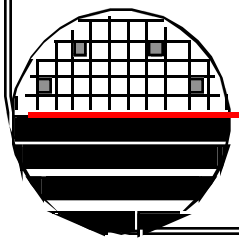
ROCHESTER INSTITUTE OF TECHNOLOGY
MICROELECTRONIC ENGINEERING

CMOS Process Variations
EEPROM Fabrication Technology

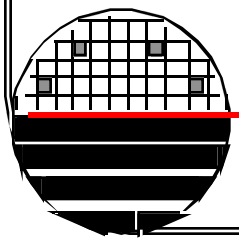
Dr. Lynn Fuller

Webpage: http://people.rit.edu/lffeee
Microelectronic Engineering
Rochester Institute of Technology
82 Lomb Memorial Drive
Rochester, NY 14623-5604
Tel (585) 475-2035
Email: Lynn.Fuller@rit.edu
Department Webpage: http://www.microe.rit.edu



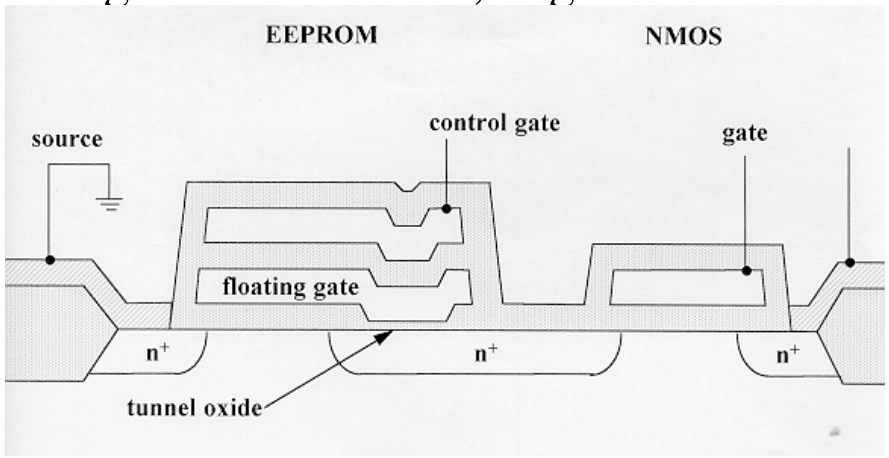
OUTLINE

Introduction
Tunneling Gate Oxide EEPROM
Programming
Reading
Test Specification/Results
Fowler-Nordheim Tunneling
Process Variation
Tunnel Oxide Recipe
Test Chip Layout
Step-by-Step Process



INTRODUCTION

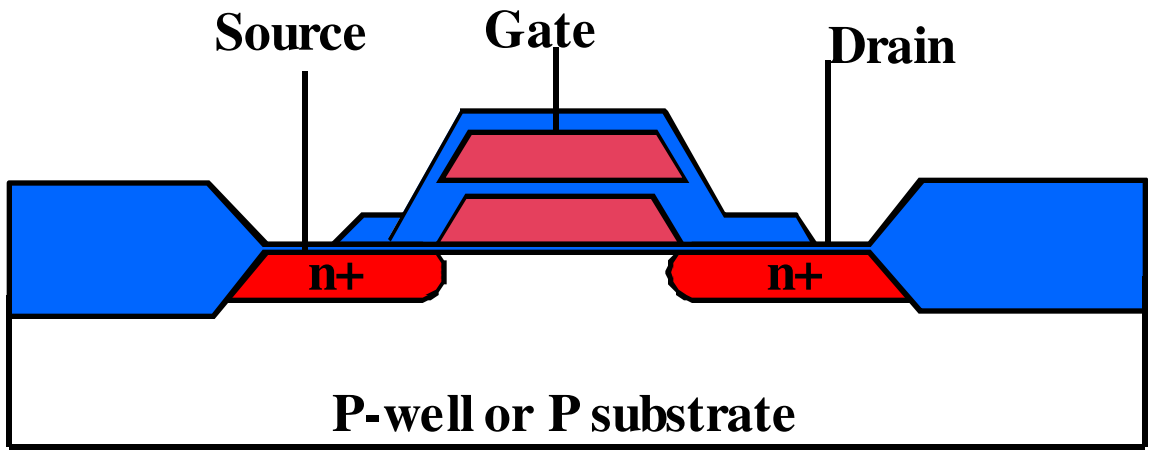
In certain applications, data must be electrically entered and erased from Read Only Memory (ROM). The procedure can involve the entire ROM sections or one memory cell at a time. From the various technologies available, we have chosen the design of a FLOTOX EEPROM (FLOating-gate Tunneling Oxide Electrically Erasable Programmable ROM, Figure 1.



This EEPROM cell has double polysilicon gates, with the top polysilicon as the control gate and the lower polysilicon as the floating gate. A thin tunneling oxide is formed above the drain in the FLOTOX Transistor.

TUNNELING GATE OXIDE EEPROM

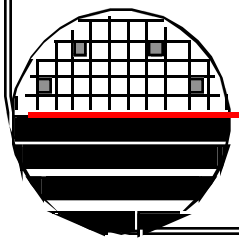
Another form of the FLOTOX Transistor is shown below. The structure is simpler and smaller but is more difficult to manufacture because of the problems associated with diffusion of phosphorous from the gate poly through the tunnel oxide into the transistor channel region. Most modern EEPROM devices use this structure.



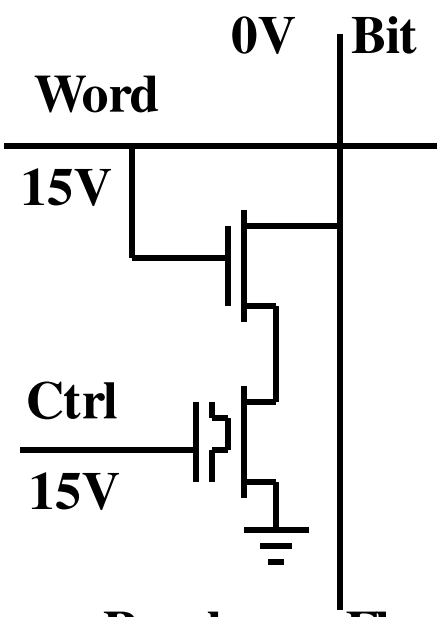
INTRODUCTION

The EEPROM is programmed by transferring electrons between the floating-gate and the substrate, through the tunneling oxide, by means of Fowler-Nordheim tunneling. There are two modes of programming the EEPROM: write and erase. First, in the write mode, the floating-gate is charged negatively by electrons that tunnel from the drain to the floating gate. The charging is done by applying a +15V voltage to the control gate and connecting both the drain and source to ground.

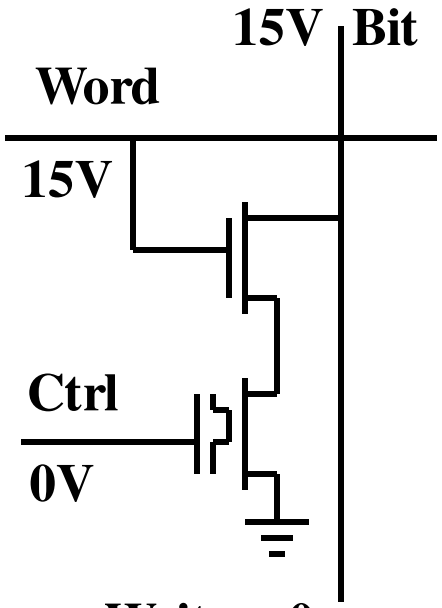
The negative charge stored on the floating gate has the effect of shifting the threshold voltage towards a more positive value. When the floating gate is charged, the normal +5V applied to the control gate during a read operation will not be sufficient for the transistor to conduct channel current. Only when the floating gate is uncharged, will the transistor be able to conduct with +5V on the control gate.



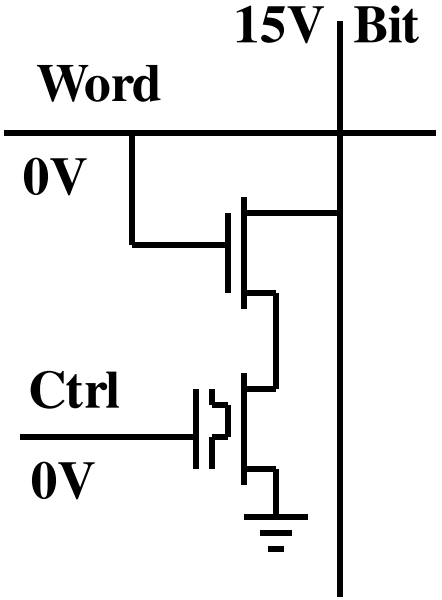
PROGRAMMING THE EEPROM CELL



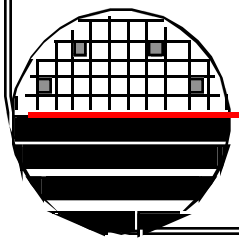
Precharge Floating Gate with Electrons



Write a 0
Remove Electrons From Floating Gate



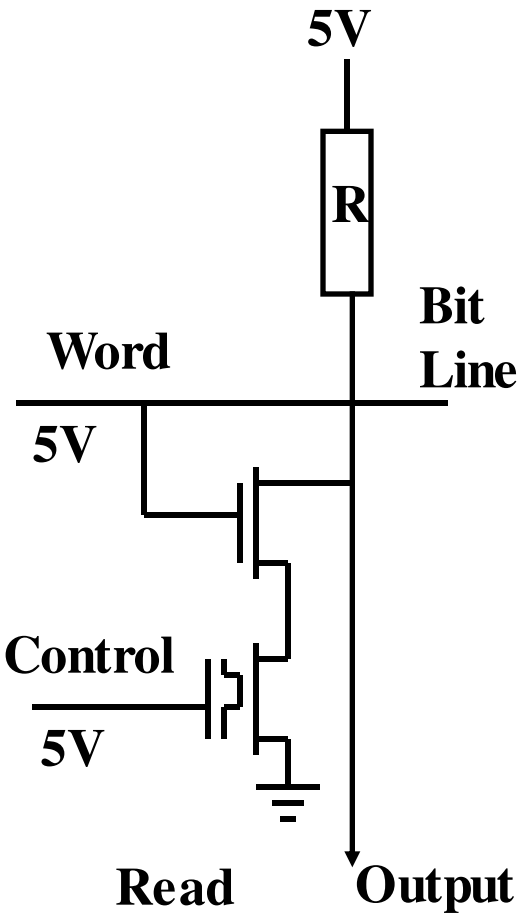
Write a 1
Leave Electrons on Floating Gate

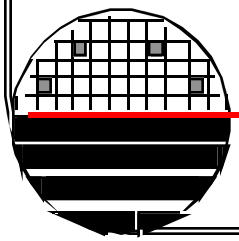


READING THE CELL

If the floating gate is charged with electrons the 5 volts on the control line will not be enough to turn that transistor on. Thus the output will be high.

If the floating gate is uncharged with electrons the 5 volts on the control line will turn that transistor on. Thus the output will be low.

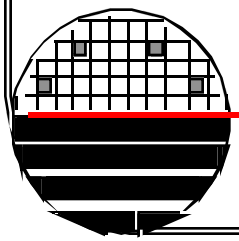




TEST SPECIFICATION

Specification

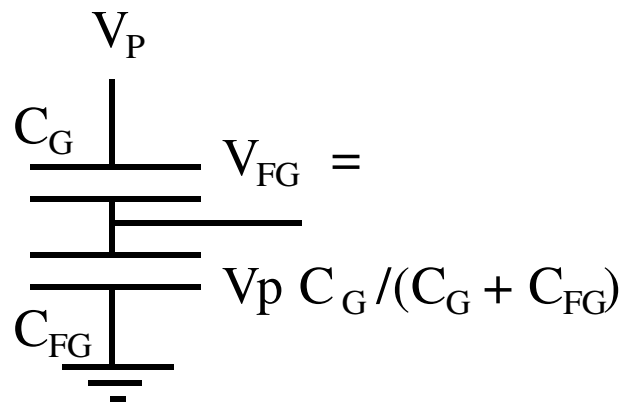
- Can FLOTOX Transistor be programmed
  - Charge the Floating Gate
  - Measure the subthreshold characteristics
  - Discharge the Floating Gate
  - Measure the subthreshold characteristics
- Can the FLOTOX Transistor hold the charge
  - Charge the Floating Gate
  - Measure the subthreshold characteristics
  - Wait 1hours, 10 hours, 100 hours, 1000hours
  - Measure the subthreshold characteristics
- Can the FLOTOX Transistor be cycled many times
- How much time does it take to charge and discharge



FOWLER-NORDHEIM TUNNELING

$$J = C_1 E_c^2 \exp(-Eo/Ec)$$

$$C_1 = q^3 m / (8\pi\hbar\phi_b m^*)$$
$$= 9.625E\text{-}7\ A/V^2$$
$$Ec = V_{FG}/tox$$
$$Eo = 8\pi(2m)^{1/2} \phi_b^{3/2} /(3hq)$$



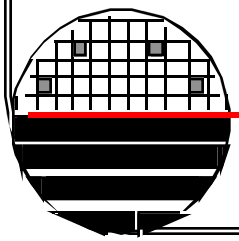
Example: let C_{FG}=0.3 pF, C_G=0.2 pF, Vp=25 volts, tox = 100 Å, m*=.5m_e , ϕ_b = 3.2
where m_e=9.11E-31, $\hbar$=(6.625E-34)/2π, q=1.6E-19

so Eo=2.765E8 V/cm and Ec = 10/100E-8 V/cm

$$J = 9.625E\text{-}7\ (Ec)^2 \exp\ (-2.765E8/Ec) = 94.4\ \mu A/cm^2$$

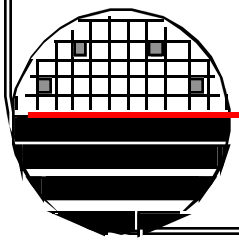
CMOS PROCESS MODIFICATIONS

The fabrication of a FLOTOX EEPROM involves a three modifications to the present CMOS process. An additional n+ drain implant is performed before the polysilicon layers are deposited. A thin 100 Å tunneling oxide is grown above the additional n+ implant region. A second polysilicon layer is deposited for the control gate.



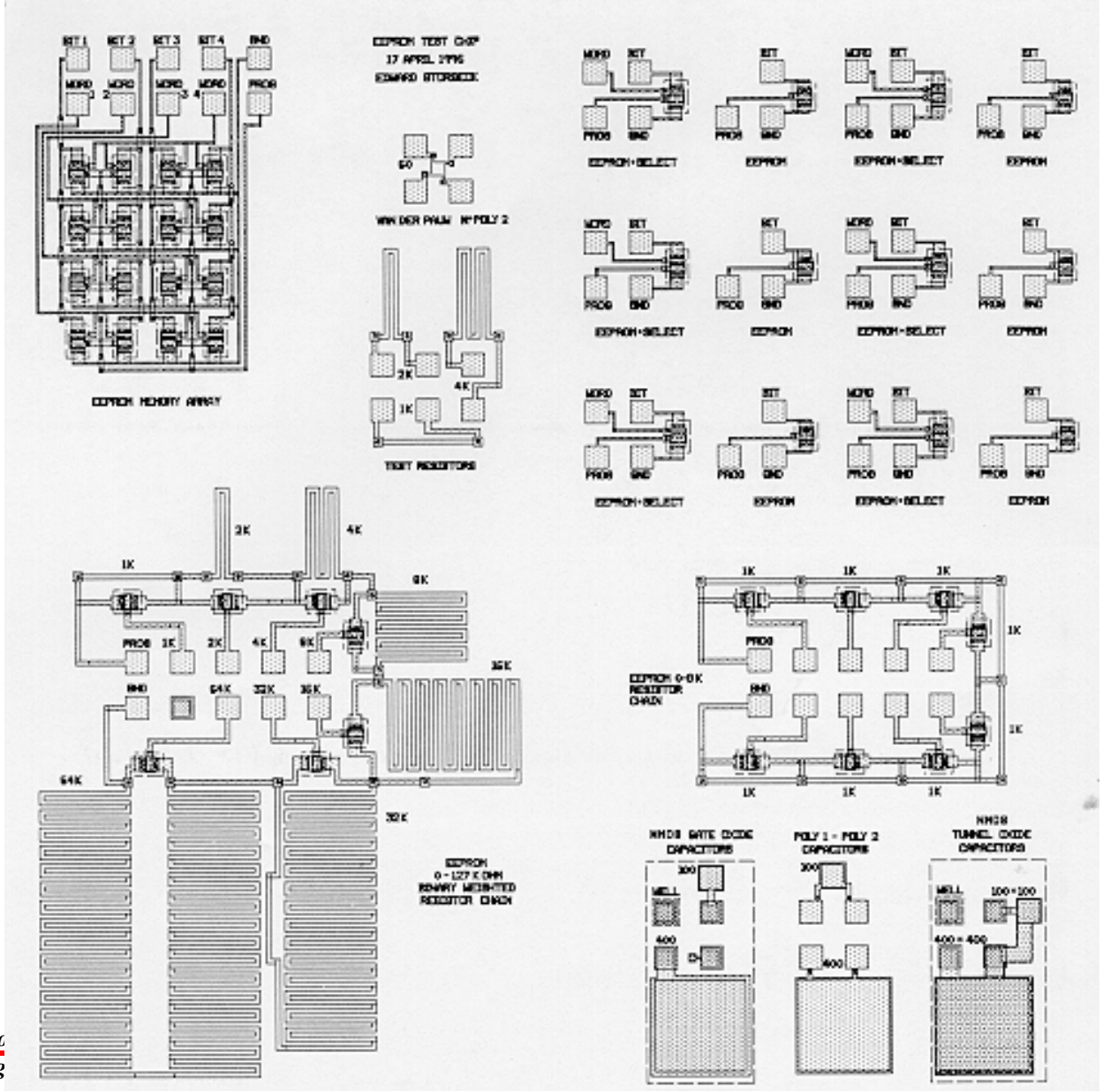
TUNNEL OXIDE RECIPE FOR ~120 Å SiO2

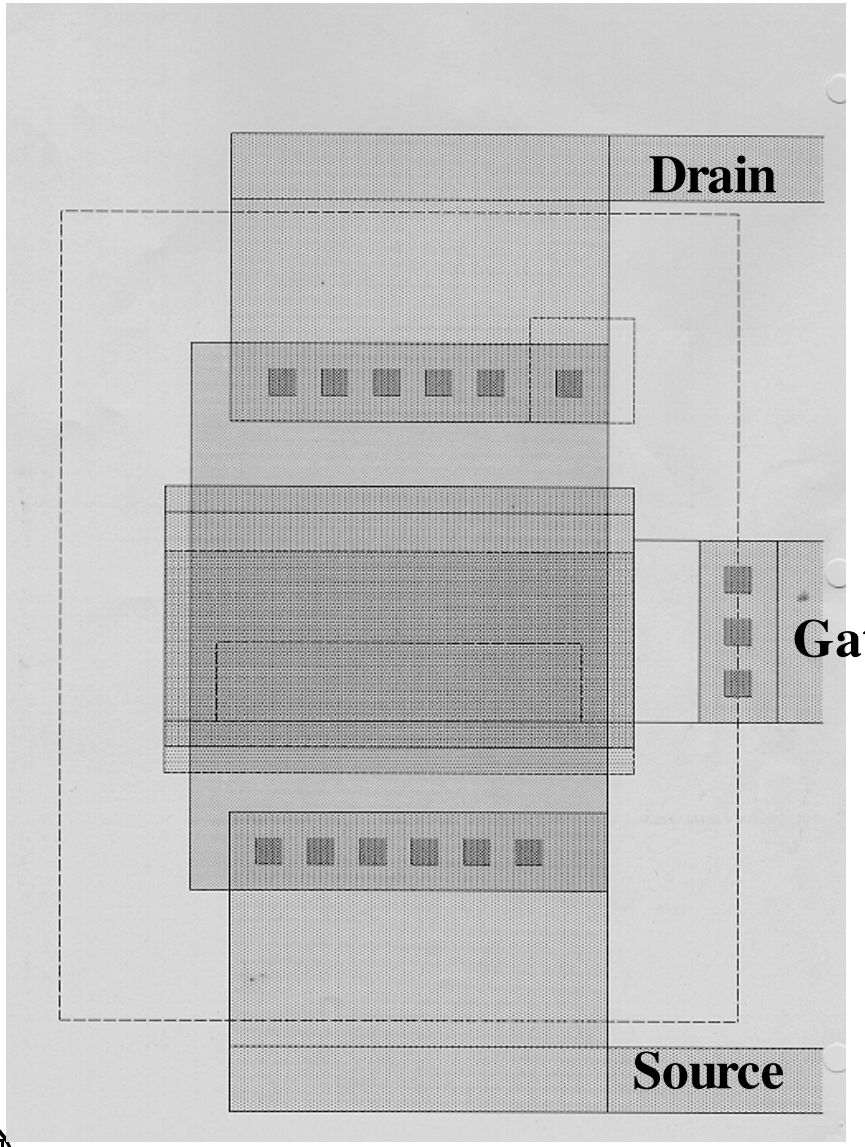
Step	Gas Flow	Temperature	Time	Boat
0 Load Tube	N2 @ 15 lpm	650 °C	?	Out
1 Push	N2 @ 15 lpm	650 °C	15 min	In
2 Stabilization	N2 @ 15 lpm	650 °C	15 min	
3 Ramp Up	N2 @ 15 lpm	650 to 950 °C	30 min	
4 First Oxide	N2 @ 15 lpm + O2 @ 5 lpm	950 °C	15 min	
5 First Anneal	N2 @ 15 lpm	1050 °C	30 min	
6 Ramp Down	N2 @ 15 lpm	950 °C	20 min	
7 2nd Oxide	N2 @ 15 lpm + O2 @ 5 lpm	950 °C	10 min	
8 2nd Anneal	N2 @ 15 lpm	950 °C	30 min	
9 Ramp Down	N2 @ 15 lpm	950 to 650 °C	60 min	
10 Pull	N2 @ 15 lpm	650 °C	15 min	Out



Entire Test Chip

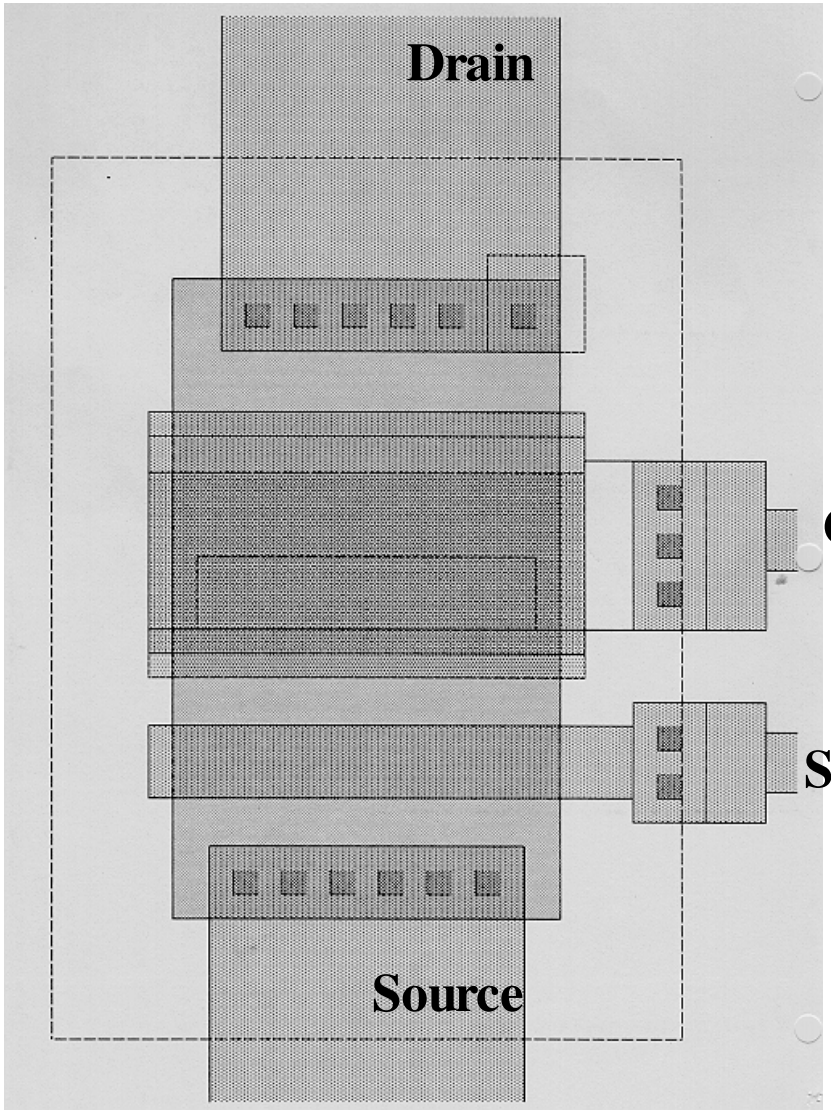
- EEPROM Transistor
- EEPROM plus Select
- EEPROM Memory Array
- Variable Programmable Resistor
- Binary-weighted Variable Programmable Resistor
- Resistors
- Capacitors





EEPROM Transistor

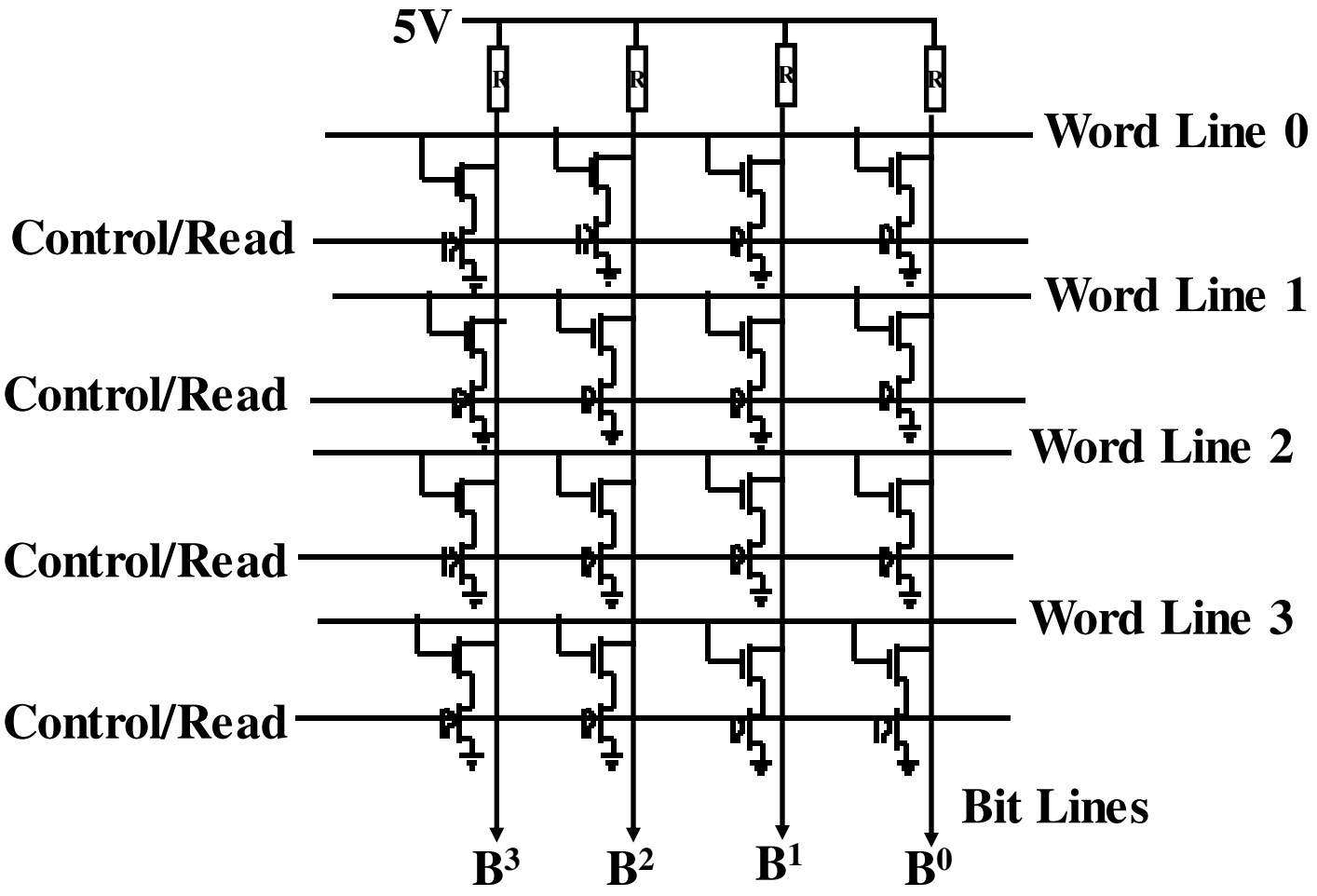
Basic transistor with two layers poly, tunnel oxide, and 3 connections.



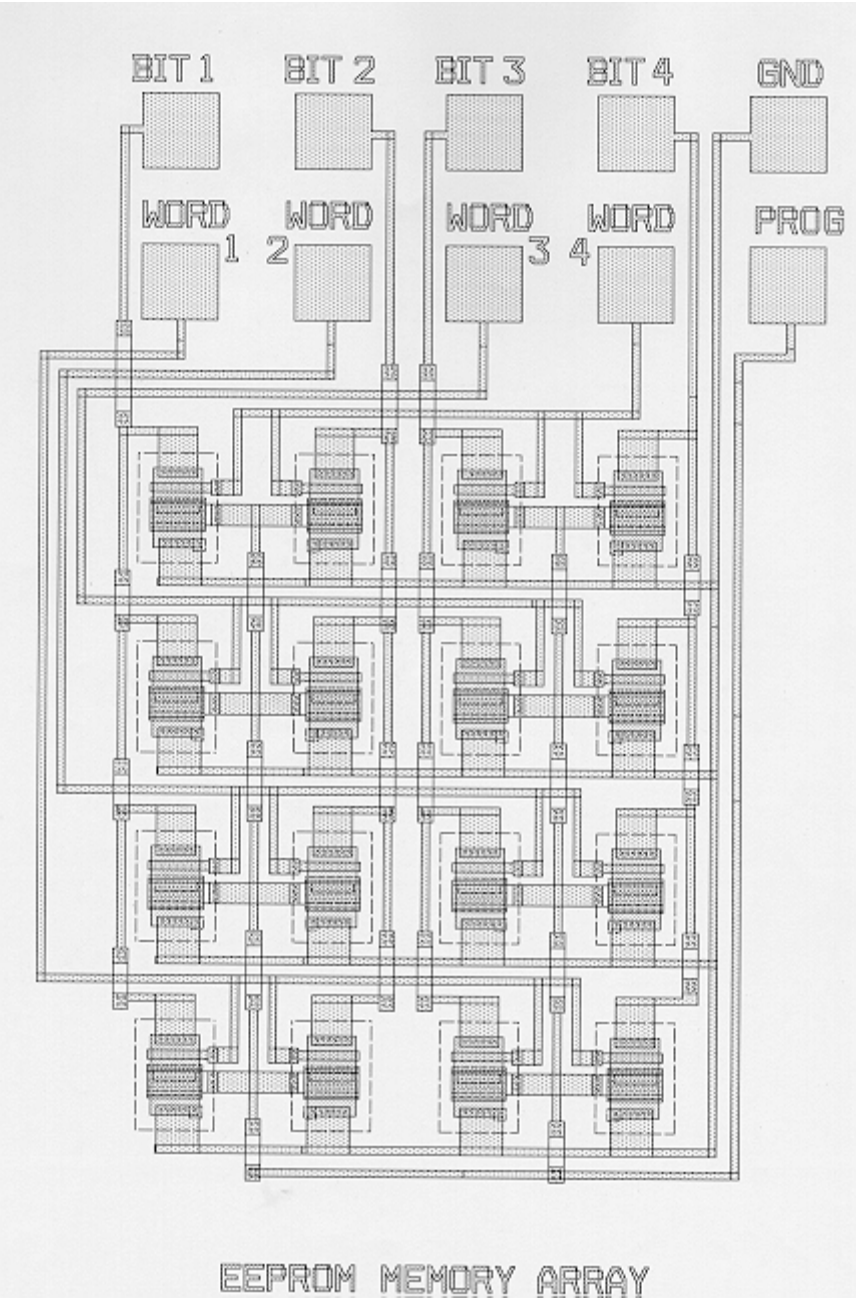
EEPROM Transistor + Select

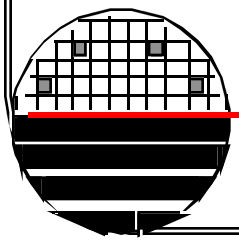
Basic transistor with two layers poly, tunnel oxide, and 4 connections.

EEPROM

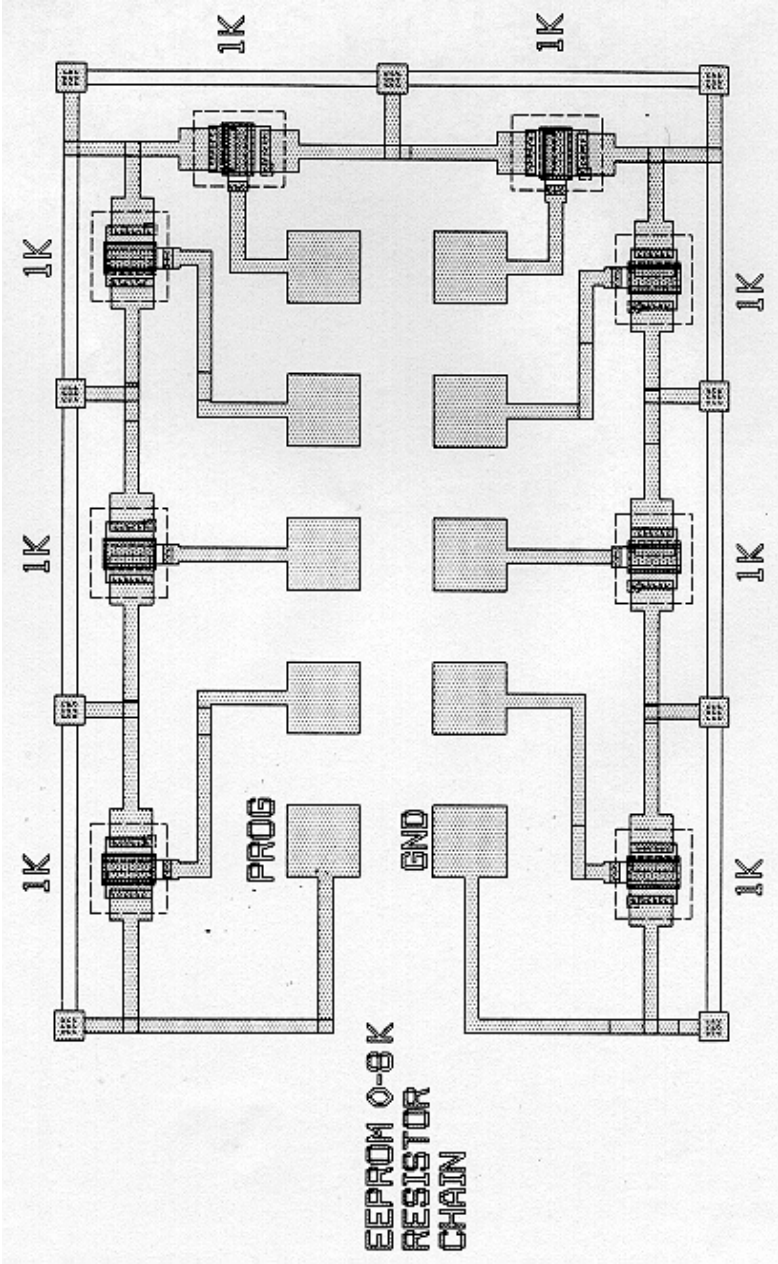


Memory array of four-by-four EEPROM cells

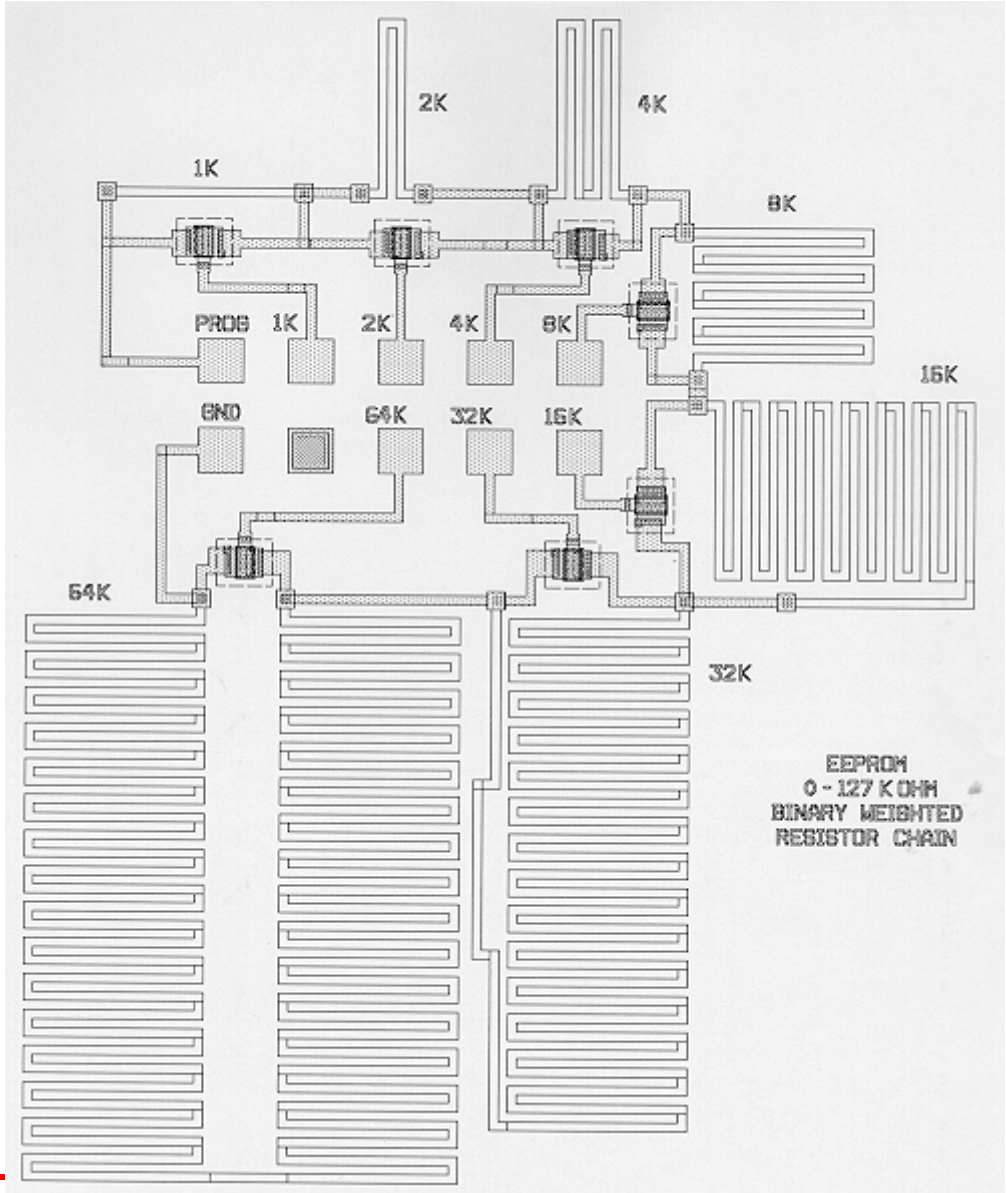




A variable programmable resistor with equal 1000 ohm resistors in series.

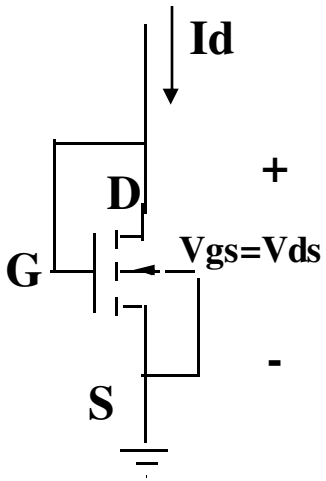


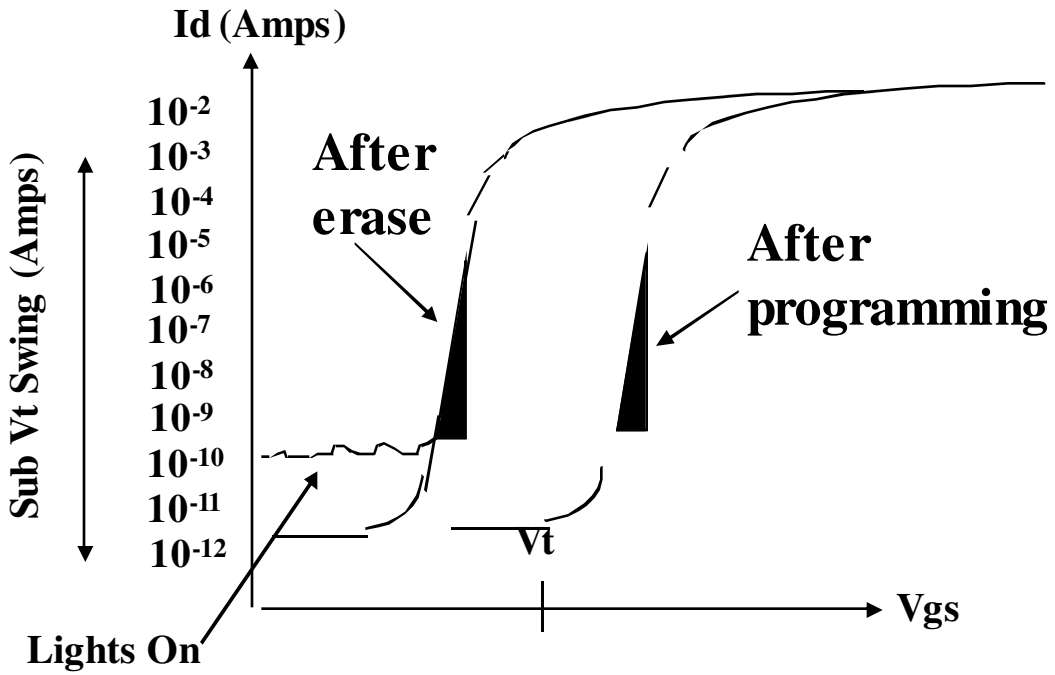
A variable programmable resistor with binary-weighted resistors, 1k, 2k, 4k, 8k .. .. ..64 k ohms



TEST RESULTS

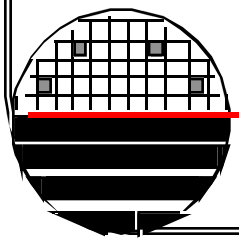
Programmed with 3 seconds of +25 volts on gate
Erase with 4 seconds of 25 volts on the Drain





REFERENCES

1. "Flash EE PROM Device", Keith Zawadski, Senior Desing Project, May 1996.
2. "EEPROMs", Edward Storbeck, Visiting Scholar from South Africa, 1996-97.
3. Device Electronics for Integrated Circuits, Richard S. Muller, Theodore I. Kamins, John Wiley and Sons, Inc.
4. Nonvolatile Semiconductor Memory Technology, Edited by William D. Brown and Joe E. Brewer, IEEE Press

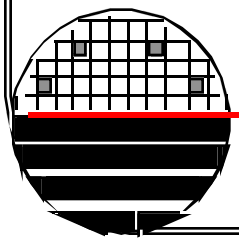


HOMEWORK - EEPROM

1.0 Calculate the current that would flow through a 10 by 10 µm tunnel oxide of 100 Å, at a control gate voltage of 10, 15 and 20 Volts, assume CFG=0.3pF and CG=0.2 pF

2.0 How long would it take to charge the floating gate to 2.5 volts. Assume the floating gate is 0.3 pF floating gate and the currents are as found in problem 1.0.

3.0 Describe the exact procedure that you would use to test the FLOTOX transistor.

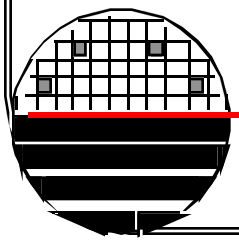


HOMEWORK - EEPROM

1.0 Calculate the current that would flow through a 10 by 10 µm tunnel oxide of 100 Å, at a control gate voltage of 10, 15 and 20 Volts, assume CFG=0.3pF and CG=0.2 pF

2.0 How long would it take to charge the floating gate to 2.5 volts. Assume the floating gate is 0.3 pF floating gate and the currents are as found in problem 1.0.

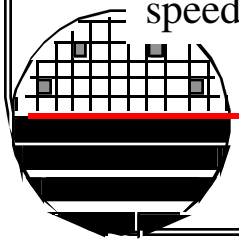
3.0 Describe the exact procedure that you would use to test the FLOTOX transistor.



NOR AND NAND FLASH

Flash memory is a non-volatile computer storage that can be electrically erased and reprogrammed. It is a technology that is primarily used in memory cards and USB flash drives for general storage and transfer of data between computers and other digital products. It is a specific type of EEPROM (Electrically Erasable Programmable Read-Only Memory) that is erased and programmed in large blocks; in early flash the entire chip had to be erased at once. Flash memory costs far less than byte-programmable EEPROM and therefore has become the dominant technology wherever a significant amount of non-volatile, solid state storage is needed. Example applications include PDAs (personal digital assistants), laptop computers, digital audio players, digital cameras and mobile phones. It has also gained popularity in console video game hardware, where it is often used instead of EEPROMs or battery-powered static RAM (SRAM) for game save data.

Since flash memory is non-volatile, no power is needed to maintain the information stored in the chip. In addition, flash memory offers fast read access times (although not as fast as volatile DRAM memory used for main memory in PCs) and better kinetic shock resistance than hard disks. These characteristics explain the popularity of flash memory in portable devices. Another feature of flash memory is that when packaged in a "memory card," it is extremely durable, being able to withstand intense pressure, extremes of temperature, and even immersion in water. Although technically a type of EEPROM, the term "EEPROM" is generally used to refer specifically to non-flash EEPROM which is erasable in small blocks, typically bytes. Because erase cycles are slow, the large block sizes used in flash memory erasing give it a significant speed advantage over old-style EEPROM when writing large amounts of data.



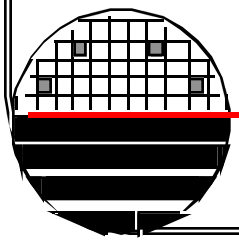
Wikipedia contributors. "Flash memory." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Feb. 2010. Web. 11 Feb. 2010.

NOR AND NAND FLASH

NOR and NAND flash differ in two important ways: the connections of the individual memory cells are different the interface provided for reading and writing the memory is different (NOR allows random-access for reading, NAND allows only page access) These two are linked by the design choices made in the development of NAND flash. A goal of NAND flash development was to reduce the chip area required to implement a given capacity of flash memory, and thereby to reduce cost per bit and increase maximum chip capacity so that flash memory could compete with magnetic storage devices like hard disks. NOR and NAND flash get their names from the structure of the interconnections between memory cells. In NOR flash, cells are connected in parallel to the bitlines, allowing cells to be read and programmed individually. The parallel connection of cells resembles the parallel connection of transistors in a CMOS NOR gate. In NAND flash, cells are connected in series, resembling a NAND gate. The series connections consume less space than parallel ones, reducing the cost of NAND flash. It does not, by itself, prevent NAND cells from being read and programmed individually. When NOR flash was developed, it was envisioned as a more economical and conveniently rewritable ROM than contemporary EPROM, EAROM, and EEPROM memories. Thus random-access reading circuitry was necessary. However, it was expected that NOR flash ROM would be read much more often than written, so the write circuitry included was fairly slow and could only erase in a block-wise fashion. On the other hand, applications that use flash as a replacement for disk drives do not require word-level write address, which would only add to the complexity and cost unnecessarily. Because of the series connection and removal of wordline contacts, a large grid of NAND flash memory cells will occupy perhaps only 60% of the area of equivalent NOR cells (assuming the same CMOS process resolution, e.g. 130nm, 90 nm, 65 nm). NAND flash's designers realized that the area of a NAND chip, and thus the cost, could be further reduced by removing the external address and data bus circuitry. Instead, external devices could communicate with NAND flash via sequential-accessed command and data registers, which would internally retrieve and output the necessary data. This design choice made random-access of NAND flash memory impossible, but the goal of NAND flash was to replace hard disks, not to replace ROMs.

Wikipedia contributors. "Flash memory." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Feb. 2010. Web. 11 Feb. 2010.

EEPROM Technology



132GBit NAND FLASH

ISSCC: SanDisk set to show highest density NAND flash

Peter Clarke

2/22/2012 7:13 AM EST

LONDON – SanDisk Corp. is expected to report one of the highlights of the International Solid-State Circuits Conference, being held in San Francisco, when it provides details of a NAND flash memory implemented in 19-nm CMOS.

The chip, set to be discussed in paper 25.8 in the non-volatile memory session of ISSCC, is a 128-Gbit monolithic device that stores 3-bits per memory cell, is the highest density IC ever produced. The chip also has a 3-bit per cell write performance of 18-Mbyte per second and a read throughput of 400-Mbits per second. The chip is a rectangle of silicon of 170 square millimeters area.

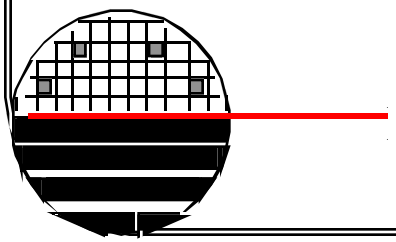
The technology is described as proof that 3-bit per cell NAND has reached a level of maturity equivalent to 1- and 2-bit per cell NAND and will be welcomed for use as motherboard memory in applications such as tablet computers and smartphones and for solid-state drives (SSDs).

In the competitive world of NAND flash memory this device is already in commericial production. Other papers at ISSCC can discuss technologies and circuit ideas that are two, four or more years away from possible commercial implementation.

SanDisk has a roadmap that sees 19-nm NAND flash ramping production in 2012, followed by a 1Y-nm process lowering the cost of 128-Gbit memory ICs in 2013 and a 1Z-nm process taking monolithic memory to 256-Gbits late in 2014. "We believe NAND technology will scale for a few more generations," Ritu Shrivastava, vice president of technology development at SanDisk told analysts recently.

The 128-Gbit NAND flash memory chip was developed jointly by teams from SanDisk and Toshiba at SanDisk's Milpitas campus. The effort was led by Yan Li, director of Memory Design at SanDisk. Products based on the 128-Gbit three-bit per cell technology began shipping late last year and have already started to ramp into high volume production. SanDisk has also developed a derivative product, a 64-Gbit NAND flash memory chip that is compatible with the industry-standard microSD format. The company has also started to ramp production of this additional chip technology.

Samsung Electronics Co. Ltd. is also set to present a sub-20nm NAND flash memory at ISSCC and with a higher read throughput of 533-Mbits per second. However the memory capacity is lower than SanDisk's at 64-Gbit.



ISSCC: SanDisk set to show highest density NAND flash

Peter Clarke

2/22/2012 7:13 AM EST

LONDON – SanDisk Corp. is expected to report one of the highlights of the International Solid-State Circuits Conference, being held in San Francisco, when it provides details of a NAND flash memory implemented in 19-nm CMOS.

The chip, set to be discussed in paper 25.8 in the non-volatile memory session of ISSCC, is a 128-Gbit monolithic device that stores 3-bits per memory cell, is the highest density IC ever produced. The chip also has a 3-bit per cell write performance of 18-Mbyte per second and a read throughput of 400-Mbits per second. The chip is a rectangle of silicon of 170 square millimeters area.

The technology is described as proof that 3-bit per cell NAND has reached a level of maturity equivalent to 1- and 2-bit per cell NAND and will be welcomed for use as motherboard memory in applications such as tablet computers and smartphones and for solid-state drives (SSDs).

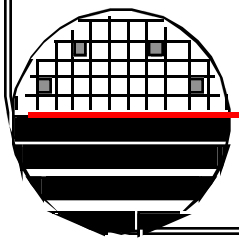
In the competitive world of NAND flash memory this device is already in commericial production. Other papers at ISSCC can discuss technologies and circuit ideas that are two, four or more years away from possible commercial implementation.

SanDisk has a roadmap that sees 19-nm NAND flash ramping production in 2012, followed by a 1Y-nm process lowering the cost of 128-Gbit memory ICs in 2013 and a 1Z-nm process taking monolithic memory to 256-Gbits late in 2014. "We believe NAND technology will scale for a few more generations," Ritu Shrivastava, vice president of technology development at SanDisk told analysts recently.

The 128-Gbit NAND flash memory chip was developed jointly by teams from SanDisk and Toshiba at SanDisk's Milpitas campus. The effort was led by Yan Li, director of Memory Design at SanDisk. Products based on the 128-Gbit three-bit per cell technology began shipping late last year and have already started to ramp into high volume production. SanDisk has also developed a derivative product, a

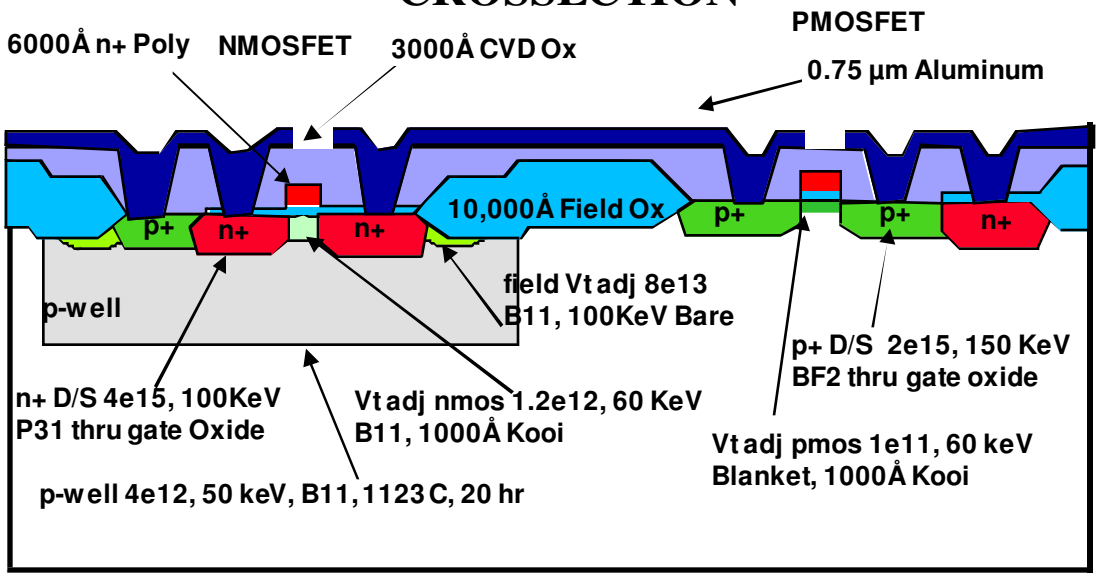
RIT EEPROM PROCESS

λ = 2.5 Micrometer Design Rules
9 Design Layers
12 Photolithography Layers
LOCOS
Extended n+ drain
n-Type Poly Floating and Control Gate
100A Tunneling Oxide over Drain



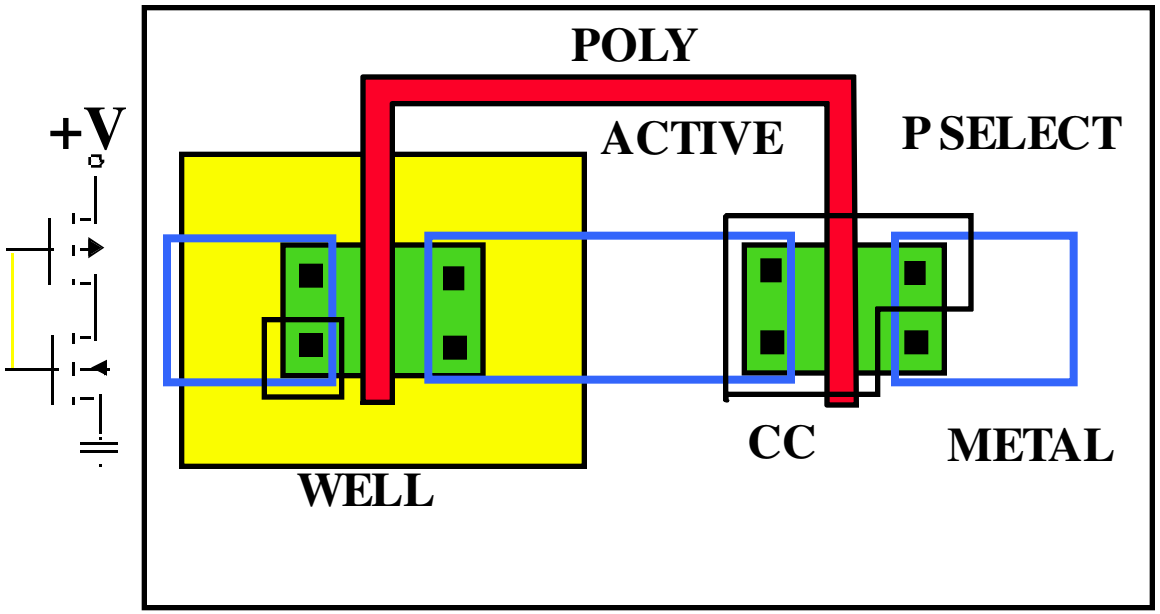
RIT P-WELL CMOS PROCESS

CROSSECTION

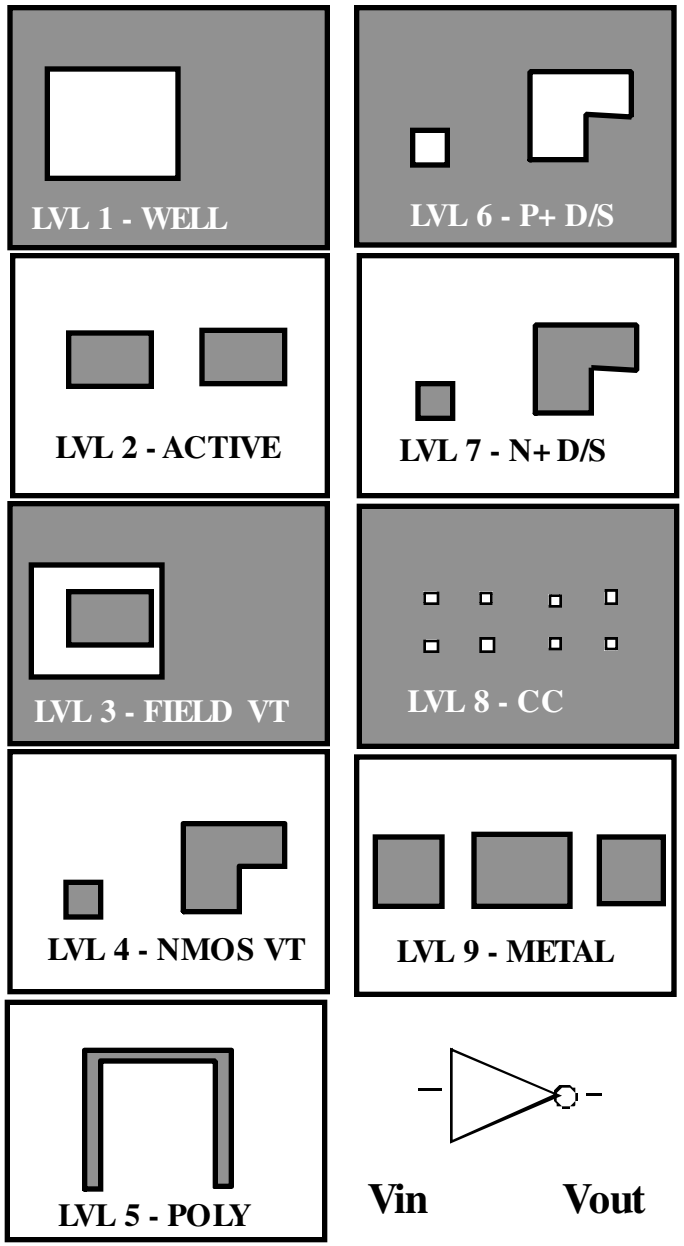


n-type substrate 10 ohm-cm (100)

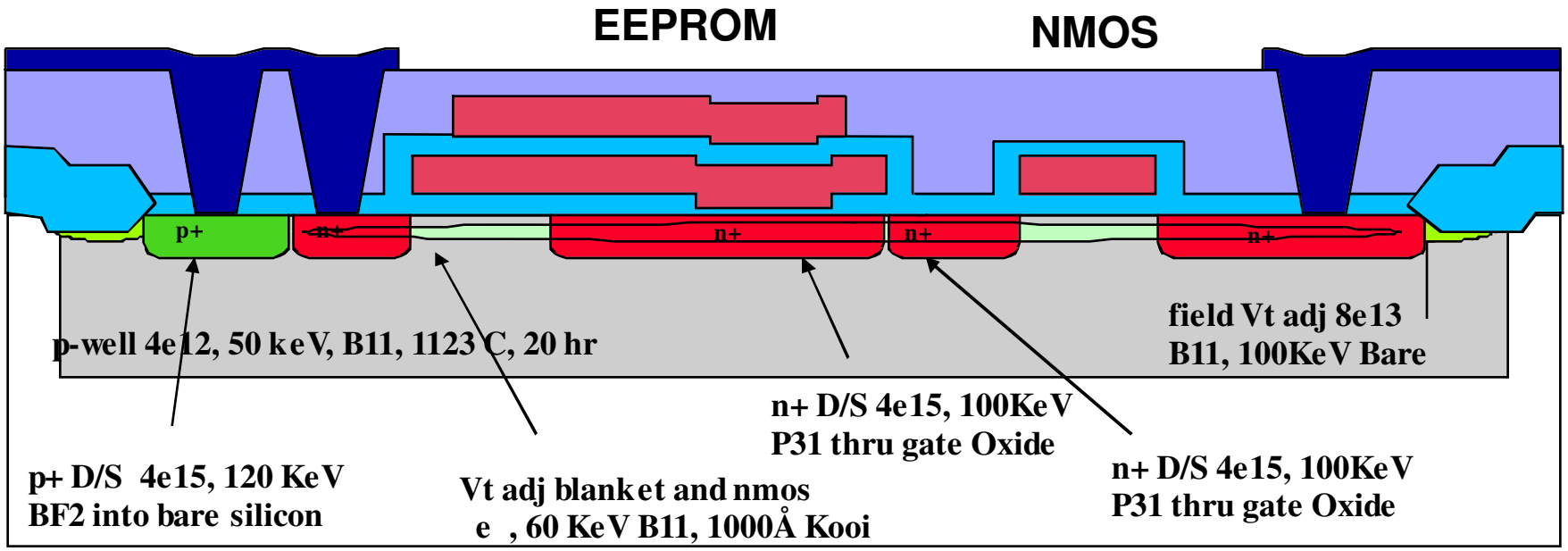
SIX LAYOUT LEVELS



NINE PHOTO LEVELS

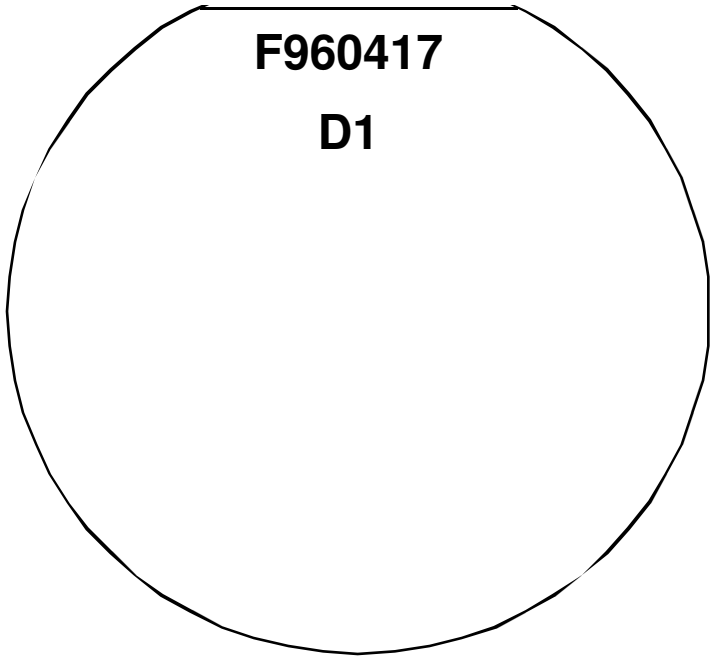


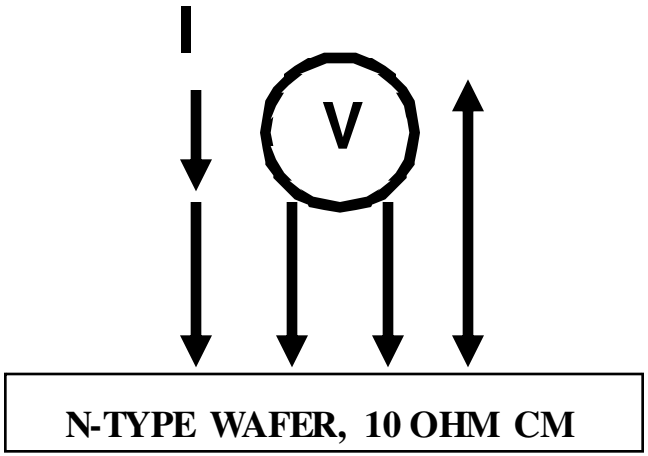
CROSS-SECTION



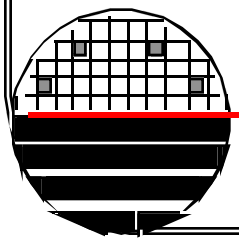
Three extra layers: **Drain Extension**
Tunnel Oxide
Floating Poly

IDENTIFY AND FOUR POINT PROBE



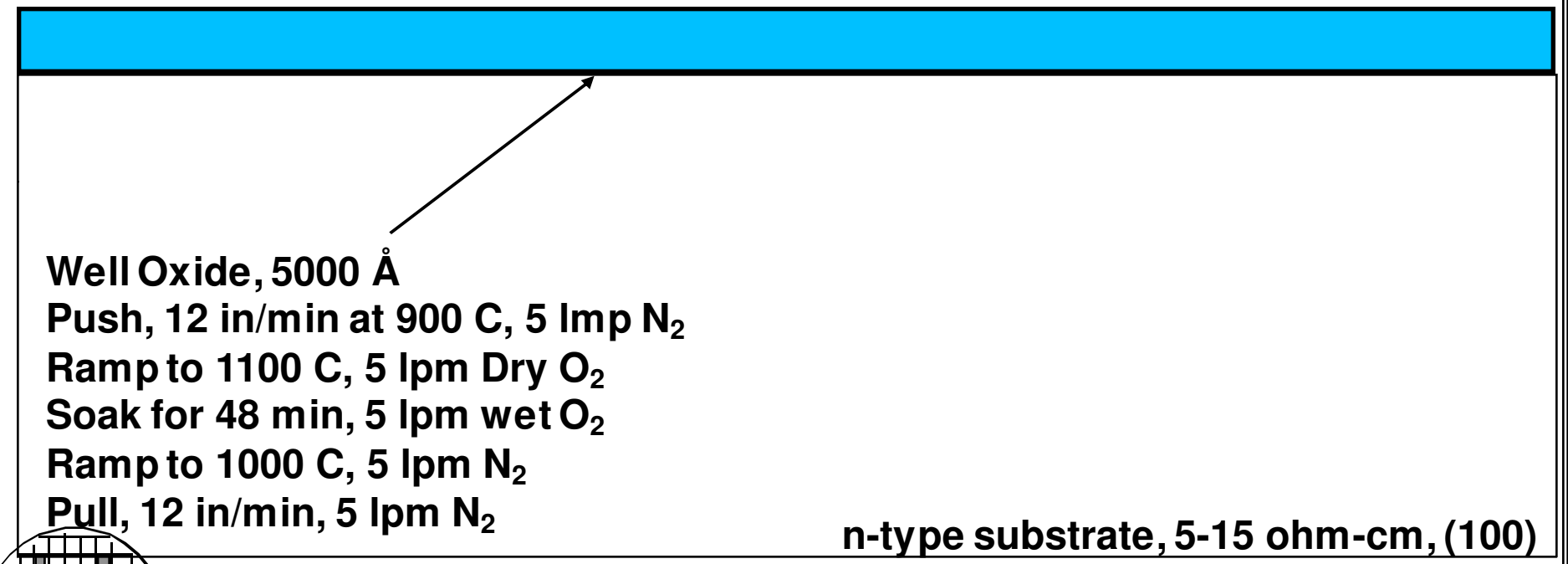


Rho = V / I * Pi / ln 2 ohm-cm



ALIGNMENT OXIDE

EEPROM **NMOS**



Well Oxide, 5000 Å
Push, 12 in/min at 900 C, 5 lmp N_2
Ramp to 1100 C, 5 lpm Dry O_2
Soak for 48 min, 5 lpm wet O_2
Ramp to 1000 C, 5 lpm N_2
Pull, 12 in/min, 5 lpm N_2

n-type substrate, 5-15 ohm-cm, (100)

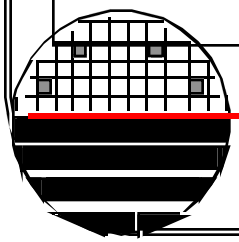
LEVEL 1 PHOTO - WELL

EEPROM NMOS

Photoresist

well Oxide, 5000 Å

n-type substrate, 5-15 ohm-cm, (100)



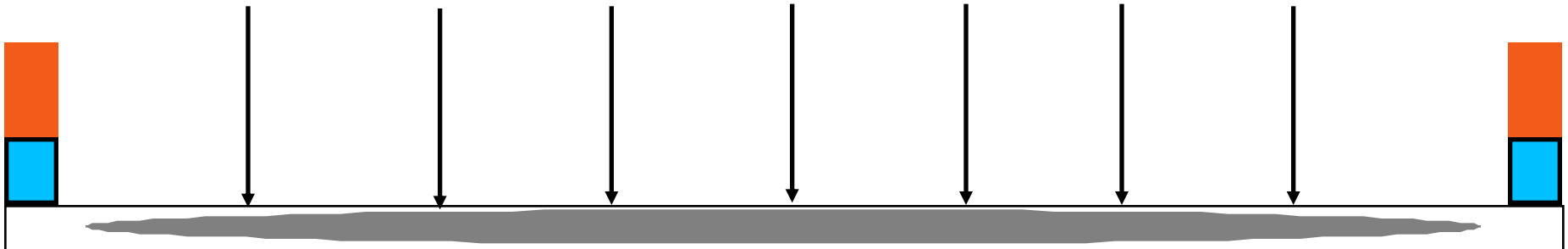
OXIDE ETCH

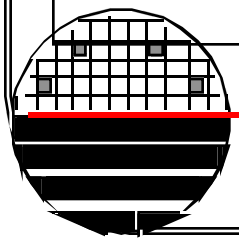
EEPROM **NMOS**

WELL IMPLANT

EEPROM NMOS

p-well implant, 4e12, 50 KeV, B11





STRIP RESIST

EEPROM **NMOS**

p-well implant, 4e12, 50 KeV, B11

POST WELL DRIVE

EEPROM NMOS

5800 Å Oxide (increase by 800 Å)

p-well, 4e12, 50 KeV, B11, 1125 C, 20 hrs
xj = 4.5 µm
Rhos = 3000 ohms

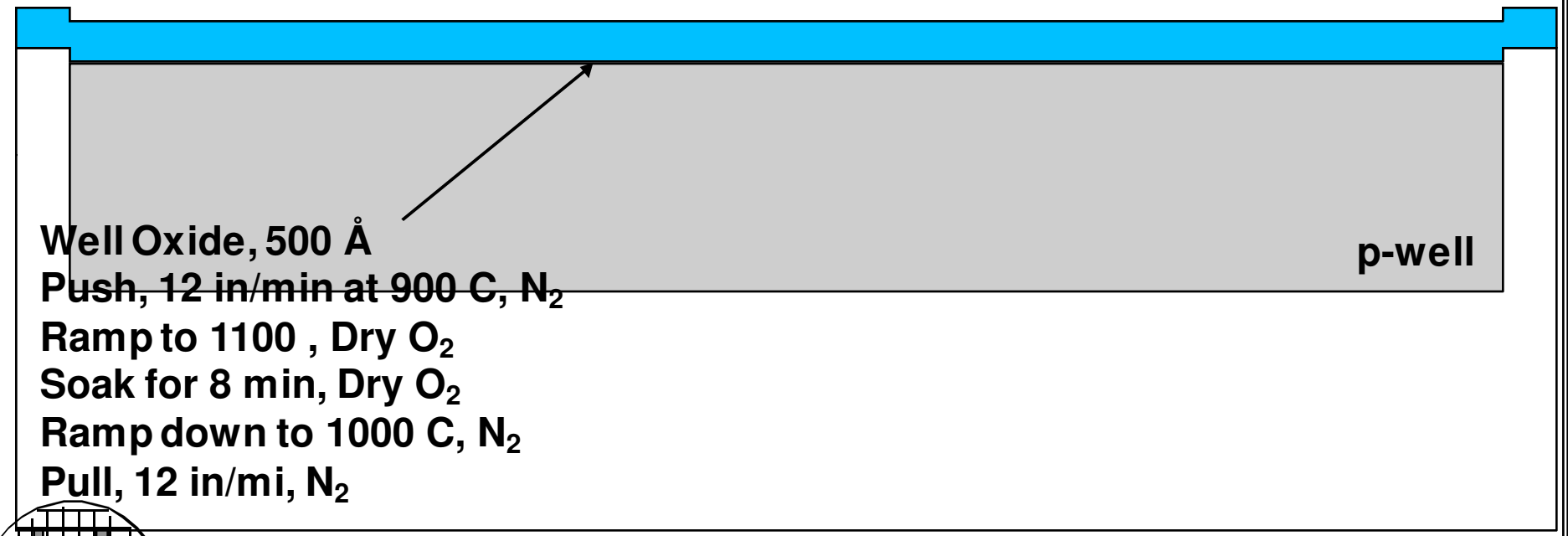
OXIDE ETCH

EEPROM **NMOS**

925 Å alignment step height

GROW PAD OXIDE

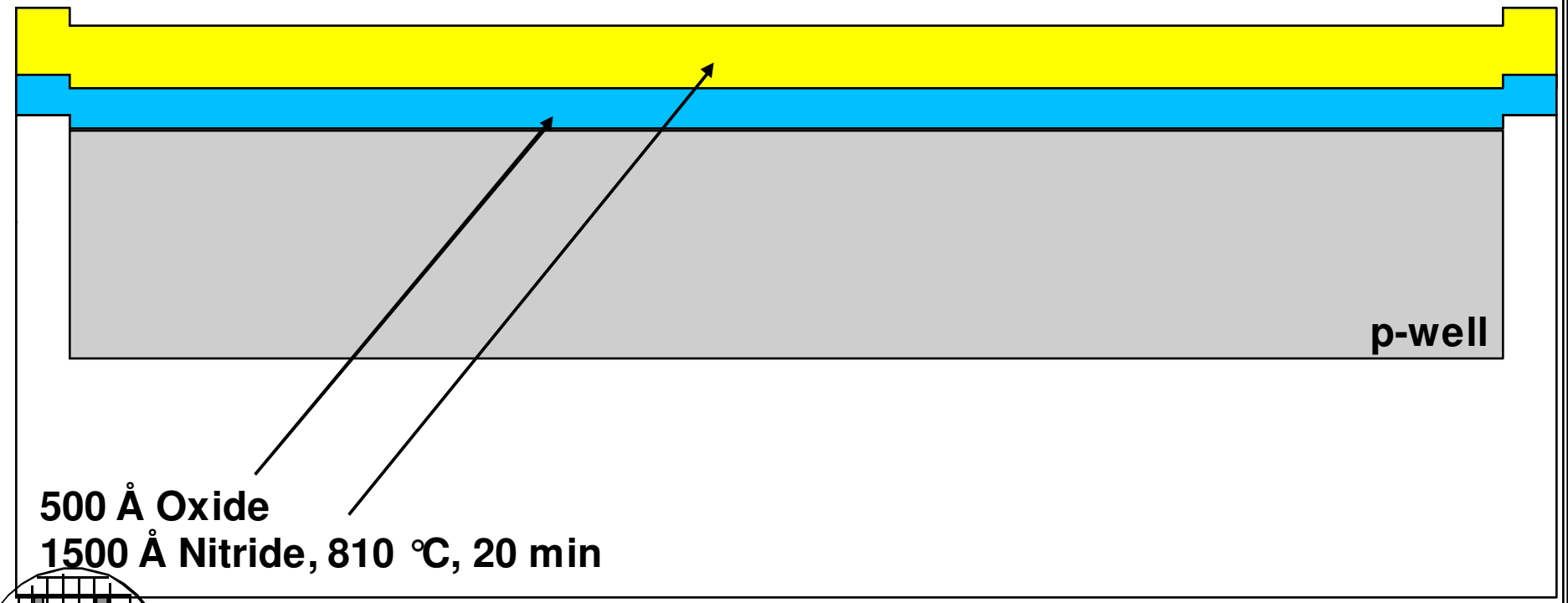
EEPROM NMOS



Well Oxide, 500 Å
~~Push, 12 in/min at 900 C, N_2~~
Ramp to 1100 , Dry O_2
Soak for 8 min, Dry O_2
Ramp down to 1000 C, N_2
Pull, 12 in/mi, N_2

LPCVD SILICON NITRIDE

EEPROM NMOS



LEVEL 2 PHOTO - ACTIVE

EEPROM NMOS

NITRIDE PLASMA ETCH

EEPROM NMOS

SF6, 300 mTorr, 30 sccm
50 W, 60 seconds

STRIP RESIST

EEPROM NMOS

LEVEL 3 PHOTO - CHANNEL STOP / FIELD VT ADJ

EEPROM **NMOS**

IMPLANT CHANNEL STOPS

11B, 100 keV, 8E13

STRIP RESIST

EEPROM NMOS

OXIDE ETCH

EEPROM NMOS

1 min in BOE HF

GROW FIELD OXIDE

EEPROM **NMOS**

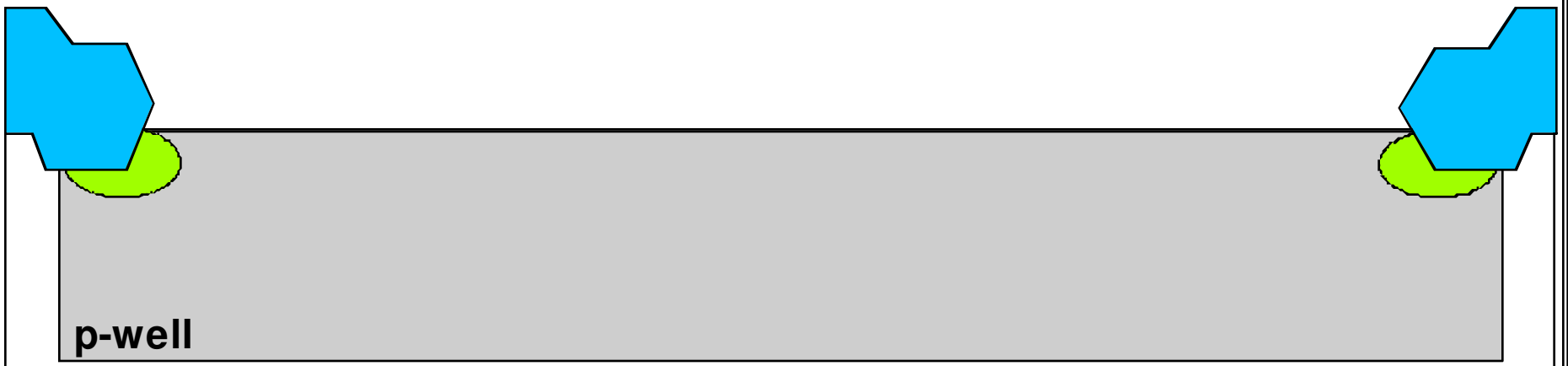
Push in 900C, ramp to 1100C in O2,
soak 210 min wet O2,
ramp down to 1000C and pull.

NITRIDE ETCH

EEPROM NMOS

OXIDE ETCH

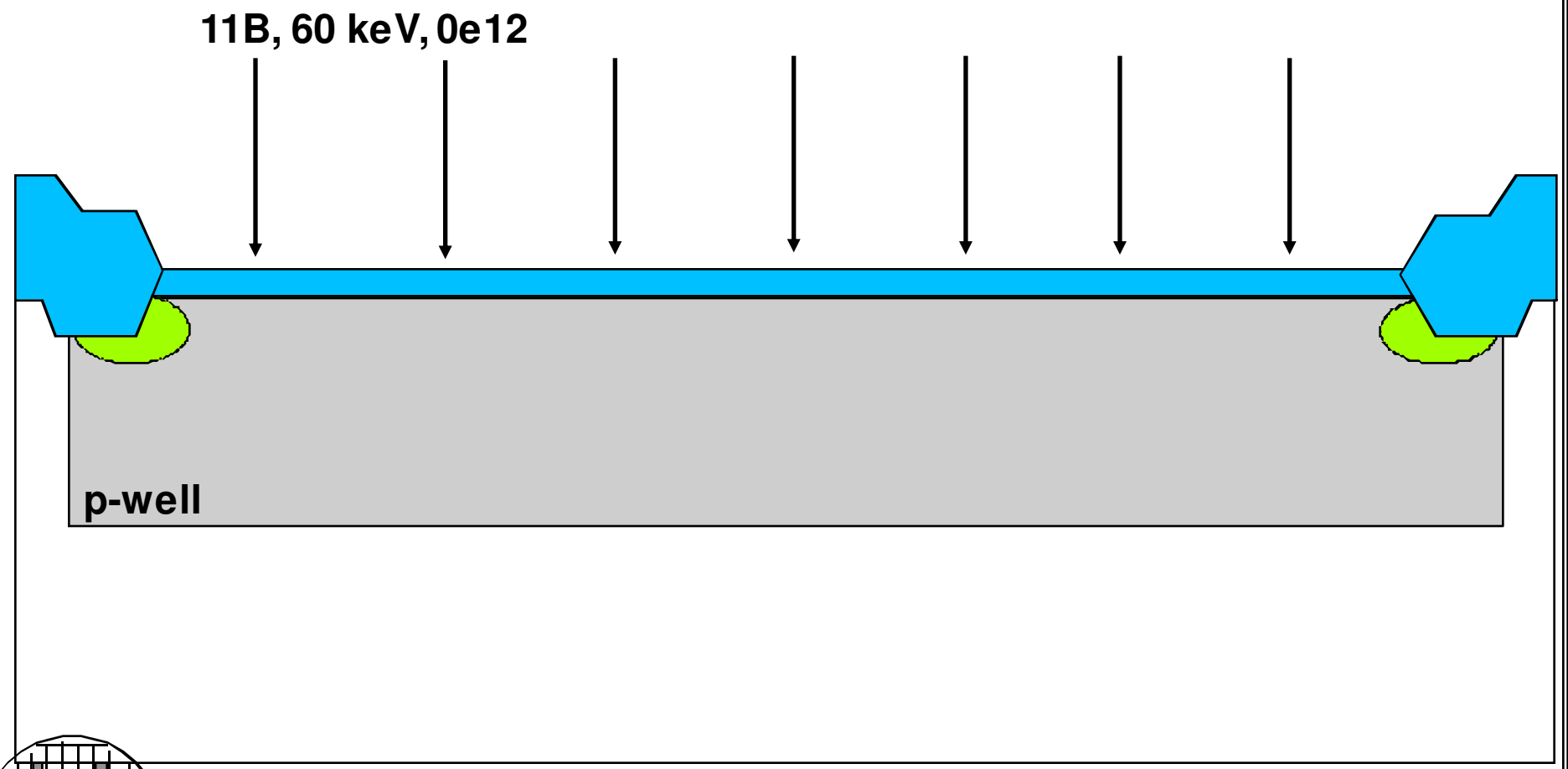
EEPROM NMOS



GROW KOOI OXIDE

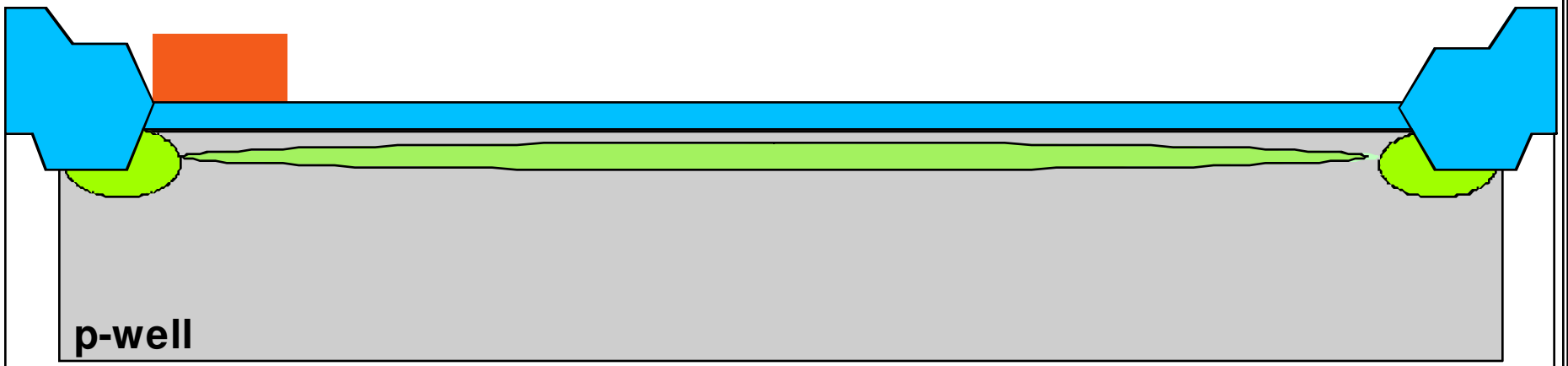
EEPROM **NMOS**

IMPLANT BLANKET Vt ADJ

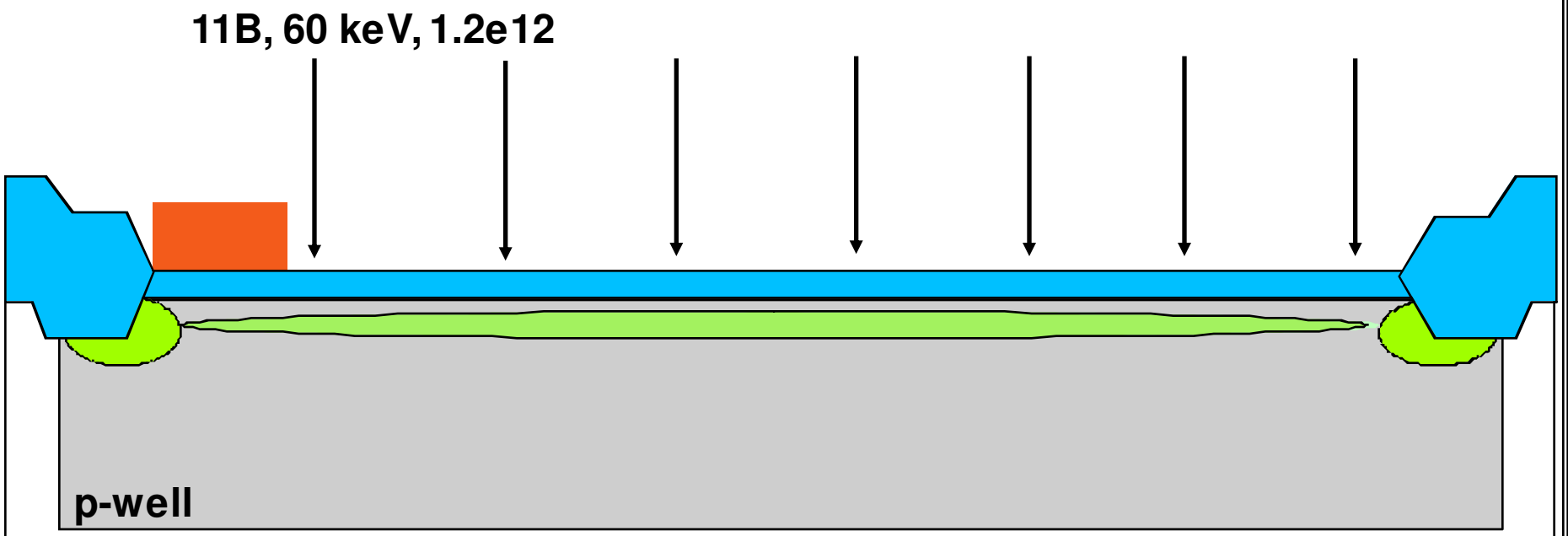


LEVEL 4 PHOTO - NMOS Vt ADJ

EEPROM **NMOS**

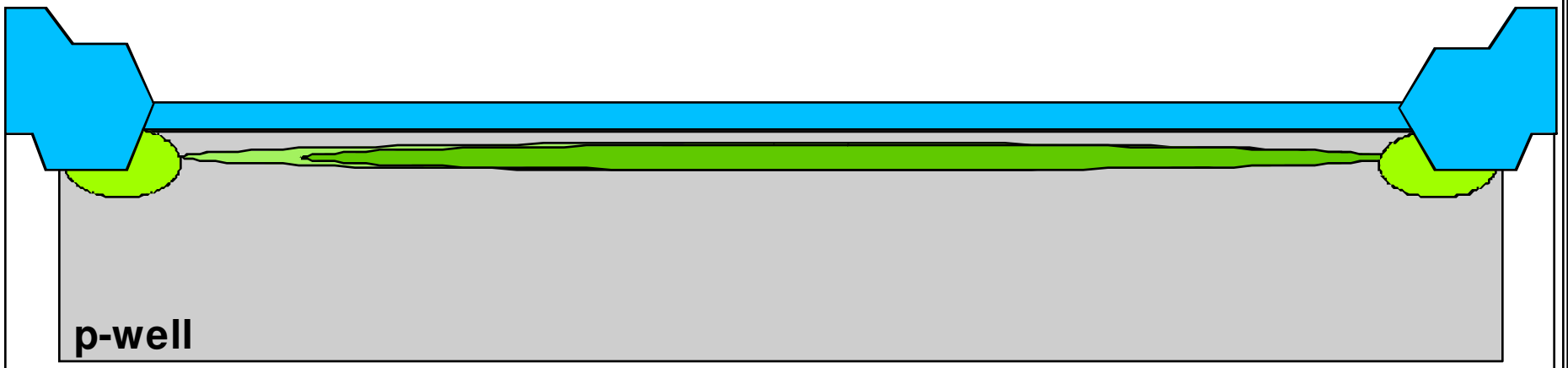


IMPLANT - NMOS Vt ADJ



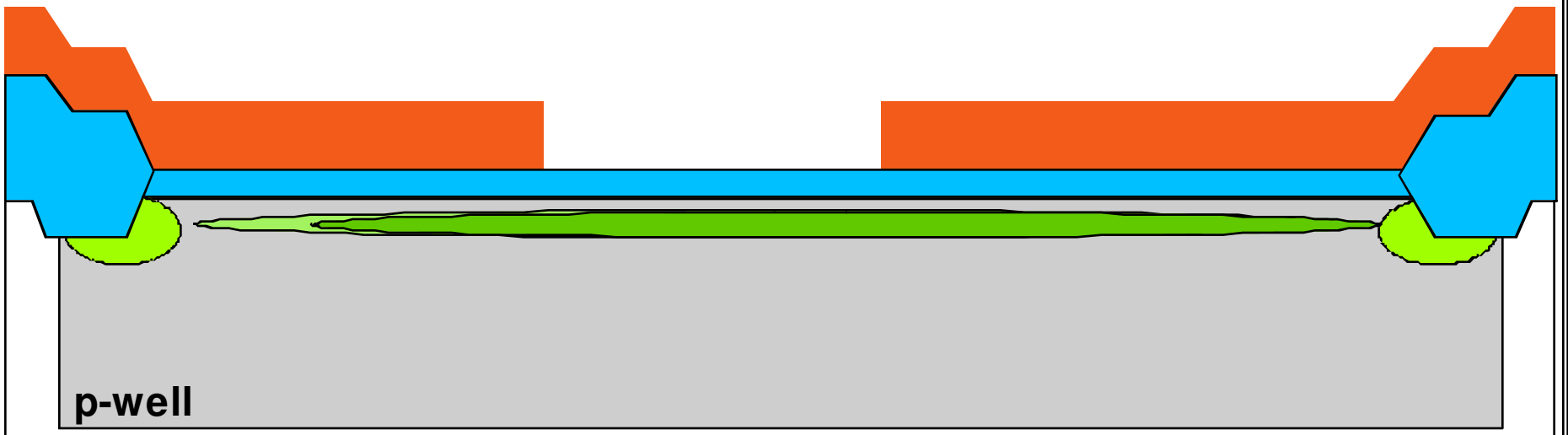
STRIP RESIST

EEPROM NMOS



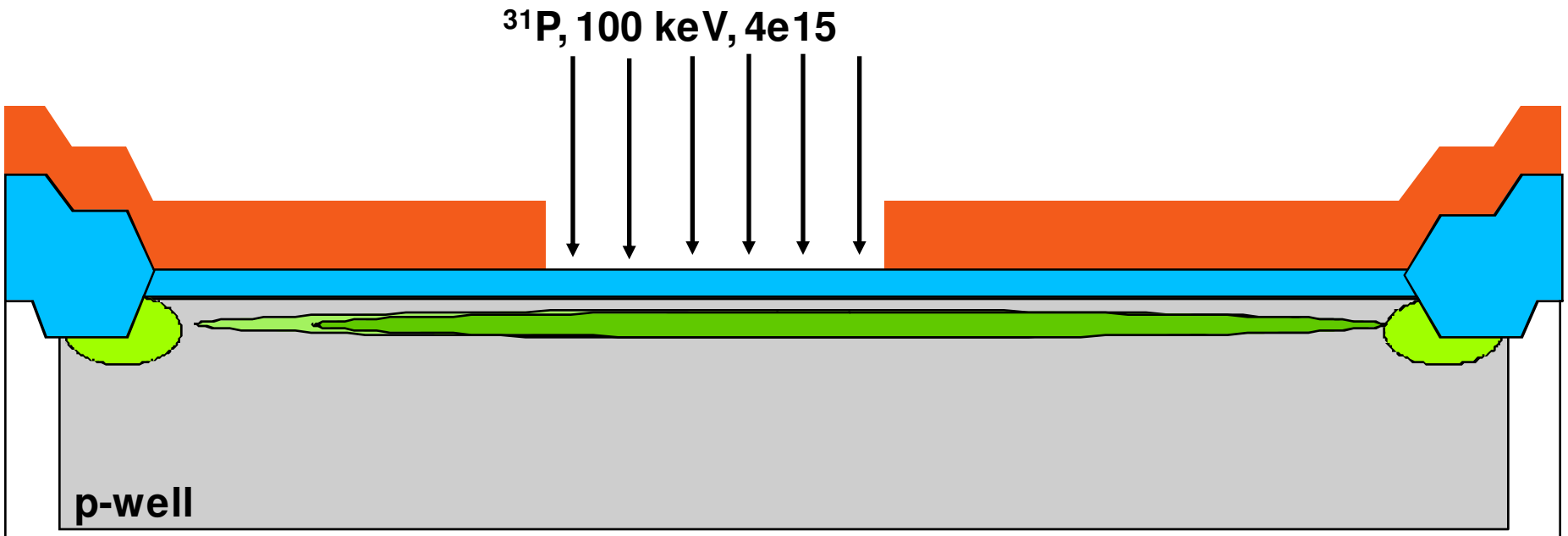
LEVEL 5 PHOTO - N+ DRAIN

EEPROM NMOS



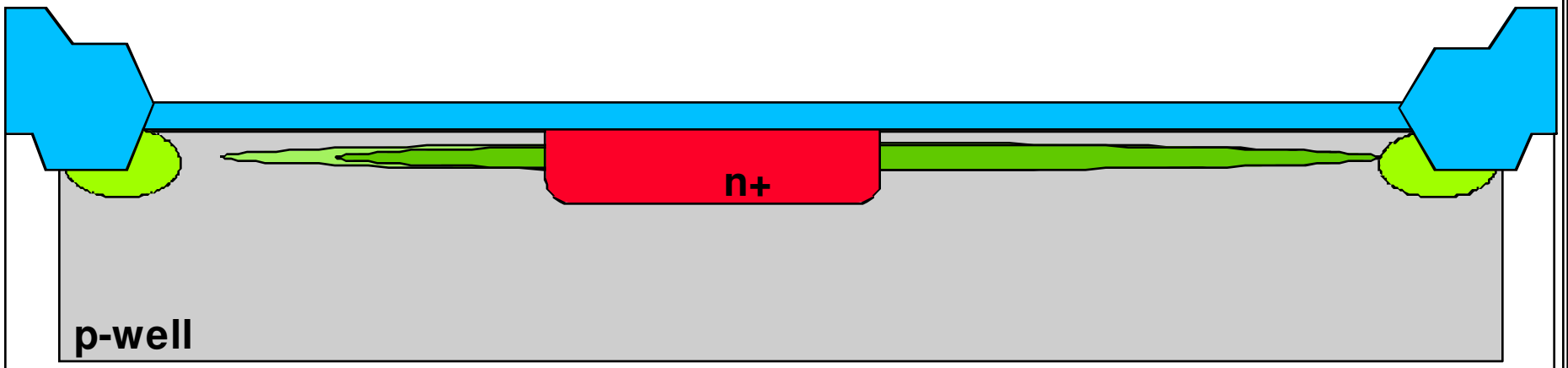
IMPLANT N+ DRAIN

EEPROM NMOS



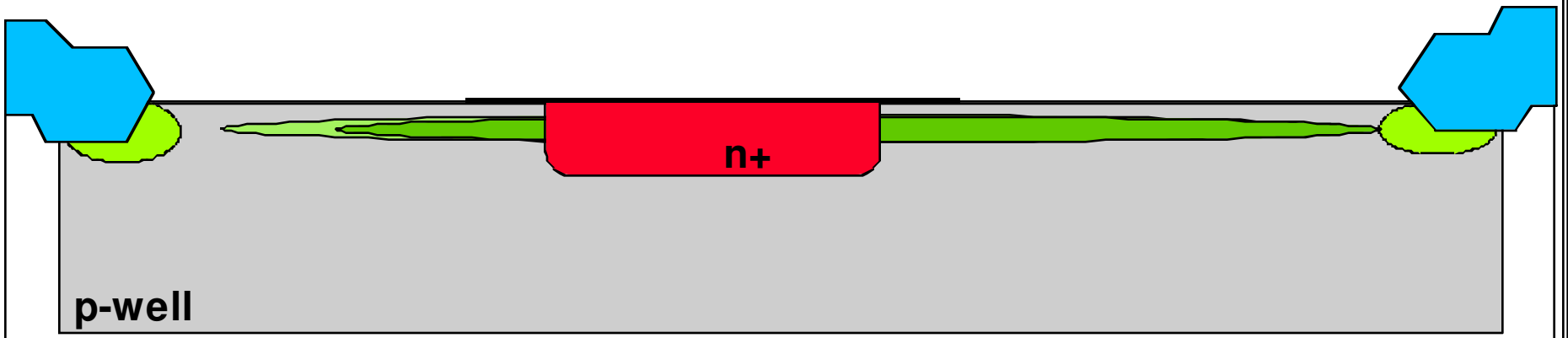
STRIP RESIST

EEPROM NMOS



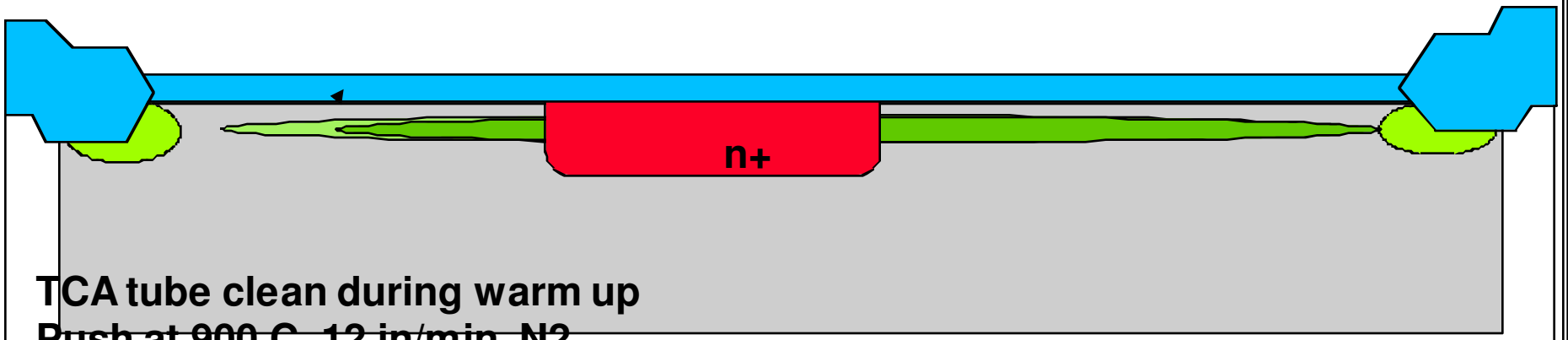
OXIDE ETCH

EEPROM NMOS



GROW GATE OXIDE

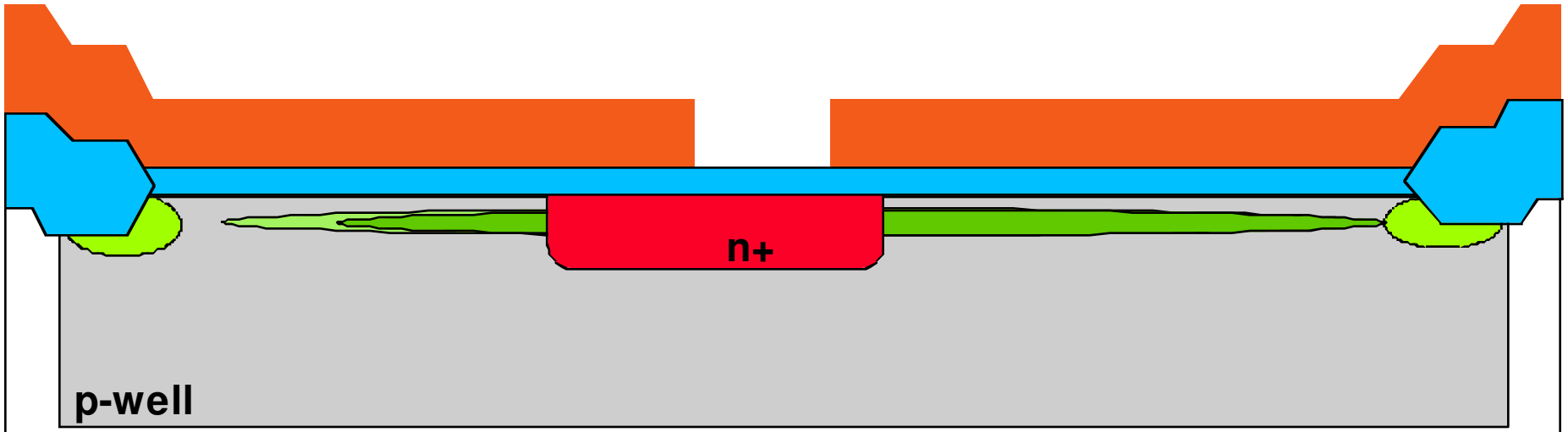
EEPROM **NMOS**



TCA tube clean during warm up
Push at 900 C, 12 in/min, N2
Ramp up to 1100 °C, dry O2
Soak 1100 C, dry O2, 8 min
Pull at 900 C, 12 in/min, N2
Xox desired = 500 Å

PHOTO 6 - TUNNEL OXIDE

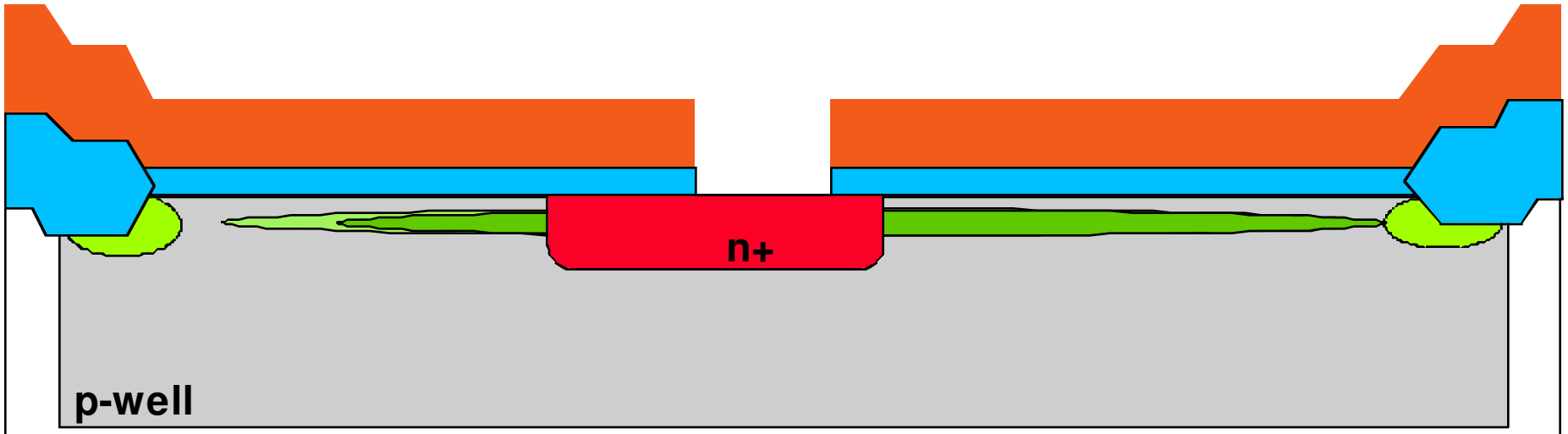
EEPROM NMOS



ET06

EEPROM

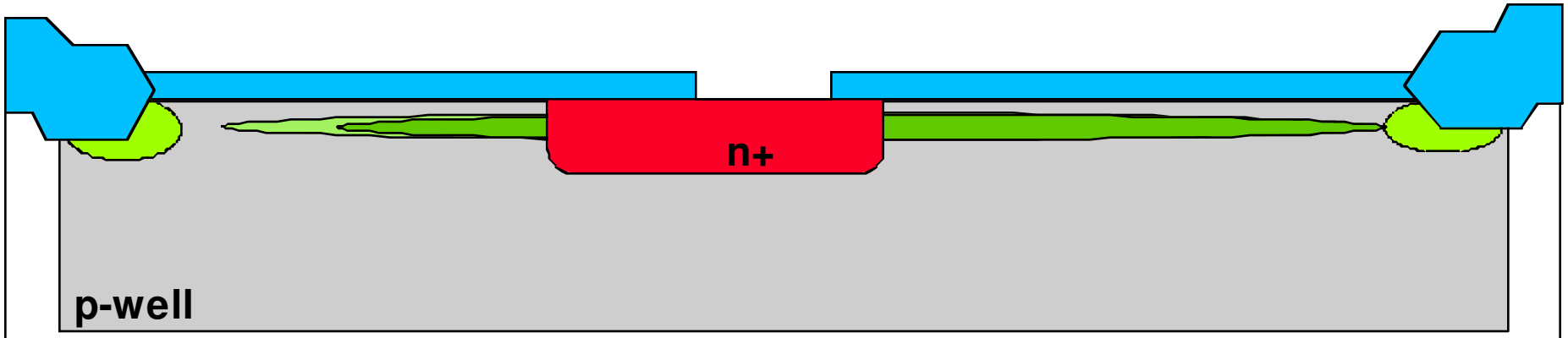
NMOS



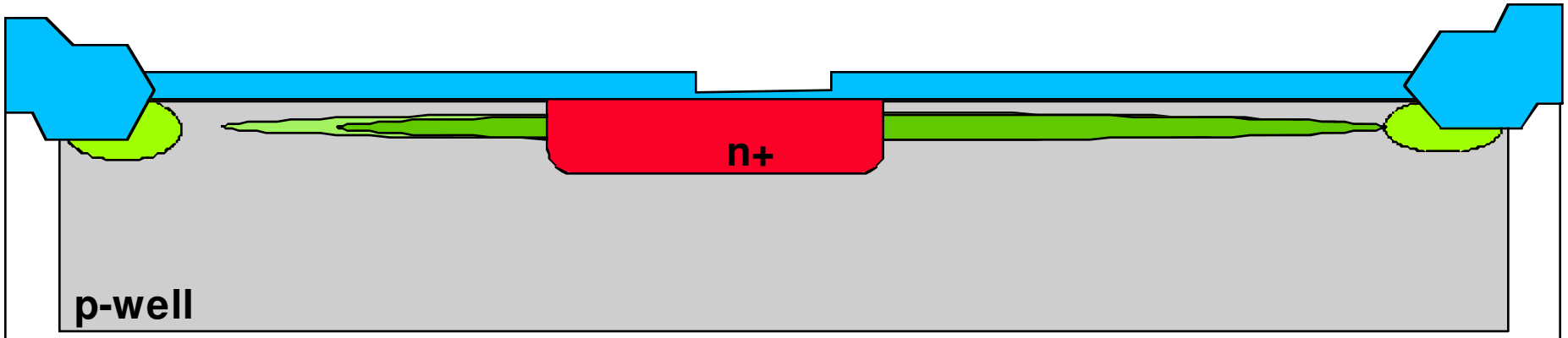
ET07

EEPROM

NMOS

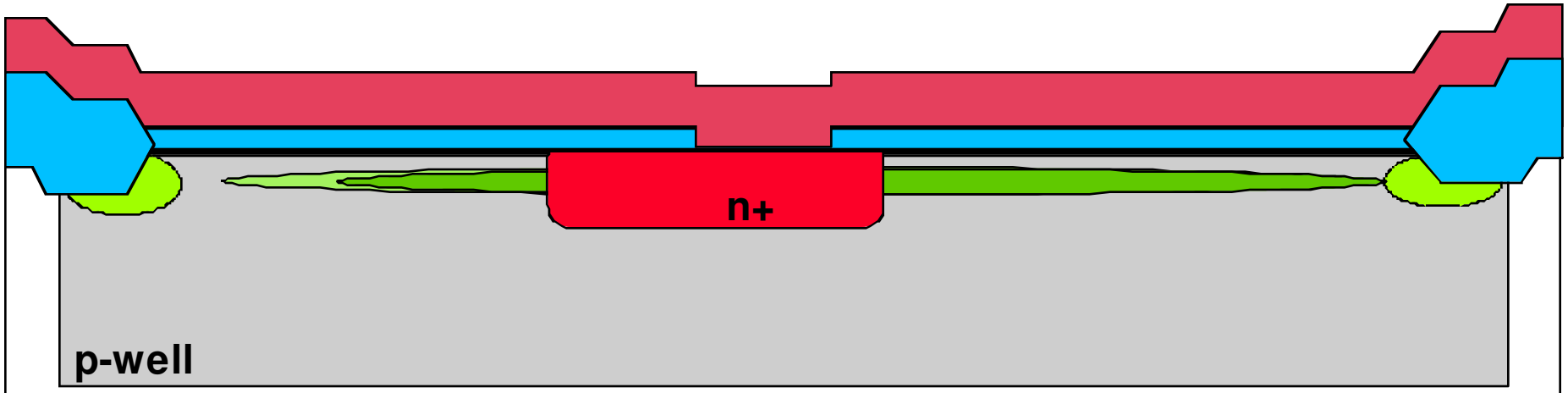


OX06-TUNOX NMOS EEPROM



POLYSILICON DEPOSITION

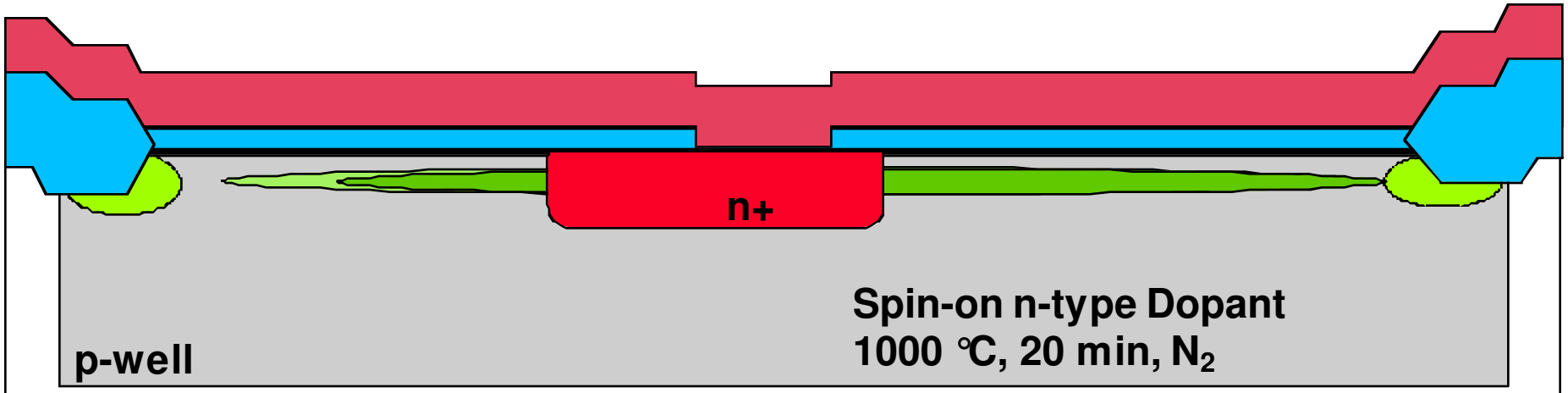
EEPROM NMOS



LPCVD Polysilicon
0.6 micrometers
610 ℃, 60 min

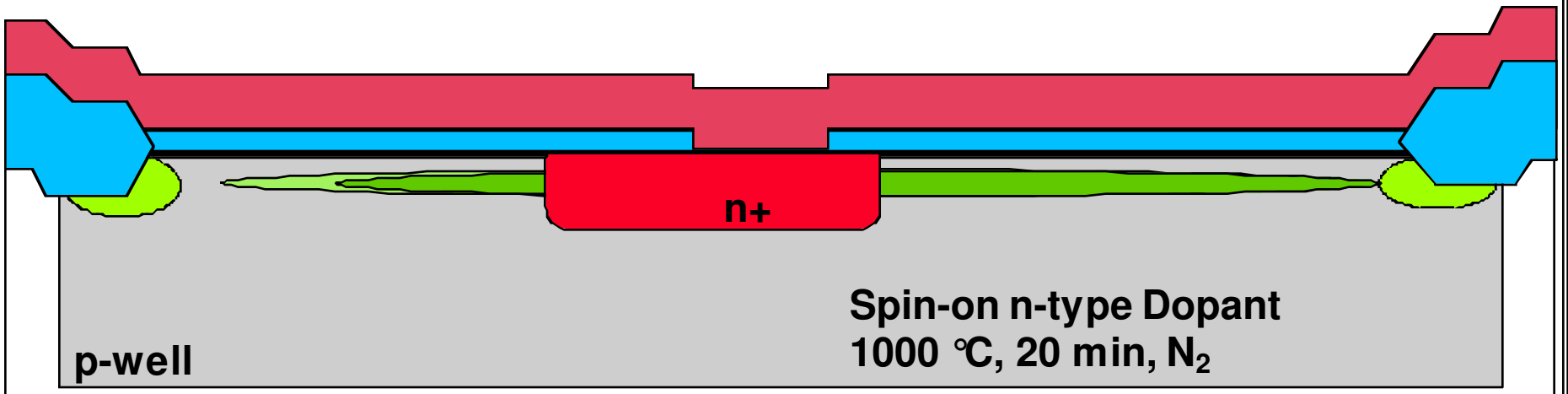
POLYSILICON DOPING

EEPROM NMOS



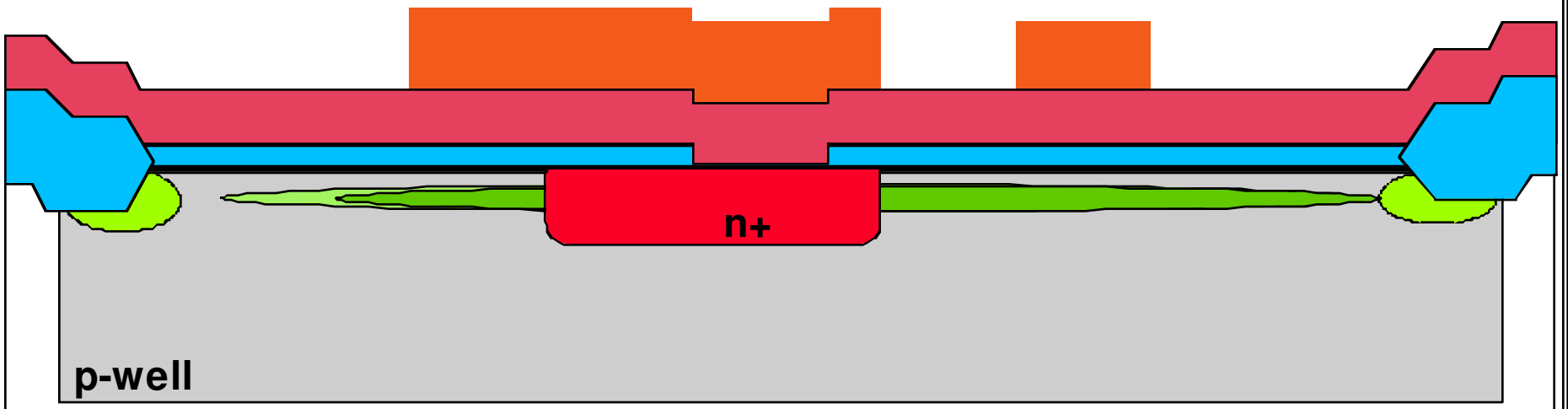
ET06

EEPROM NMOS



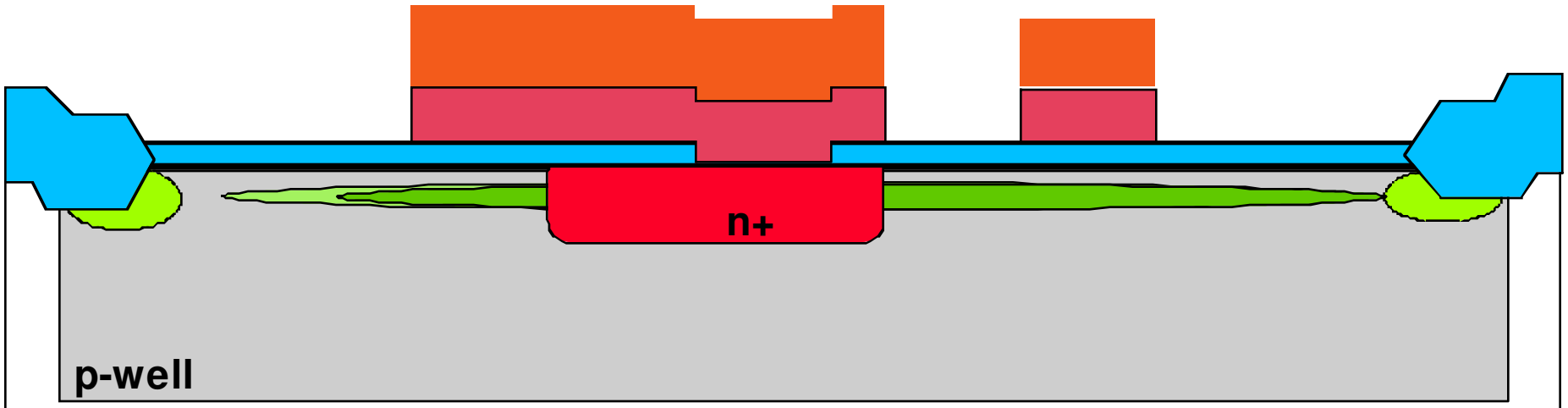
LEVEL 5 PHOTO - POLYSILICON

EEPROM NMOS



PLASMA ETCH POLYSILICON

EEPROM NMOS

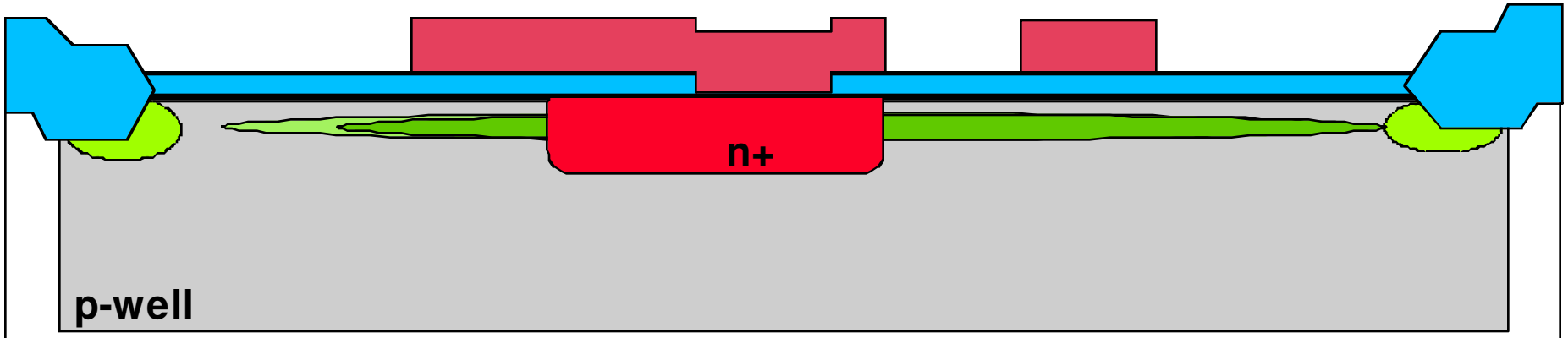


SF6 + O2, 30 sccm + 3 sccm, 75 mT, 75 watts

ET07

EEPROM

NMOS

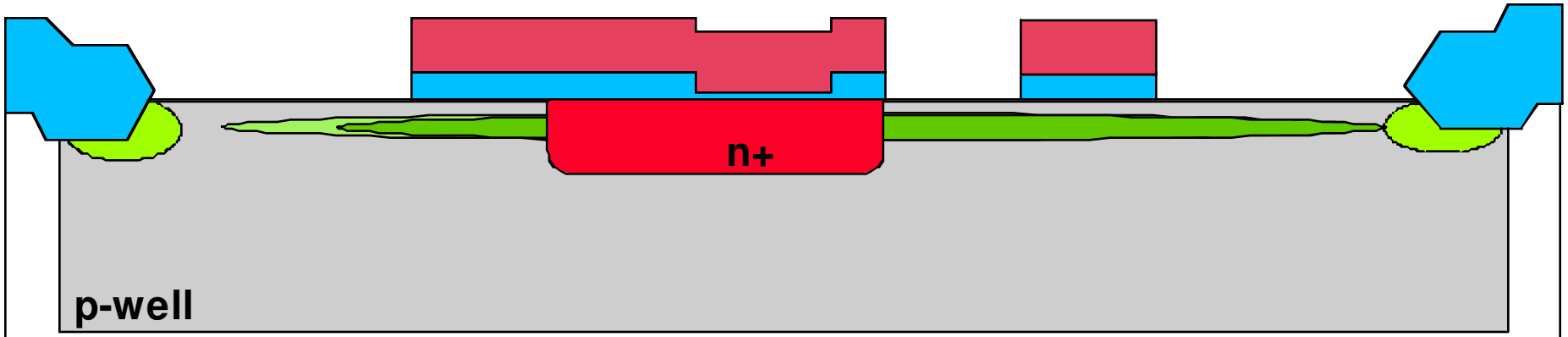


SF6 + O2, 30 sccm + 3 sccm, 75 mT, 75 watts

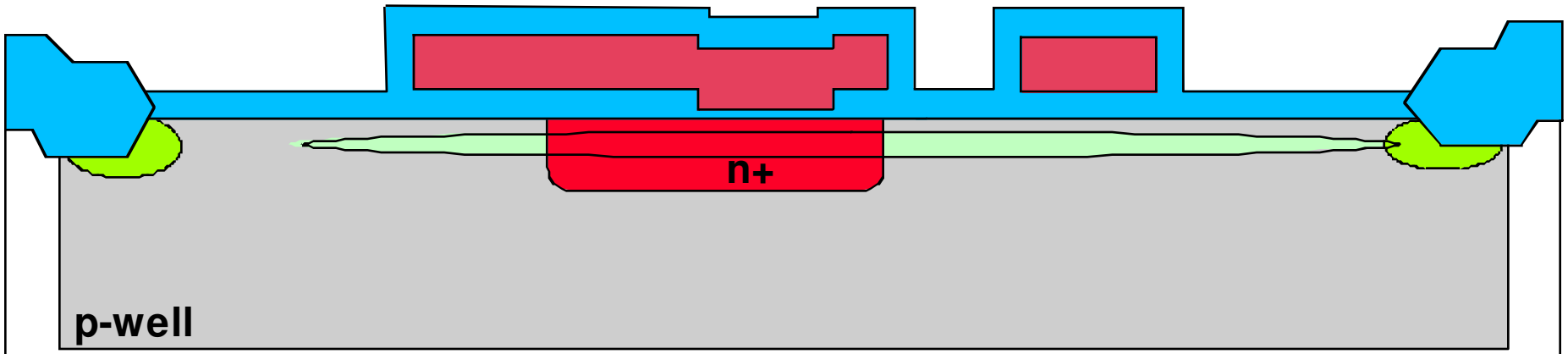
ET06

EEPROM

NMOS



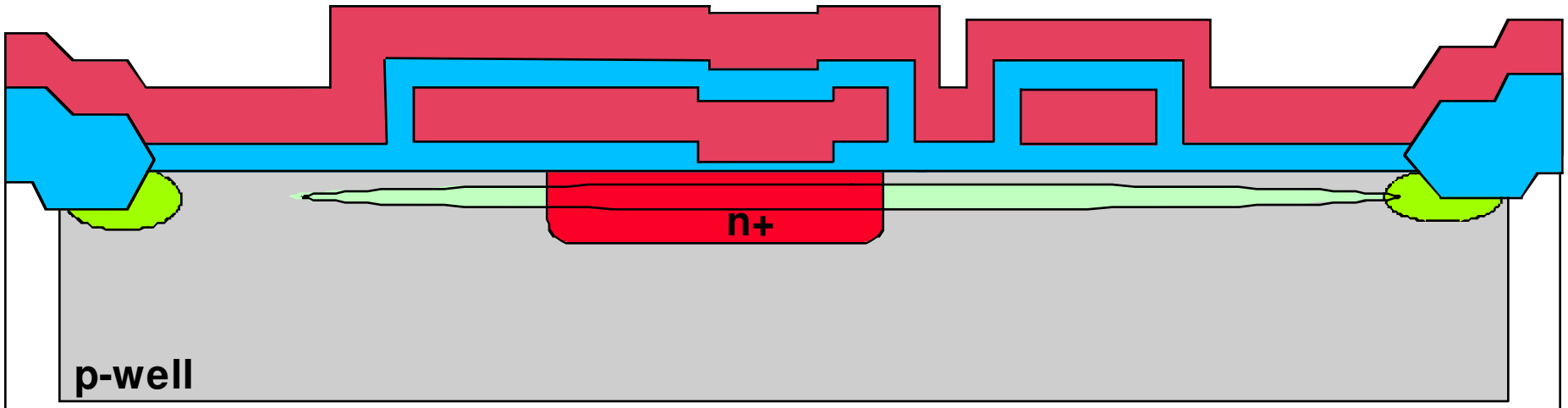
CL01 & OX06 NMOS EEPROM



CV01

EEPROM

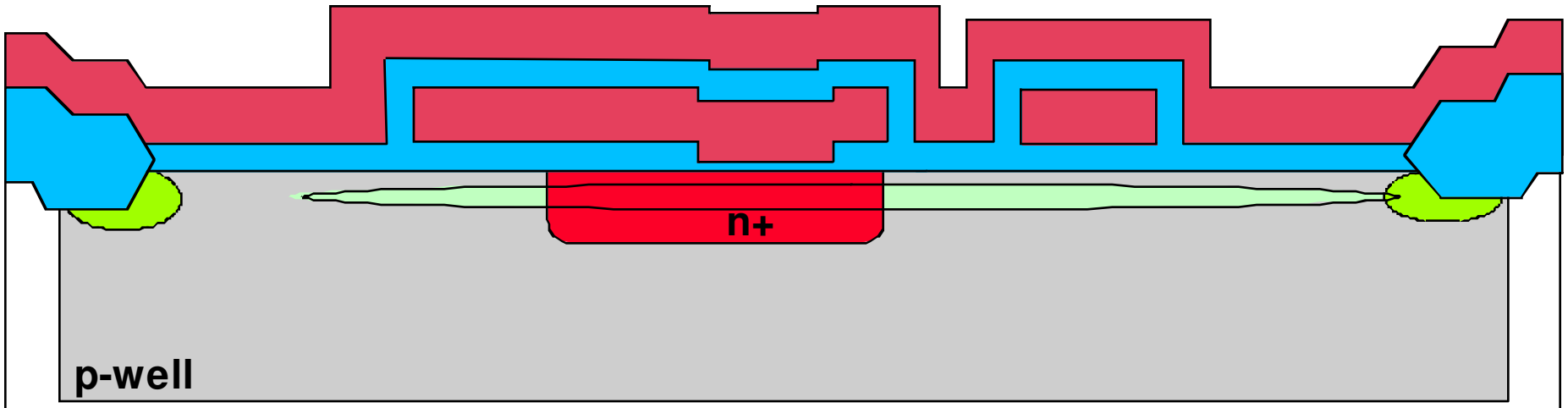
NMOS



DI04

EEPROM

NMOS



ET06

EEPROM

NMOS

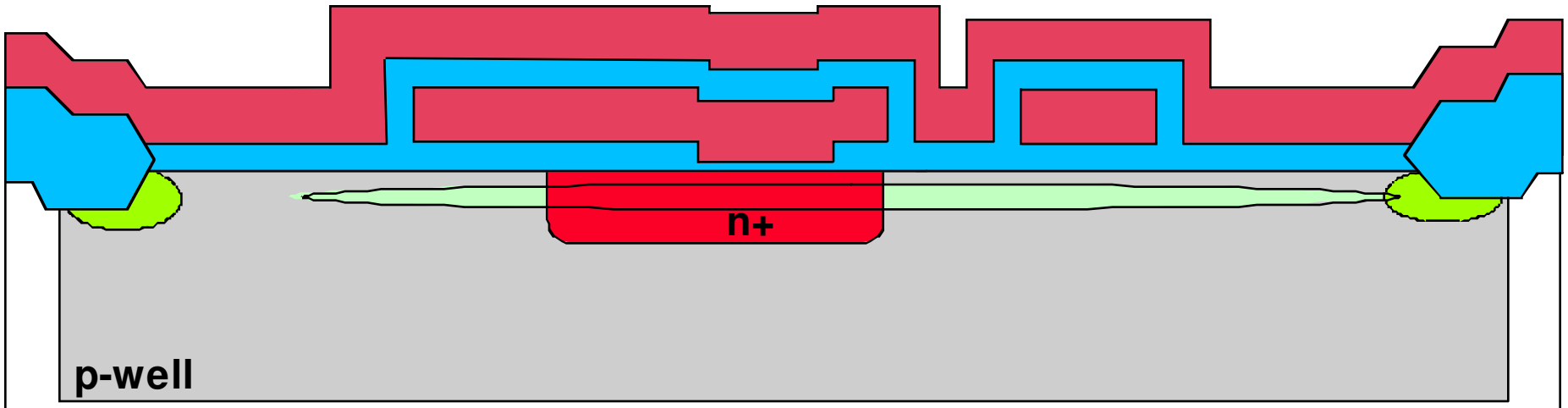
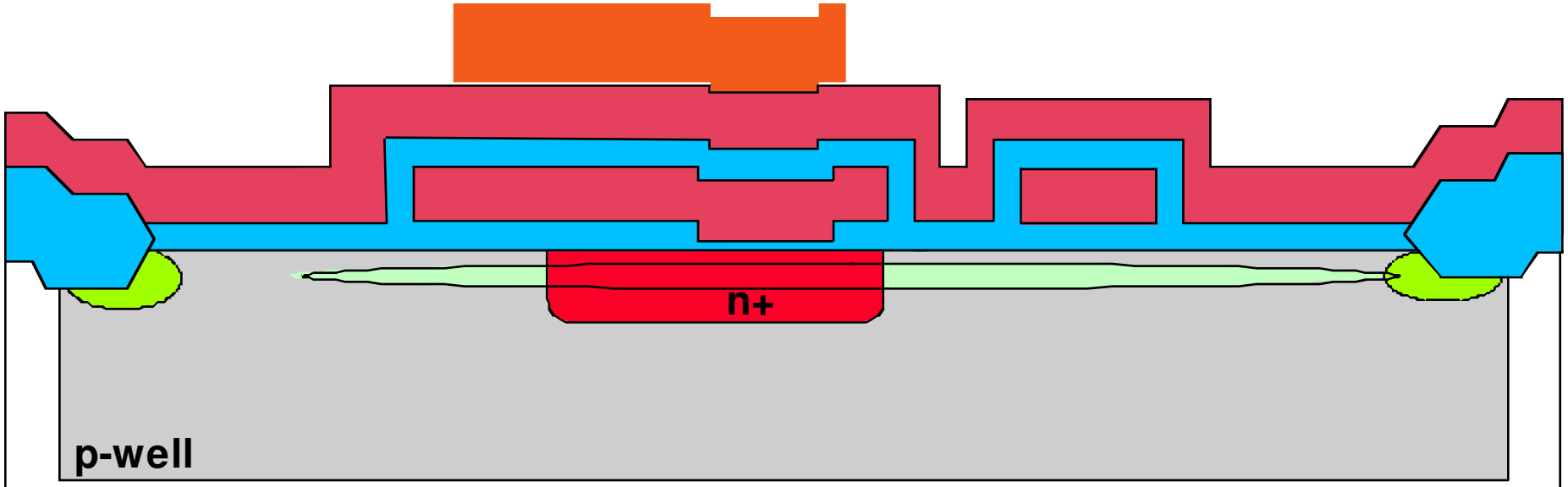


PHOTO 8 - SECOND GATE

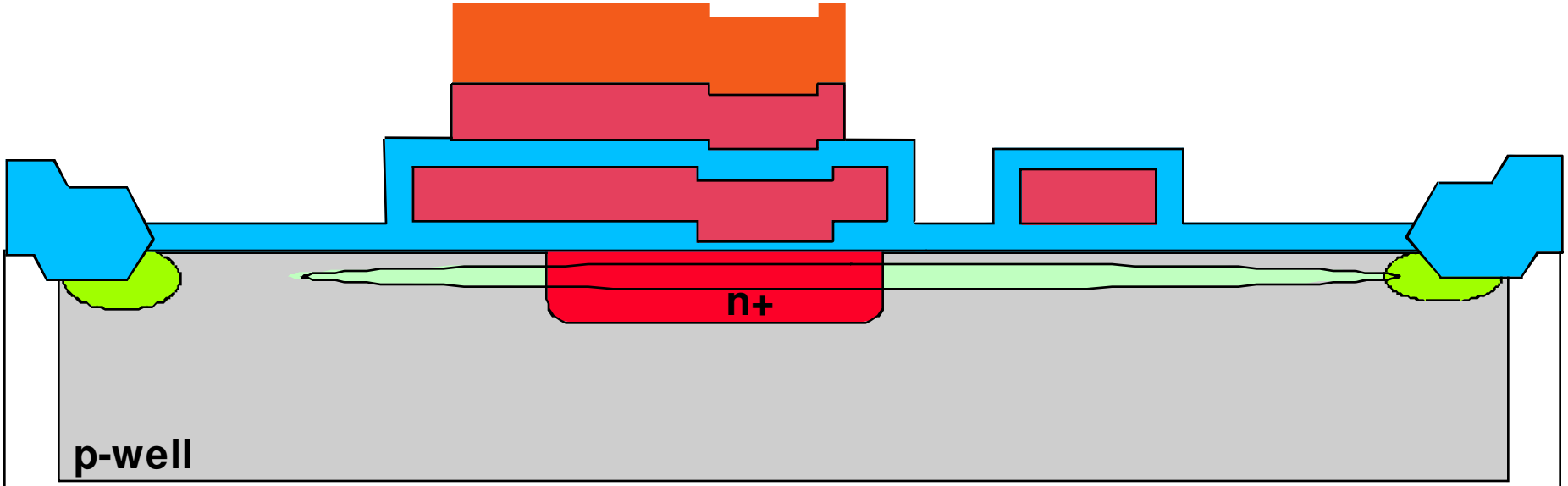
EEPROM NMOS



ET08

EEPROM

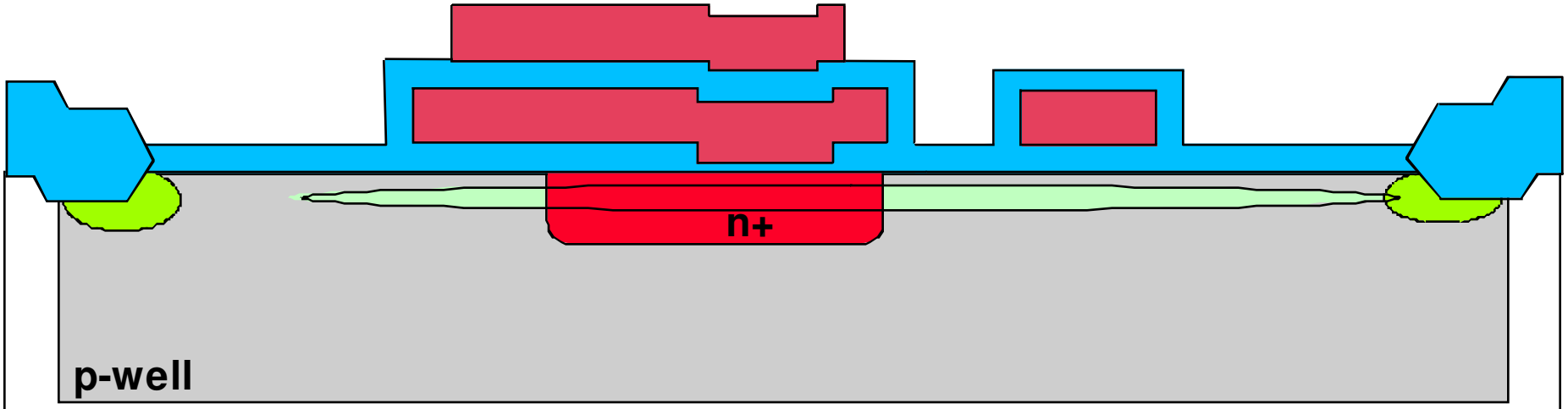
NMOS



ET07

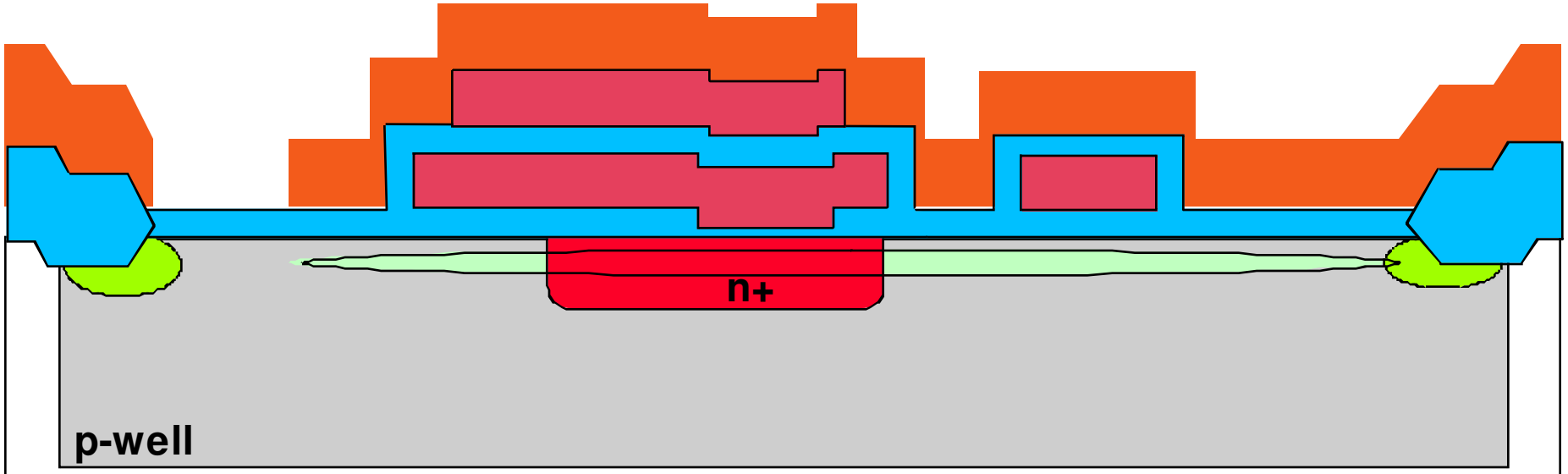
EEPROM

NMOS



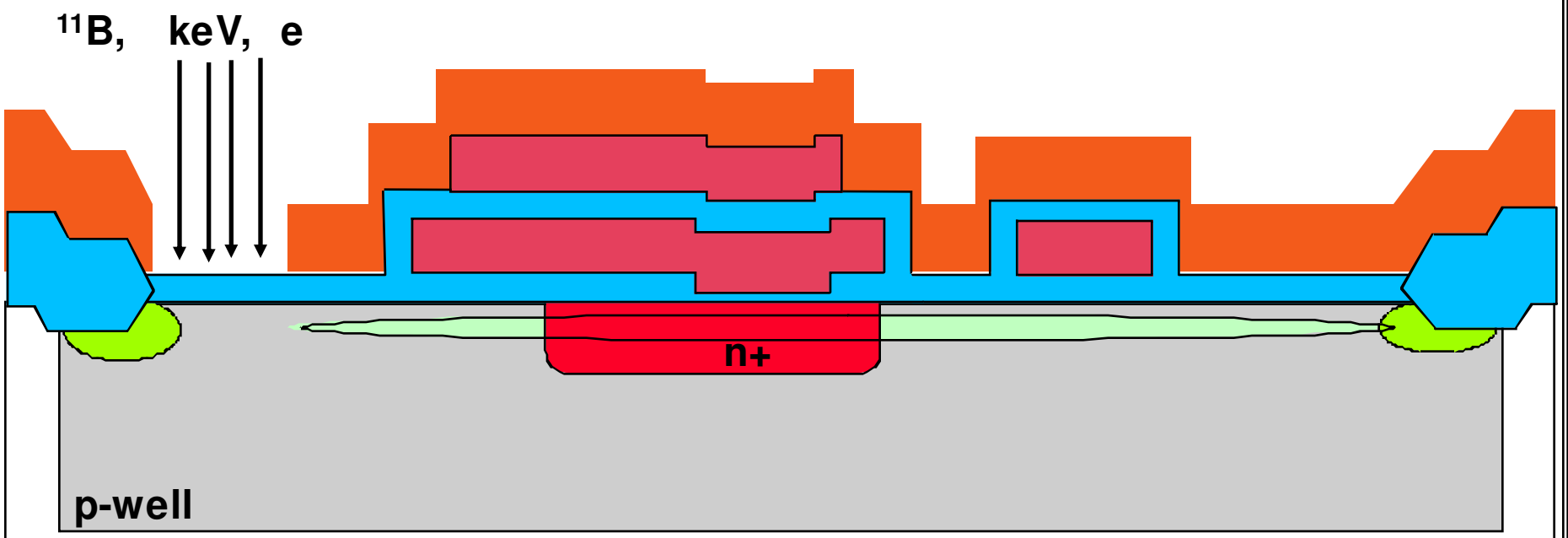
LEVEL 6 PHOTO - P+ D/S, OXIDE ETCH

EEPROM NMOS



ION IMPLANT P+ D/S

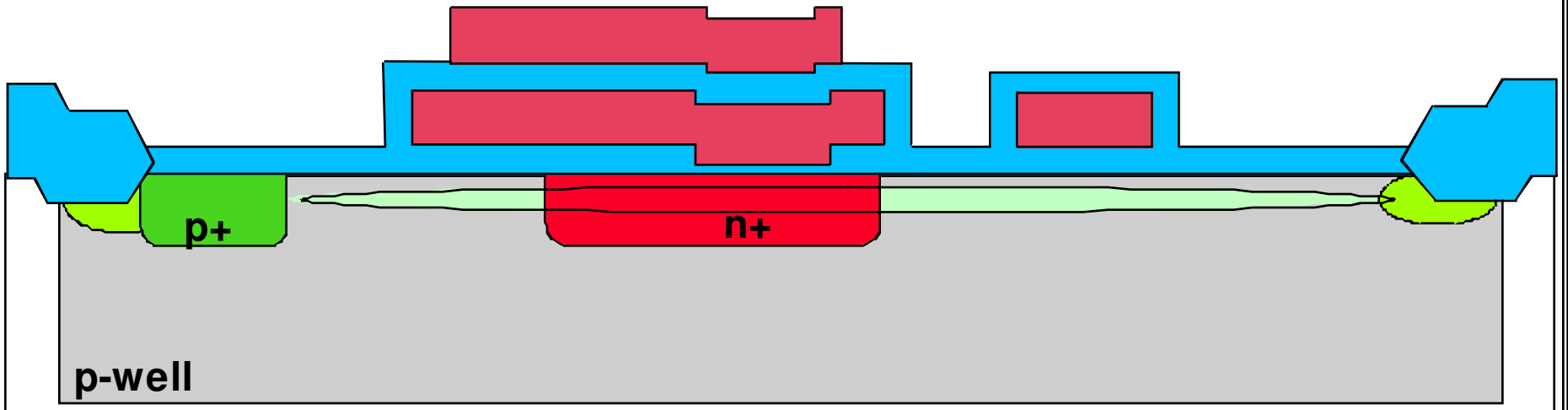
EEPROM NMOS



4e15, 120 KeV, BF_2 into bare silicon

ET07

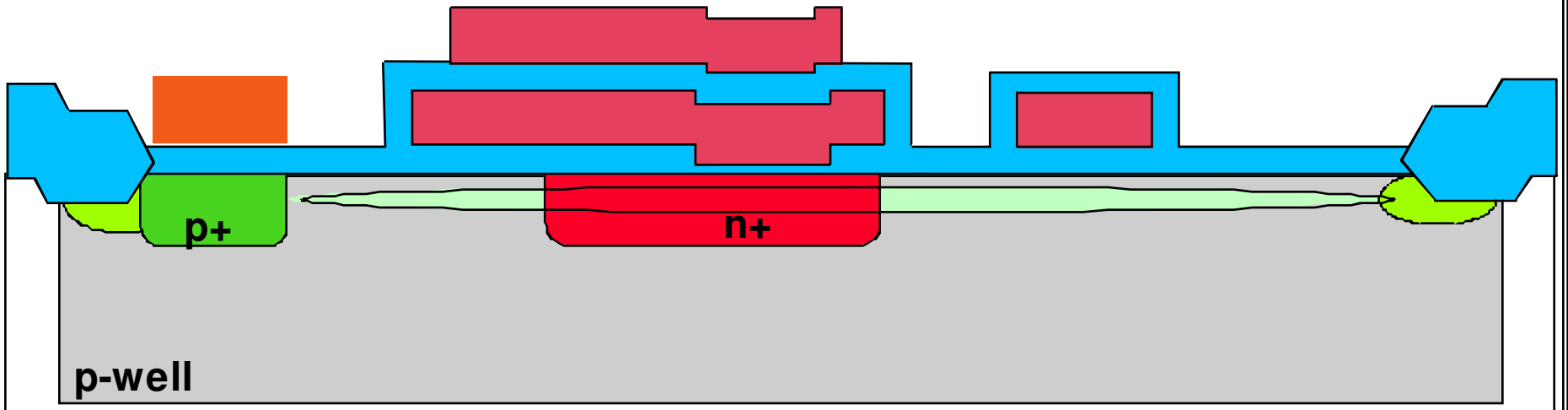
EEPROM NMOS



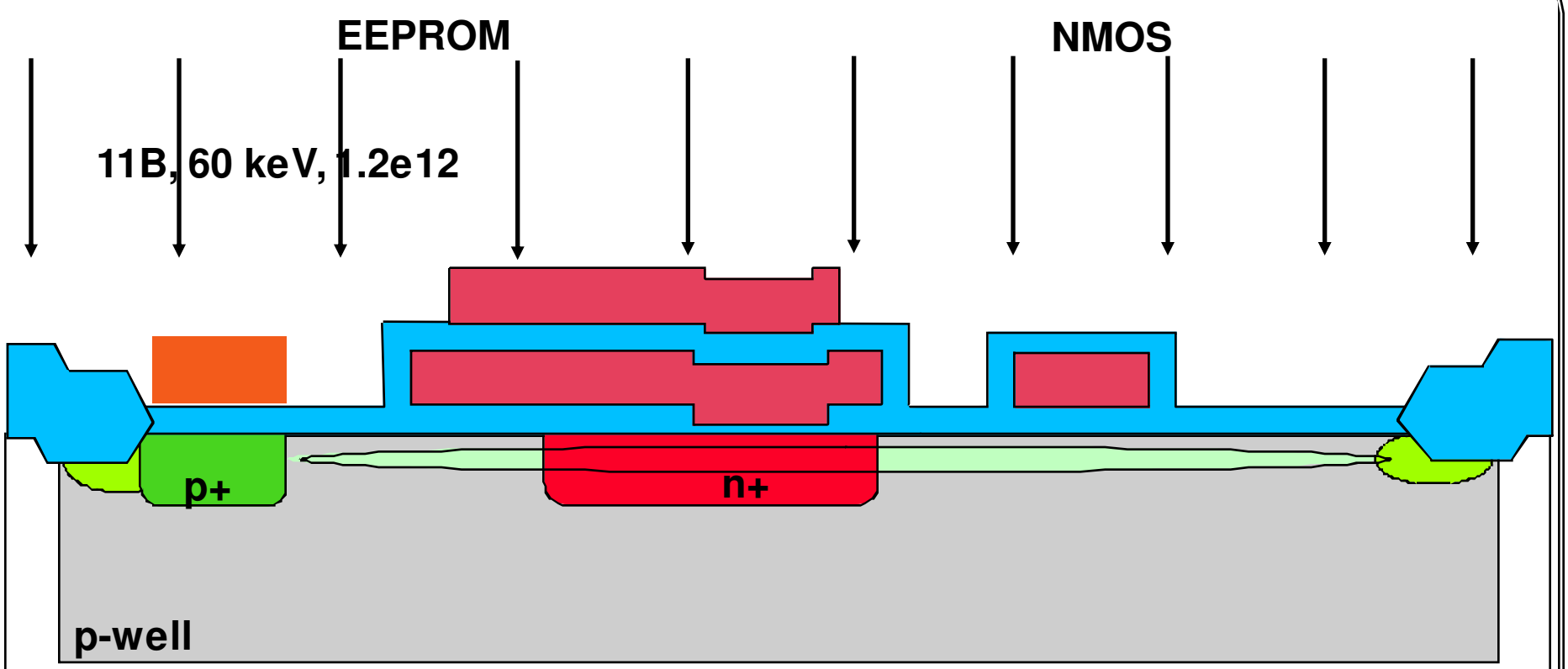
4e15, 120 KeV, BF_2 into bare silicon

LEVEL 7 PHOTO - N+ D/S

EEPROM NMOS



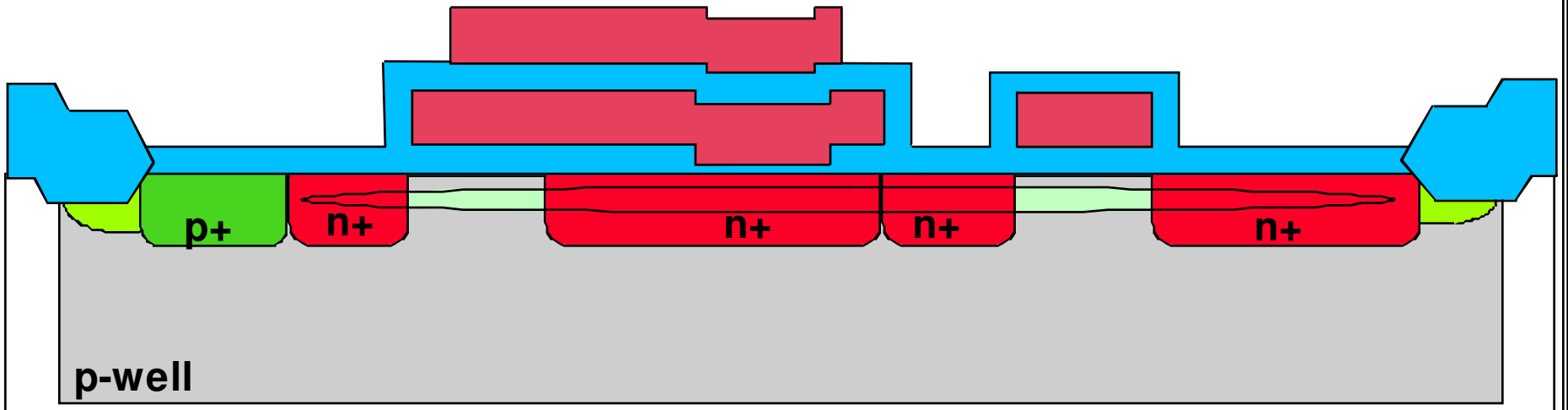
ION IMPLANT N+ D/S



4e15, 100KeV, P31 thru gate Oxide

ET07

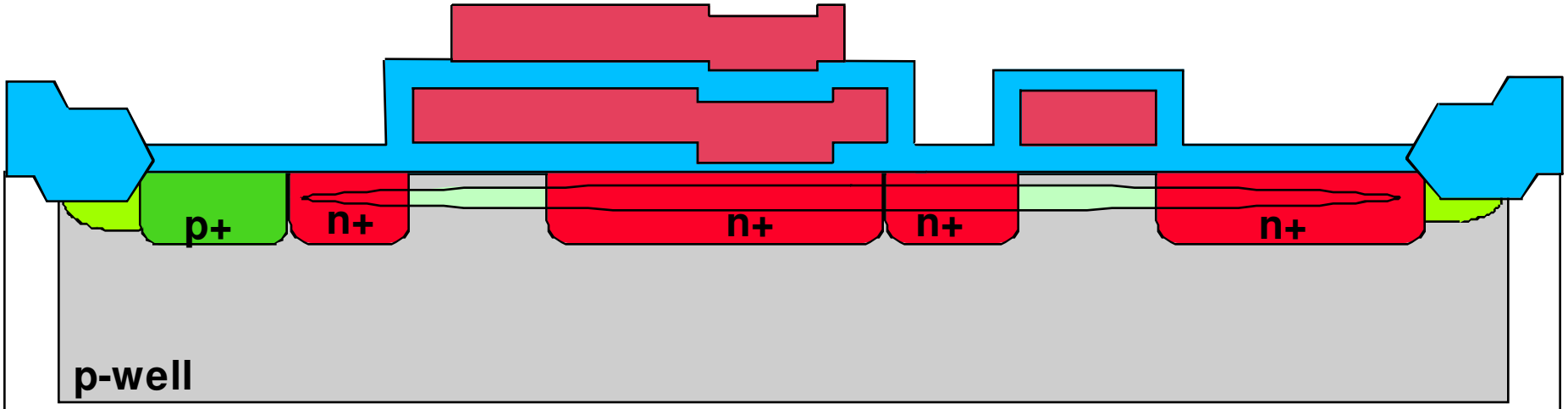
EEPROM NMOS



4e15, 100KeV, P31 thru gate Oxide

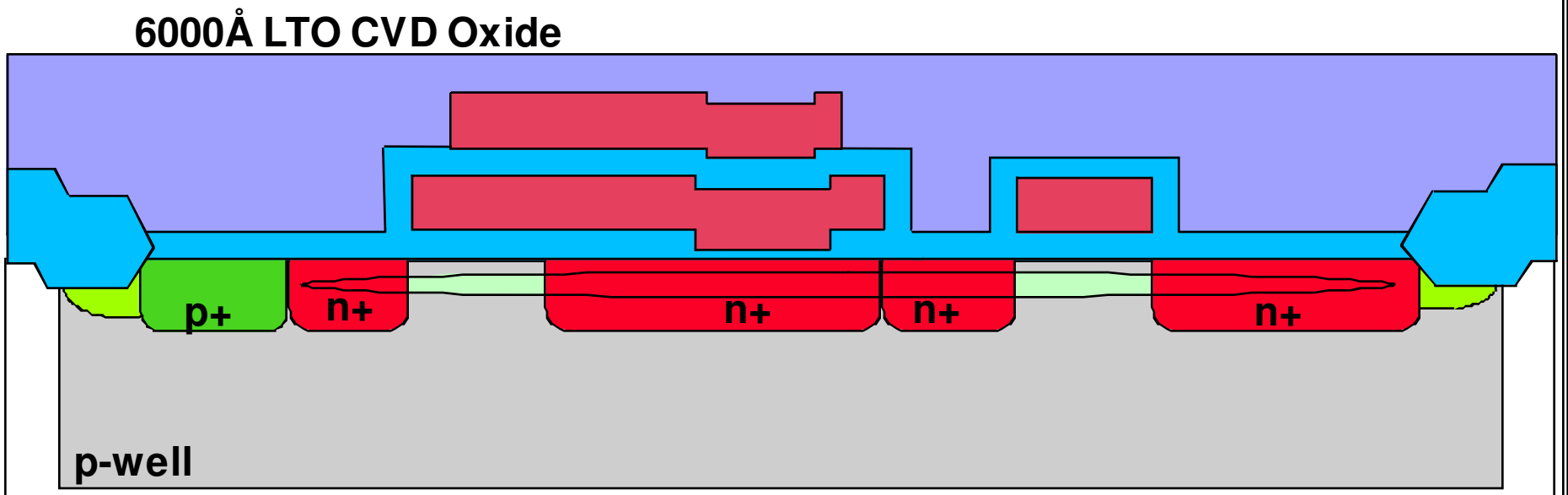
ANNEAL, 950 °C, 30 MIN, N_2

EEPROM NMOS



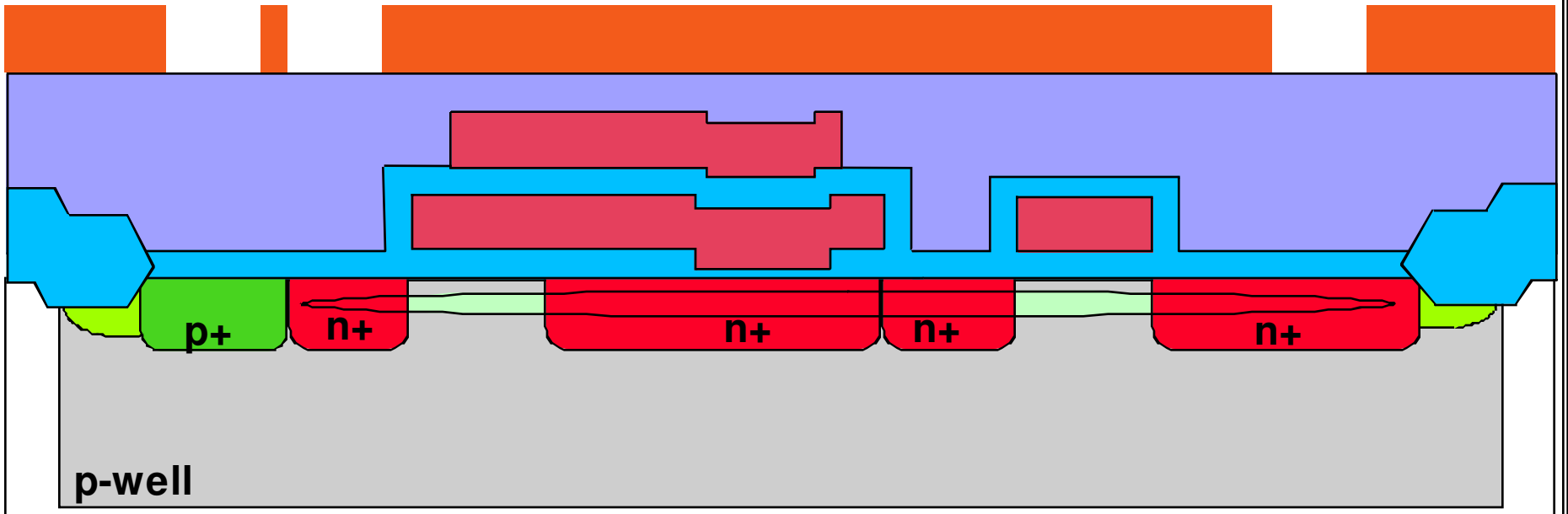
CVD (LTO) GLASS DEPOSITION

EEPROM NMOS



LEVEL 8 PHOTO - CONTACT CUT

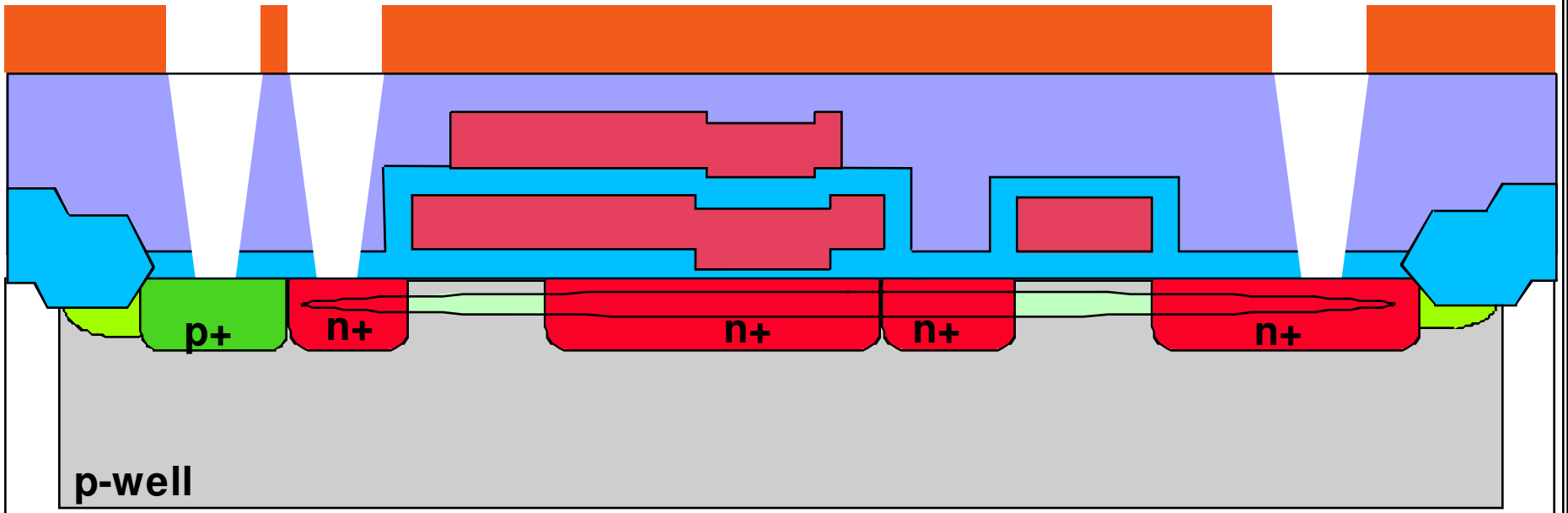
EEPROM NMOS



Etch in BHF 7min

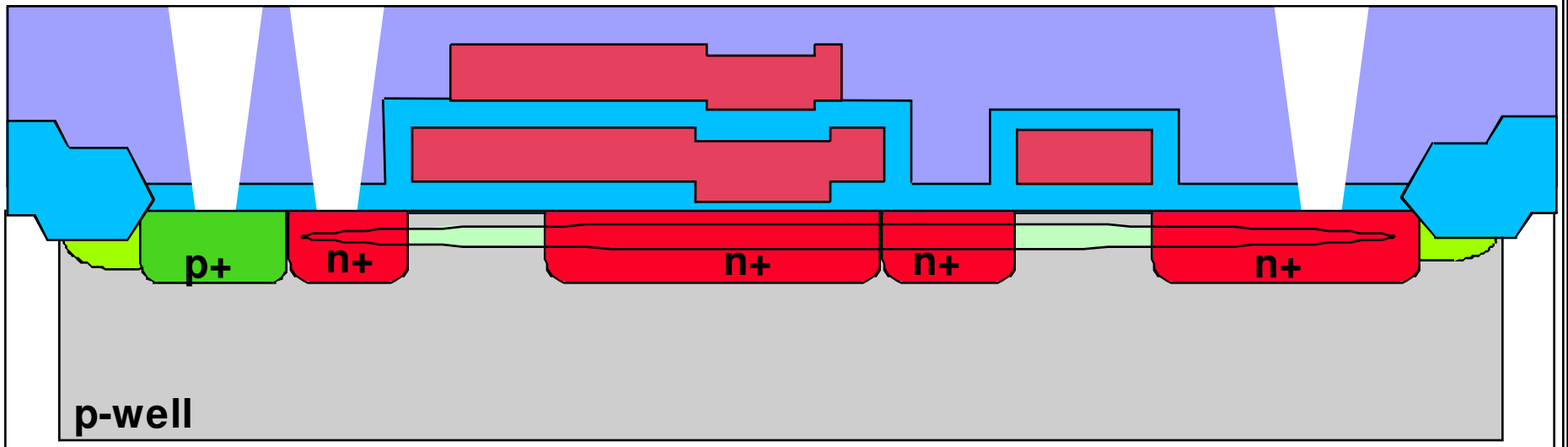
CONTACT CUT ETCH

EEPROM NMOS



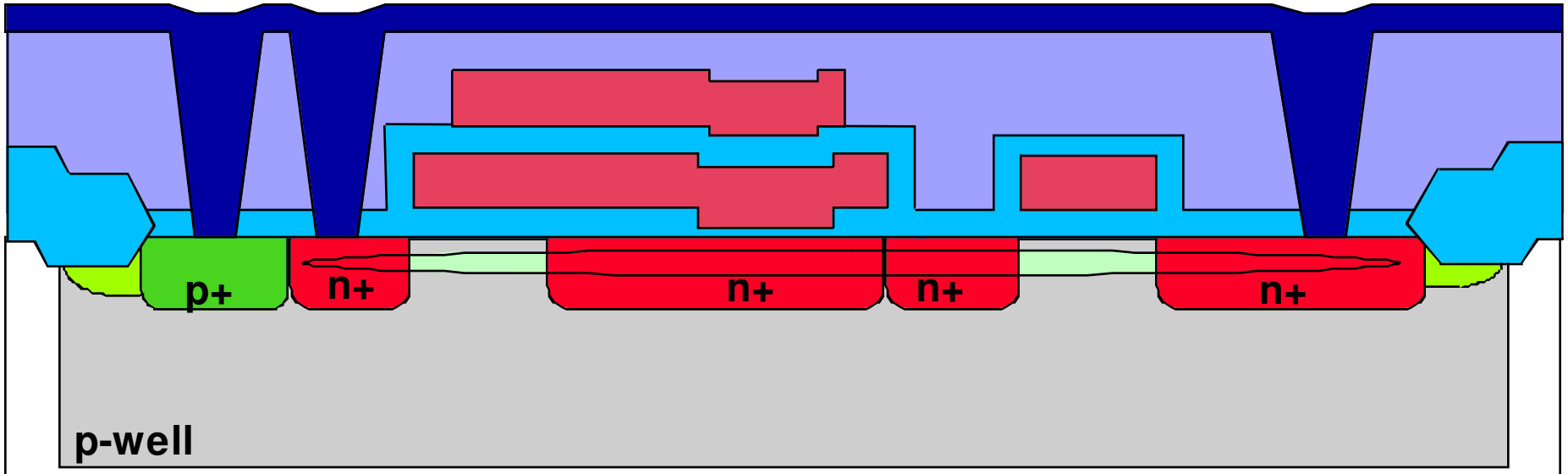
STRIP RESIST

EEPROM NMOS



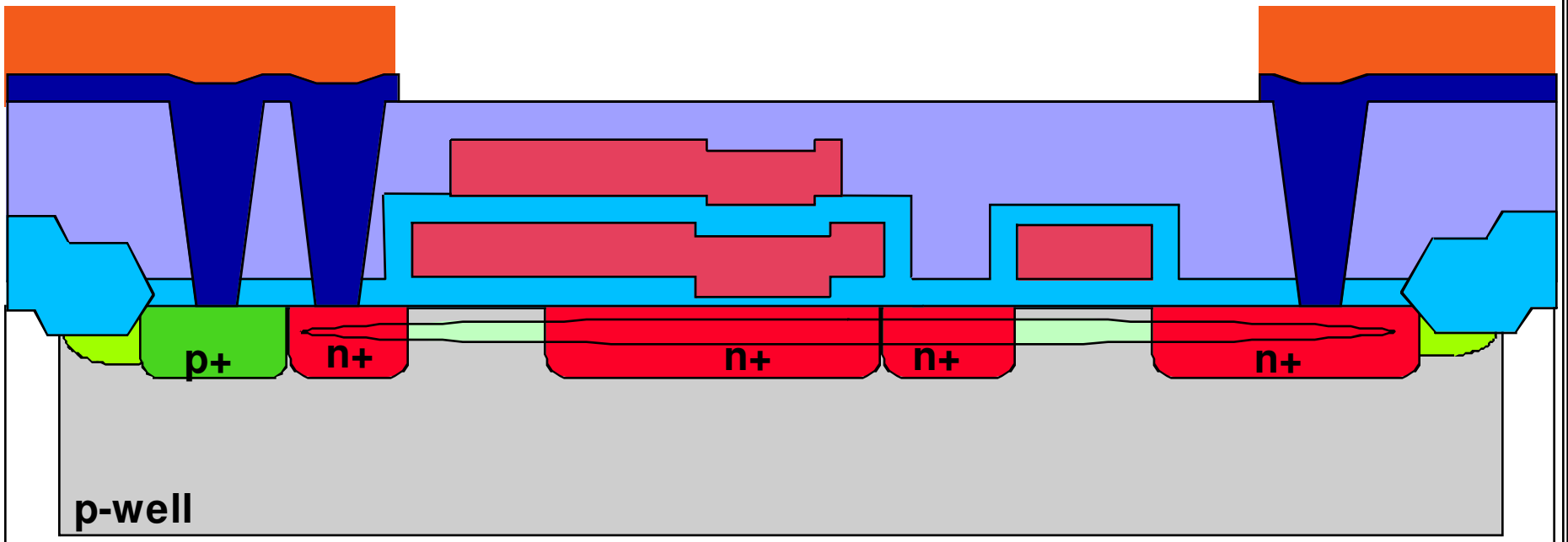
METAL DEPOSITION

EEPROM NMOS



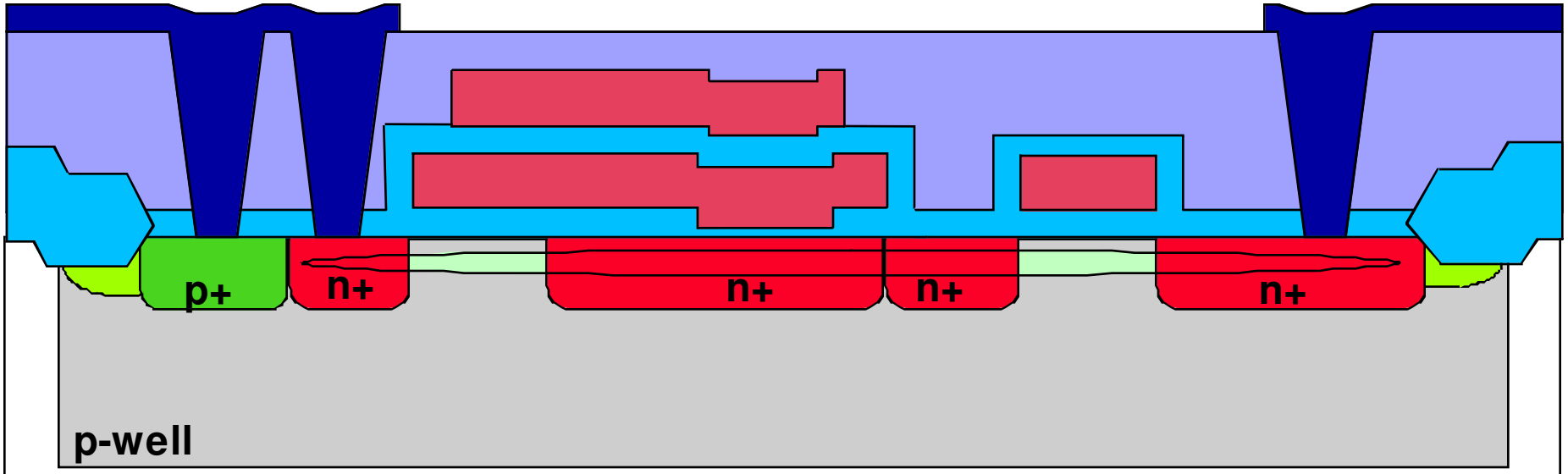
1µm Aluminum

LEVEL 9 PHOTO - METAL



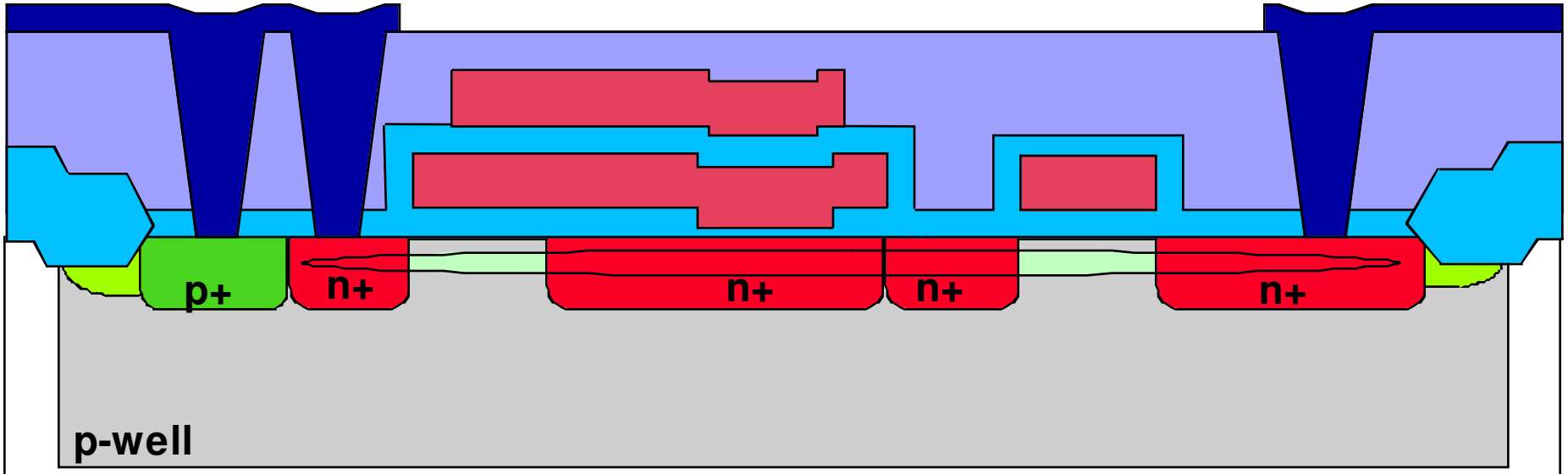
ETCH METAL AND STRIP RESIST

EEPROM NMOS



SINTER

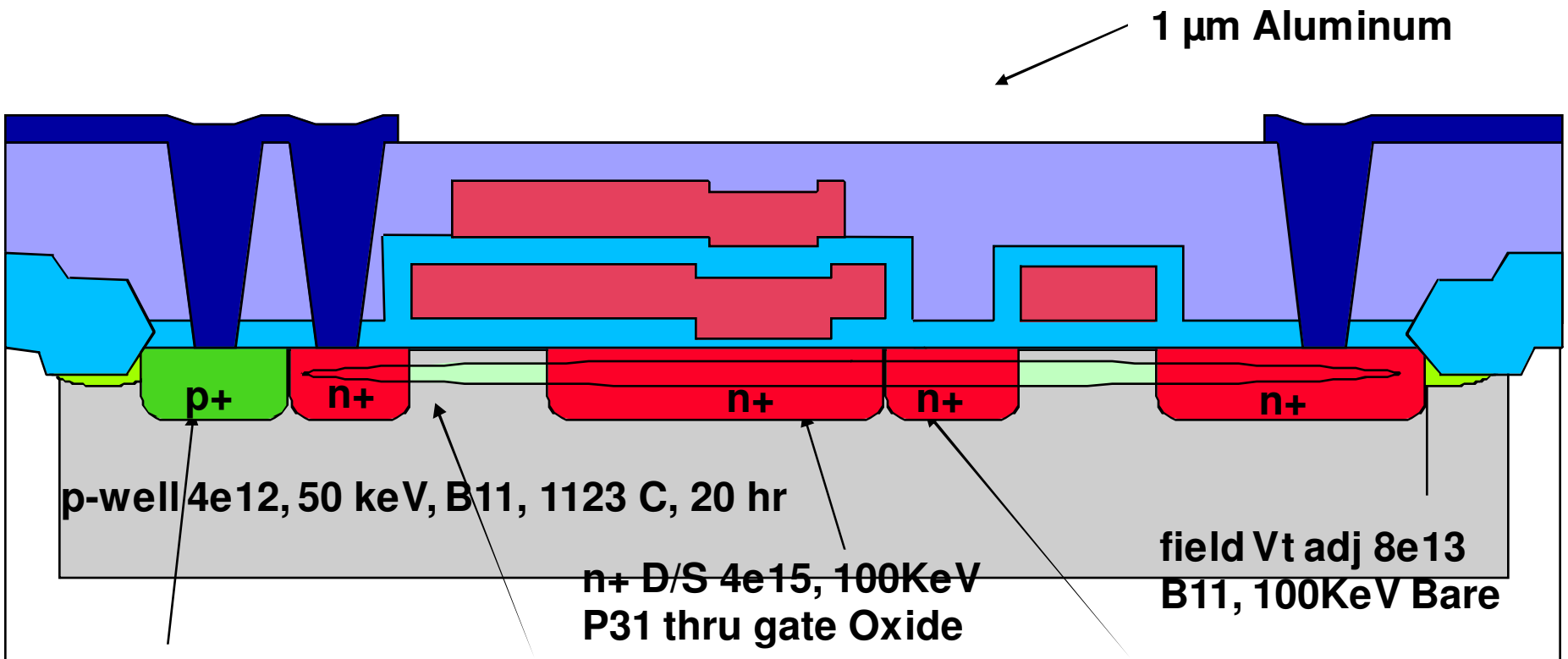
EEPROM NMOS



EEPROM

EEPROM

NMOS



Nonvolatile Semiconductor Memory Technology

A Comprehensive Guide to Understanding and Using NVSM Devices

Edited by

William D. Brown
University of Arkansas

Joe E. Brewer
Northrop Grumman Corporation

IEEE PRESS

IEEE Press Series on Microelectronic Systems
Stu Tewksbury, *Series Editor*
IEEE Solid-State Circuits Council, *Sponsor*
IEEE Components, Packaging, and Manufacturing Technology Society, *Sponsor*

The Institute of Electrical and Electronics Engineers, Inc., New York

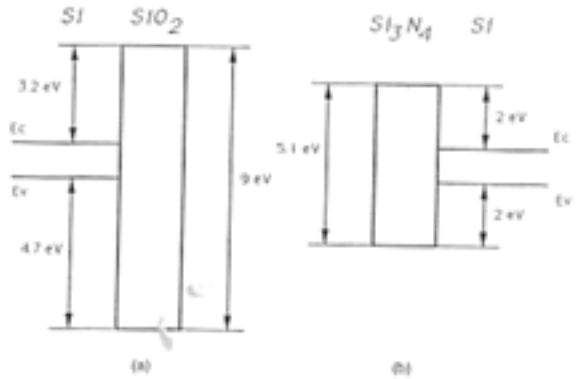


Figure 1.9 Energy band structures of (*a*) the Si–SiO_2 system and (*b*) the Si–Si_3N_4 system.

[1.20, 1.21], and source-side injection (SSI) [1.22, 1.23]. The first two are based on a quantum mechanical tunneling mechanism through an oxide layer, whereas the last two are based on injection of carriers that are heated in a large electric field in the silicon, followed by injection over the energy barrier of SiO_2. In order to change the charge content in charge-trapping devices, direct band-to-band tunneling and modified Fowler–Nordheim tunneling mechanisms are used. In the following sections, these six mechanisms are discussed briefly.

1.2.1 Fowler–Nordheim Tunneling

One of the most important injection mechanisms used in floating gate devices is the so-called Fowler–Nordheim tunneling, which, in fact, is a field-assisted electron tunneling mechanism [1.24]. When a large voltage is applied across a polysilicon–SiO_2–silicon structure, its band structure will be influenced as indicated in Fig. 1.10. Due to the high electrical field, electrons in the silicon conduction band see a triangular energy barrier with a width dependent on the applied field. The height of the barrier is determined by the electrode material and the band structure of SiO_2. At sufficiently high fields, the width of the barrier becomes small enough that electrons can tunnel through the barrier from the silicon conduction band into the oxide conduction band. This mechanism had already been identified by Fowler and Nordheim for the case of electrons tunneling through a vacuum barrier, and was later described by Lenzlinger and Snow for oxide tunneling. The Fowler–Nordheim current density is given by [1.24]:

Figure 1.10 Energy band representation of Fowler–Nordheim tunneling through thin oxides: the injection field equals the average thin oxide field. Electrons in the silicon conduction band tunnel through the triangular energy barrier.

$$J = \alpha\, E_{inj}^2 \exp\left[\frac{-E_c}{E_{inj}}\right] \tag{1.3}$$

with

$$\alpha = \frac{q^3}{8\pi h \phi_b} \frac{m}{m^*} \tag{1.4}$$

and

$$E_c = 4\sqrt{2m^*}\,\frac{\phi_b^{3/2}}{3hq} \tag{1.5}$$

where h = Planck's constant
ϕ_b = the energy barrier at the injecting interface (3.2 eV for Si–SiO_2)
E_{inj} = the electric field at the injecting interface
q = the charge of a single electron (1.6×10^{-19} C)
m = the mass of a free electron (9.1×10^{-31} kg)
m^* = the effective mass of an electron in the band gap of SiO_2 (0.42m [1.22])
h = h/2π

Equation (1.3) is the simplest form for the Fowler–Nordheim tunnel current density and is quite adequate for use with nonvolatile memory devices. A complete expression for the tunnel current density takes into account two second-order effects: image force barrier lowering and the influence of temperature.

The image force lowers the effective barrier height due to the electrostatic influence of an electron approaching the interface. Two correction factors $t(\Delta\phi_b)$ and $v(\Delta\phi_b)$, have to be introduced into Eq. (1.3), both of which are tabulated elliptic integrals and slowly varying functions. The reduction in energy barrier height ($\Delta\phi_b$) is given by [1.24]:

$$\Delta\phi_b = \frac{1}{\phi_b}\sqrt{\frac{q^3 E_{inj}}{4\pi\varepsilon_{ox}}} \tag{1.6}$$

Although tunneling is essentially independent of temperature, the number of electrons in the conduction band, available for tunneling, is dependent on the temperature. This dependence can be taken into account by a correction factor f(T), given by [1.24]:

$$f(T) = \frac{\pi c k T}{\sin(\pi c k t)} \tag{1.7}$$

with

$$c = \frac{2\sqrt{2m^*}\; t(\Delta\phi_b)}{hqE_{inj}} \tag{1.8}$$

Taking these two corrections into account, we see that the expression for the Fowler–Nordheim tunnel current density becomes:

$$J = \alpha\, E_{inj}^2 \frac{1}{t^2(\Delta\phi_b)} f(T) \exp\left[\frac{-E_c}{E_{inj}}\, v(\Delta\phi_b)\right] \tag{1.9}$$

The influence of the correction factors is small, however, and, for most practical calculations, the basic Eq. (1.3) is sufficiently accurate.

The Fowler–Nordheim tunnel current density is, thus, almost exponentially dependent on the applied field. This dependence is shown in Fig. 1.11*a* for the monocrystalline silicon–SiO_2 interface. The Fowler–Noedheim current is usually plotted as $\log(J/E^2)$ versus 1/E, which should yield a straight line with a slope proportional to the oxide barrier, as shown in Fig. 1.11*b*. In this case, the numerical expression is

$$J\,[A/m^2] = 1.15\,10^{-6}\, E_{inj}^2 \exp\left[\frac{-2.54\,10^{10}}{E_{inj}}\right] \tag{1.10}$$

which, at an injection field of 10 MV/cm, leads to a current density of approximately 10^7 A/m^2 or 10^7 pA/μm^2. This high value of injection field is of the order of that needed across the oxide during the programming of a nonvolatile memory device. The breakdown field of these oxides should, of course, be significantly larger than this value. In order to reach these high-field values and limit the voltages needed during programming, very thin tunnel oxides are used; an injection field of 10 MV/cm is attained by applying a voltage of 10 V across an oxide of 10 nm thickness. In order to reduce the programming voltage, the tunnel oxide should become even thinner. A thickness of 6 nm, however, is the lower limit for good retention behavior. But these thin oxides are difficult to grow with low defect densities, as is required for floating gate devices. Moreover, below these values, other injection mechanisms, such as direct tunneling, can become important. Yield considerations now limit the usable oxide thicknesses to 8 to 10 nm [1.23].

It should be noted that the tunnel current density is totally controlled by the field at the injecting interface, and not by the characteristics of the bulk oxide. Once

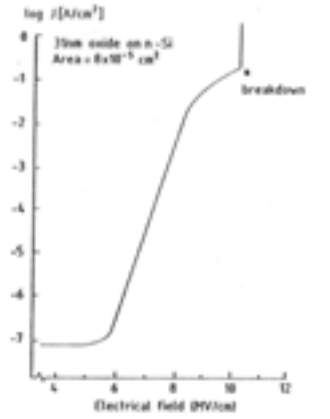


Figure 1.11 (a) Fowler–Nordheim tunneling current as a function of applied field across the oxide. The current is exponentially dependent on the field. Breakdown occurs around 10 MV/cm. (b) Fowler–Nordheim plot. I/E^2 as a function of $1/E$, extracted from the data (a). A straight line is obtained.

the electrons have tunneled through the barrier, they are traveling in the conduction band of the oxide with a rather high saturated drift velocity of about 10^7 cm/s [1.26].

For the calculation of the injection field at a silicon–SiO_2 interface, however, the flatband voltage has to be taken into account as seen by

$$E_{ox} = \frac{V_{app} - V_{fb}}{t_{ox}} \tag{1.11}$$

where V_{app} = the voltage applied across the oxide
V_{fb} = the flatband voltage
t_{ox} = the thickness of the oxide

When voltages are applied so that the silicon is driven into depletion, a voltage drop in the induced depletion layer must be accounted for in the calculation of the oxide field.

The tunnel current for a given applied voltage can be calculated as the product of the tunnel current density and the injecting area only if the tunnel current density has the same value over the whole injecting surface—that is, if the injection occurs uniformly over the area of the tunnel oxide. This assumes a perfectly plane injecting interface which, in many practical devices, will not be the case. Special cases of nonuniform injection are discussed in Sections 1.2.2 and 1.2.3.

1.2.2 Polyoxide Conduction

Fowler–Nordheim tunneling requires injection fields on the order of 10 MV/cm to narrow the Si–SiO_2 energy barrier so that electrons can tunnel from the silicon into the SiO_2 conduction band, as discussed in the previous section.

In oxides thermally grown on monocrystalline silicon, the injection field is equal to the average field in the SiO_2; therefore, thin oxides have to be used to achieve large injection fields at moderate voltages. Oxides thermally grown on polysilicon, called polyoxides, however, show an interface covered with asperities due to the rough texture of the polysilicon surface [1.27, 1.28]. This has led to the name "textured polyoxide." These asperities give rise to a local field enhancement at the interface and an enhanced tunneling of electrons [1.29, 1.30]. In polyoxides, the field at the injecting interface is, therefore, much larger than the average oxide field. Consequently, the band diagram of a polysilicon–polyoxide interface is as shown schematically in Fig. 1.12. Average oxide fields of the order of 2 MV/cm are sufficient to yield injection fields of the order of 10 MV/cm. This has the big advantage that large injection fields at the interface can be obtained at moderate voltages using relatively thick oxides, which can be grown much more reliably than the thin oxides necessary for Fowler–Nordheim injection from monocrystalline silicon.

A quantitative analysis of the tunnel current–voltage relations for polyoxides is rather complex. Although the tunnel mechanism itself is described by the same formula (Eq. 1.3), discussed in a previous section, the difficulty lies in the accurate determination of the injection fields to be used. It is no longer possible to use a single value for this injection field because of the nonuniformity of the field enhancement