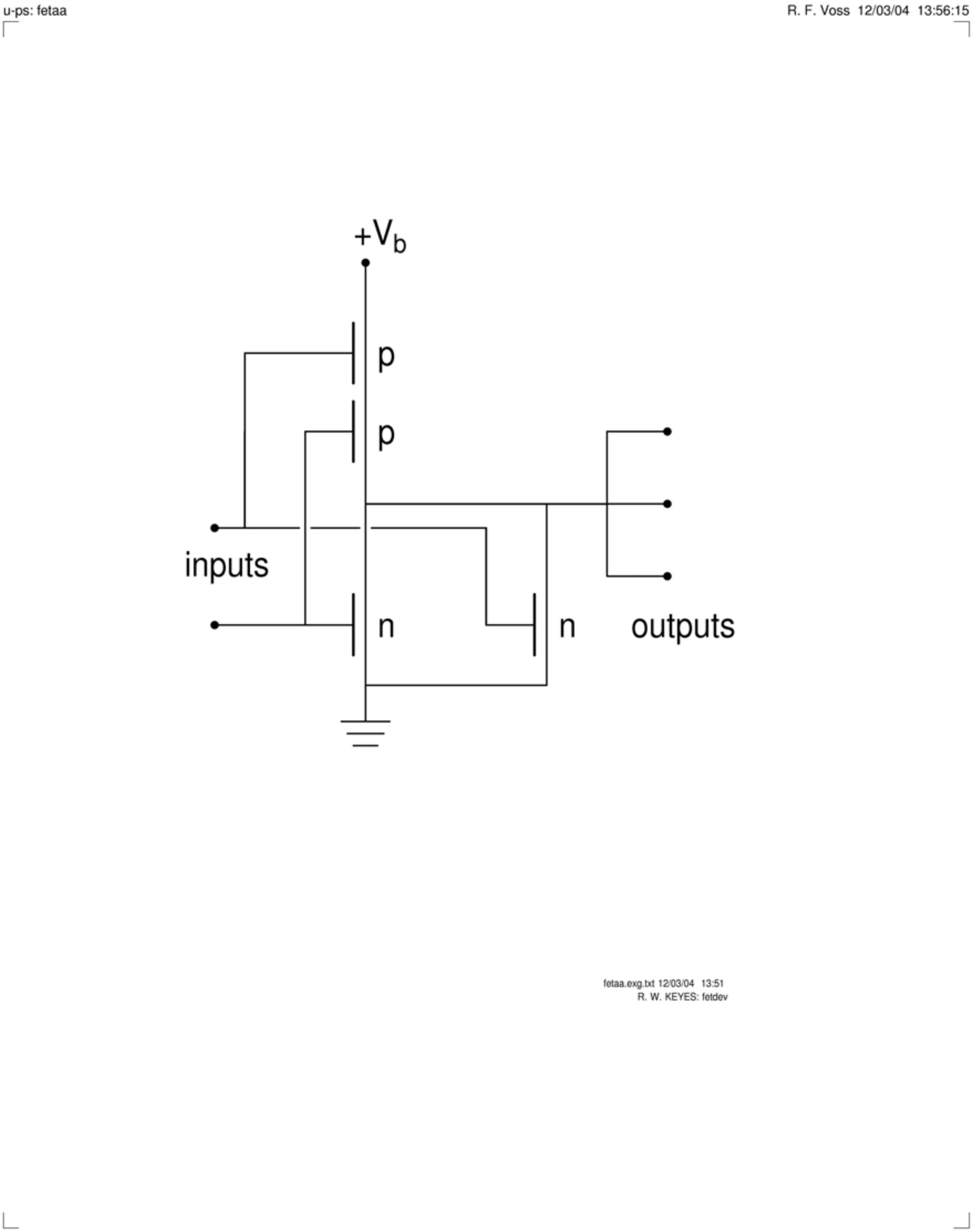


Fig. 1. The FET NOR circuit.

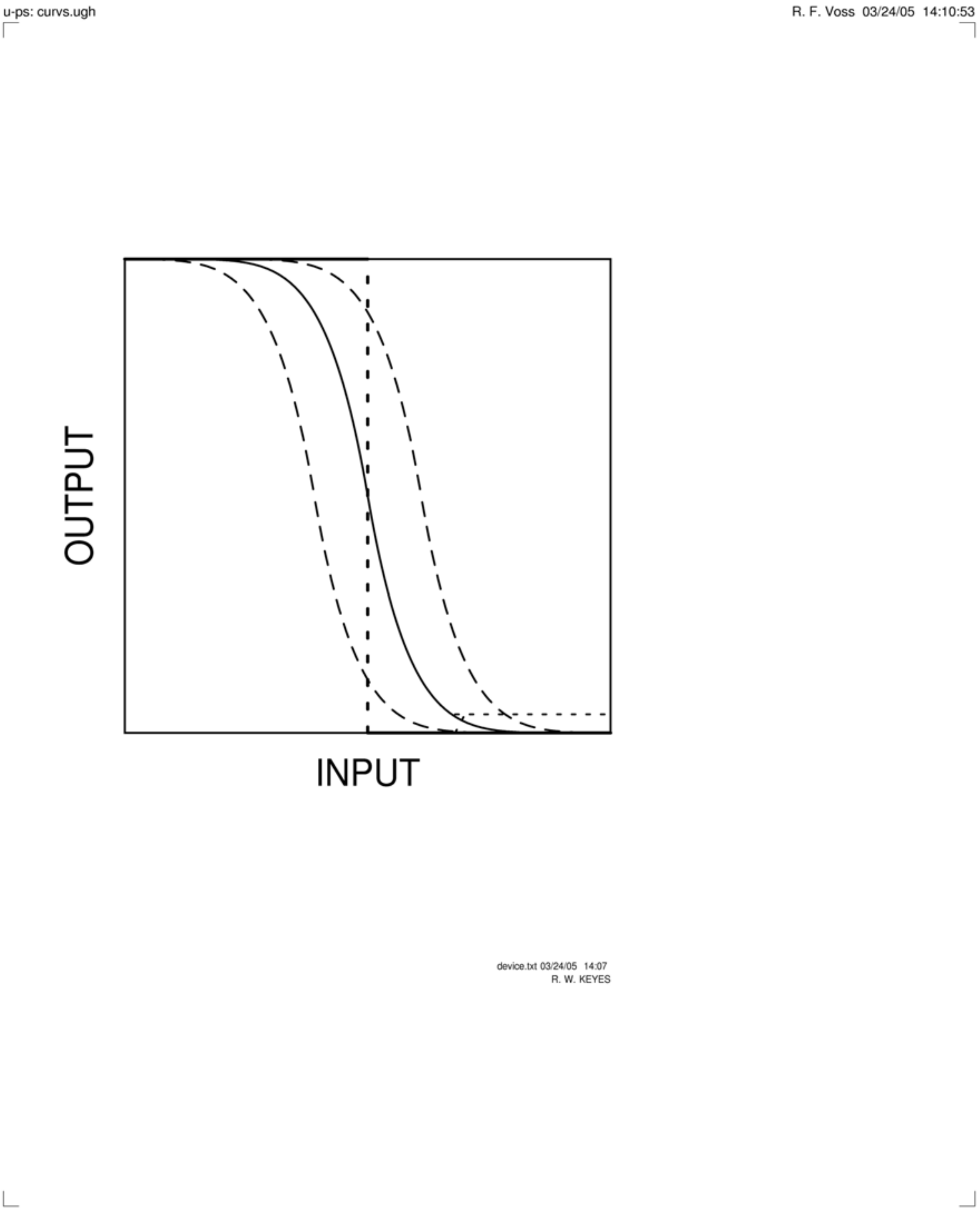
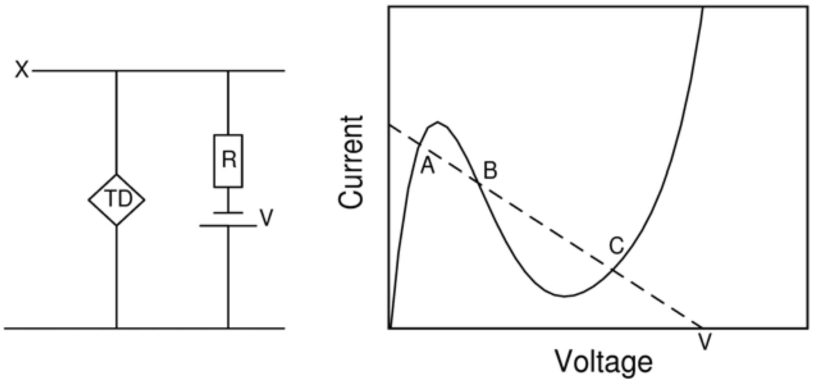


Fig. 2. The response of a transistor to a voltage applied to its gate.

u-ps: tunnl1.ugh R. F. Voss 12/07/05 16:55:21



tunnel.txt 12/07/05 16:48
R. W. KEYES

Fig. 3. The tunnel diode circuit intended for logical processing. The intersections of the tunnel diode current-voltage characteristic with the resistive load line A and C are stable, state B is unstable.