Journal of
Low Power Electronics and Applications

MDPI

Article

Scaling Floating-Gate Devices Predicting Behavior for Programmable and Configurable Circuits and Systems

Jennifer Hasler *, Sihwan Kim and Farhan Adil

Electrical and Computer Engineering (ECE), Georgia Institute of Technology, Atlanta, GA 30332-250, USA; k.sihwan@gmail.com (S.K.); farhan467@gmail.com(F.A.)

* Correspondence: jennifer.hasler@ece.gatech.edu; Tel.: +1-404-894-2944; Fax: +1-404-894-4641

Academic Editor: Tony Tae-Hyoung Kim
Received: 4 March 2016; Accepted: 19 July 2016; Published: 27 July 2016

Abstract: This paper presents scaling of Floating-Gate (FG) devices, and the resulting implication to large-scale Field Programmable Analog Arrays (FPAA) systems. The properties of FG circuits and systems in one technology (e.g., 350 nm CMOS) are experimentally shown to roughly translate to FG circuits in scaled down processes in a way predictable through MOSFET physics concepts. Scaling FG devices results in higher frequency response, (e.g., FPAA fabric) as well as lower parasitic capacitance and lower power consumption. FPAA architectures, limited to 50–100 MHz frequency ranges could be envisioned to operate at 500 MHz–1 GHz for 130 nm line widths, and operate around 4 GHz for 40 nm line widths.

Keywords: floating-gate devices; FPAA; hot-electron injection; electron tunneling; scaling FPAA architectures

This paper discusses scaling of Floating-Gate (FG) devices, and the resulting implication to larger systems, such as large-scale Field Programmable Analog Arrays (FPAA). Figure 1 shows our high level figure, connecting the properties of FG circuits and systems in one technology (e.g., 350 nm CMOS), and predicting the behavior and advantages in smaller technologies. FG devices have been essential in demonstrating programmable and configurable analog and mixed mode computation, although typically at processes like 350 nm CMOS. The question of scaling these devices to more modern processes (e.g., 130 nm, 40 nm CMOS), typical of other system Integrated Circuits (IC) (e.g., FPGAs) remains, even though Electrically Erasable ROMs (EEPROMs) have moved to smaller and smaller linewidths (<20 nm gate length), and continued growth is expected.

Current EEPROM devices already store four bits (16 levels) in a single transistor of 100 nm $\times$ 100 nm area in 32 nm process [1,2]. A good overview of EEPROM/Flash history was presented at ISSCC2012 [3]. Recent data on EEPROM devices shows commercially announced devices at 15 nm and 19 nm [4–6]) as well as production of 32 nm devices. From the current EEPROM progress, such devices are expected to migrate to 7 nm and 11 nm technology nodes; therefore, the risk that the industry will not commercially produce a 10 nm floating-gate device is very low.

Figure 1 shows scaling FG devices to smaller line widths results in higher frequencies enabled through an FPAA fabric, as well as lower FPAA power consumption, due to lower parasitic capacitances. Scaling of FG devices is a key issue when working to improve the density as well as raising the density of FG based memories, computing in memory systems, and FPAA. For example, how will an FPAA's operating frequency improve as the IC technology process is scaled down? Figure 1 shows a modeling summary of the capability at the frequency of a particular FPAA device architecture as a function of process geometry used. Although the initial FPAA devices, built in 350 nm process, have achieved frequencies in the 50–100 MHz range (i.e., [7]), scaled FPAA devices should enable significantly higher frequencies, enabling RF type signals at 40 nm and smaller IC processes.

Therefore, the potential of scaled down devices, and the resulting computation, from a 350 nm process down to a 40 nm process, requires investigating both experimentally and analytically the effects of a 40 nm process.

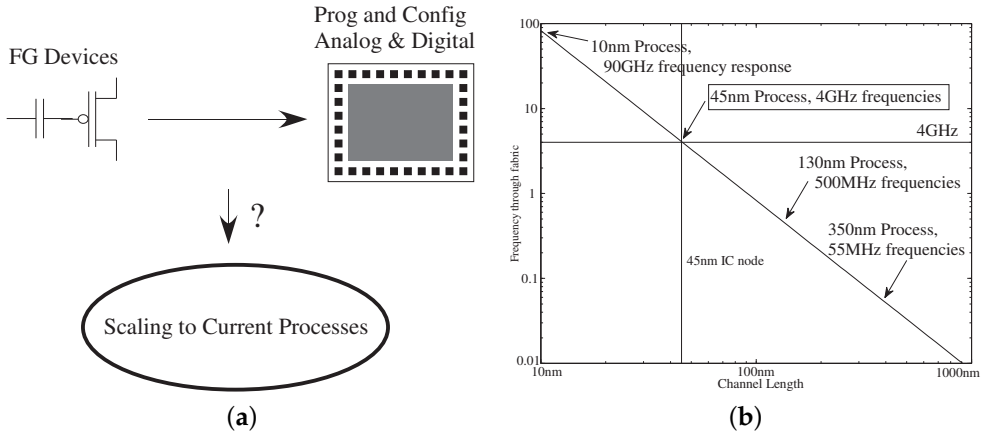


Figure 1. Scaling of Floating-Gate (FG) devices. (**a**) FG devices have been essential in demonstrating programmable and configurable analog and mixed mode computation, typically demonstrated using 350 nm CMOS processes. The question of scaling these devices to more modern processes (e.g., 130 nm, 45 nm CMOS); (**b**) frequency response of FPAA architectures as a function of minimum channel length. The results come from FPAA architecture modeling, CMOS process modeling, and experimental data where available (350 nm, 130 nm, 40 nm).

1. 350 nm FG Devices as Baseline for Scaling Performance

This section addresses what is needed for a functional FG device, typical of a wide range of circuit applications [7–9], as well as how these processes are characterized. Figure 2 shows the fundamental plots to characterize the resulting FG devices, enabling an automated FG algorithm [10]. Figure 2a shows that the current–voltage relationship is programmed through stored FG charge, resulting in a programmable weighting factor (i.e., subthreshold) and/or a programmable threshold voltage (V_{T0}, i.e., above-threshold). Although one has a capacitive divider, we have typical current–voltage relationships for a single curve. Changing the FG charge moves to a different curve, either increasing it by electron tunneling, or decreasing it by hot-electron injection. The resulting charge results in a voltage change on the FG node by the total capacitance at the FG node, or C_T, the sum of all capacitances at the FG node. An FG pFET transistor has a similar behavior to a pFET device, but with a different effective value for the subthreshold slope (U_T/κ for a typical FET device, where κ is the capacitive voltage divider between gate and surface potential, and U_T is the thermal voltage (kT/q)) due to the capacitive divider, the incoming (*gate*) capacitance and (C_T), and a programmable flatband voltage that can move the curve throughout the voltage range.

Because of the high quality gate insulators, the FG charge, once programmed, will remain roughly unchanged months and years later (at the same temperature) (e.g., [8,9]). We expect a typical FG device to change 1–100 μV at room temperature over a 10 year device lifetime as characterized by accelerated temperature measurements for this 350 nm CMOS process [9]. Long-term charge loss in floating-gate transistors occurs due to the phenomenon of thermionic emission, classically described by the simple model [11–14]

$$Q(t) = Q(0)\exp\left(-ve^{-q\phi_b/U_T}t\right) \quad (1)$$

where $Q(0)$ is the initial charge on the floating-gate, $Q(t)$ is the floating-gate charge at time t, v is the relaxation frequency of electrons in polysilicon, and $q\phi_b$ is the effective Si-insulator barrier potential (Volts). In [8], it has been shown that this model overestimates the charge loss, and that it does not follow such a simple curve, but a classical starting point.

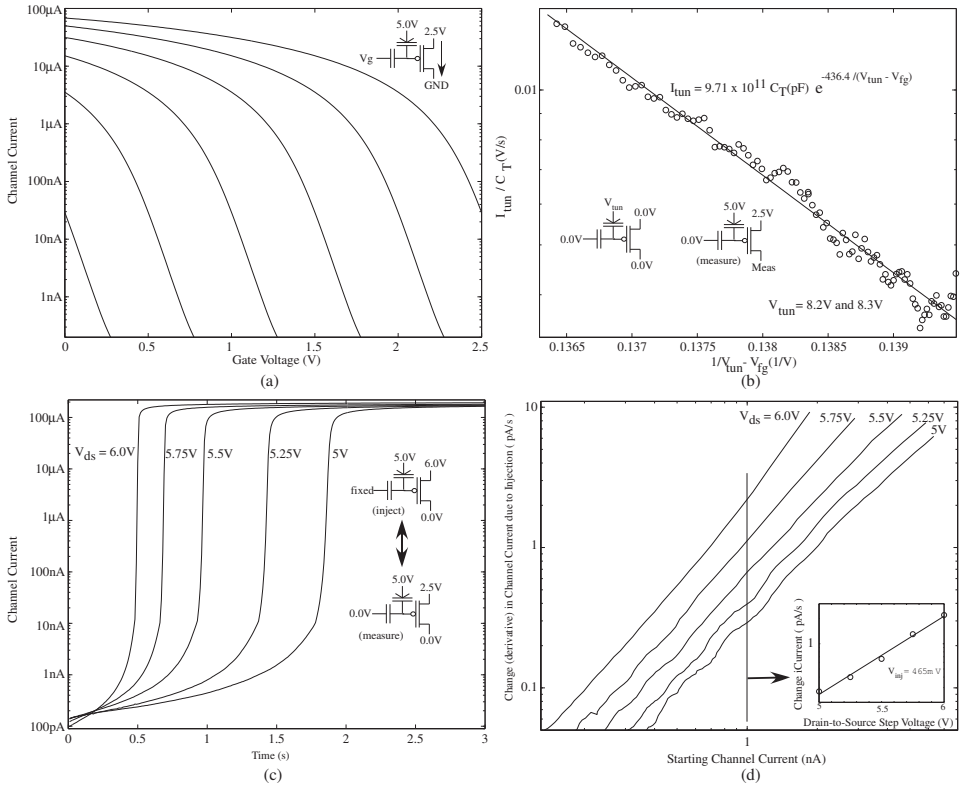


Figure 2. Fundamental measured data and resulting data regression sufficient for developing arrays of programmable FG devices consistent with characterization of previous FG devices [15,16]. The measured data is from a 350 nm commercially available CMOS process. (**a**) Channel Current versus Gate Voltage: an FG pFET transistor has a similar behavior to a pFET device, but with a different effective value for the subthreshold slope (U_T/κ); (**b**) Tunneling Current versus Gate and Tunneling Voltage: electron-tunneling erases our FG devices; therefore, tunneling characterization finds the right applied erasing voltage, V_{tun}, and the corresponding time required for erasing a device or an array; (**c**) Injection Current versus Time: channel current measurement sequence (*S* curve) showing effect of successive fixed-drain pulse injection, for multiple drain voltages. Started at a low current (≈100 pA), the hot-electron injection FG current increases the channel current to a nearly converged current of ≈100 µA; (**d**) change in Injection Current versus Injection Current: the change in the measured channel current versus channel current from *S* curve measurements. This data representation enables characterization of the exponential dependance on drain voltage (V_{inj}) on the change in channel current.

Figure 2b shows characterization of electron tunneling for an FG pFET in this 350 nm CMOS process. Electron tunneling adds a charge at the floating gate [15,17]. Tunneling current increases the resulting FG voltage, decreasing the resulting current measured from a pFET device connected to this FG [15]. The tunneling line sets the tunneling voltage (V_{tun}) controlling the tunneling current; thus, we can increase the floating-gate charge by raising the tunneling line voltage. Tunneling arises from the fact that an electron wavefunction has finite spatial extent [18,19]. For a thin enough barrier, this spatial extent is sufficient for an electron to pass through the barrier. Tunneling current depends on the exponential of a term proportional to the thickness and proportional to the square-root of barrier energy ($E_{barrier}$); the classic expression for tunneling through a square barrier [18–20]:

$$I_{tun} = I_{tun0} \exp\left(-\frac{2\sqrt{2m^*}}{\hbar}\sqrt{E_{barrier}}t_1\right), \tag{2}$$

where t_1 is the insulator thickness, m^* is the effective mass of an electron, and I_{tun0} is an experimentally determined constant for the particular insulator. An electric field across the insulator, created by the

voltage difference, reduces the thickness of the barrier to the electrons on the floating gate, allowing some electrons to move through the oxide. Fowler–Nordheim tunneling, or tunneling through a triangle barrier, models electron tunneling current as [18]

$$I_{tun} = I_{tun0} \exp\left(-\frac{4\sqrt{2m^*}}{3\hbar}\frac{E_{barrier}^{3/2}}{q\mathcal{E}}\right) = I_{tun0} \exp\left(-\frac{V_o}{V_{tun} - V_{fg}}\right), \quad (3)$$

where q is the charge of an electron, and $\mathcal{E}$ is the electric field in the insulator, $\mathcal{E}t_1 = V_{tun} - V_{fg}$, and $V_o = \frac{4\sqrt{2m^*}}{3q\hbar}E_{barrier}^{3/2}t_1$ is typically an experimentally measured parameter. Note that we can relate the pFET voltages as $V_{tun} - V_{fg} = V_{tun} - V_{dd} + V_{T0} + (V_{dd} - V_{fg} - V_{T0})$.

Figure 2c shows measurement for hot-electron injection sweeping through current for an FG pFET in this 350 nm CMOS process. Hot-electron injection enables programming FG devices by decreasing the FG voltage to the particular target location. Our approach for hot-electron injection is based around fundamental physics [21], as well as fundamental FG devices and circuit innovations using transistors operating with subthreshold or near subthreshold bias currents [16]. The fundamental model for hot-electron injection current (I_{inj}) is [21]

$$I_{inj} = I_s e^{f(V_{fg}, \Phi_{dc})}, \quad (4)$$

where I_s is the channel current, V_{fg} is the floating-gate voltage, and Φ_{dc} is the drain-to-channel potential for the pFET device; we often use a linearized exponential function for injection current

$$I_{inj} \approx I_{inj0}\left(\frac{I_s}{I_{s0}}\right) e^{\Phi_{dc}/V_{inj}}, \quad (5)$$

where V_{inj} represents the one parameter for this linearization. The exponential dependance of drain voltage on injection current will be utilized to enable a wide dynamic range of programming step sizes with linearly-scaled, lower-precision gate and drain voltages.

Figure 2c shows the characteristic positive feedback process for subthreshold channel currents, and the eventually saturating behavior for above-threshold channel currents, which we designate as an S curve for hot-electron injection given the shape of the response [10,16]. For a starting drain current, injection decreases the FG voltage, increasing the drain current, further decreasing the FG voltage [17]. The process slows down as the current moves to above-threshold operation (defined as significantly greater than threshold current, or I_{th}) because as the FG voltage decreases, the increased drain current decreases the drain-to-channel voltage available for injection due to additional voltage drop across the channel [16]. Eventually, the resulting injection current slows down, resulting in minimal change in FG voltage.

Figure 2d shows that we can get more information by extracting measured current changes as a result of injection, the values required in FG programming algorithms. From these S curves, we can find the change in current as a function of current, which yields a straight line for subthreshold currents [10]. From these curves, one can, for fixed current levels, curve fit to extract out the value of V_{inj} for that device at that particular bias condition.

2. Scaling of FG Devices

The following sections illustrate measured data for characterized FG devices fabricated in 130 nm and 40 nm CMOS processes as carefully chosen representative processes to show the impact of device scaling. Other CMOS processes from 350 nm to 14 nm CMOS follow along predictable expectations based on the approaches for these two devices. Finally, the last subsection discusses the theory of FG scaling to understand this data as well as extensions to other CMOS processes. The FG device uses a thicker insulator MOSFET, available starting in 350 nm CMOS processes. Thicker oxide or effective insulators enable long (i.e., 10 year) charge storage lifetimes.

Figure 3a shows scaled pictures of different transistor sizes; the thicker insulator device for 45 nm process, although having a gate similar to a 250 nm process enabling long-term storage, has drain-source parasitic capacitance similar to a 45 nm process. Minimizing these parasitics is critical for frequency performance for any implementation, as well as important for keeping routing fabric as small as possible. Decreasing the entire size to a typical 45 nm device with a thicker insulator, typical of EEPROM type devices, is possible.

Figure 3b shows the top level (e.g., layout) generic view of a single-poly FG device. Practical devices have additional improvements; some are process dependent. This core structure, as shown, is used in every FG test structure since it characterizes baseline performance of these devices, starting from its initial introduction [22]. The structure is similar to the double-poly structure shown in Figure 2 of [8]. These devices do not put the gate coupling directly above the gate electrode, as in EEPROM devices, but rather the gate is brought out for *analog* control of the resulting device. One should never have contacts to the gate electrode; contacts can significantly increase the resulting gate leakage.

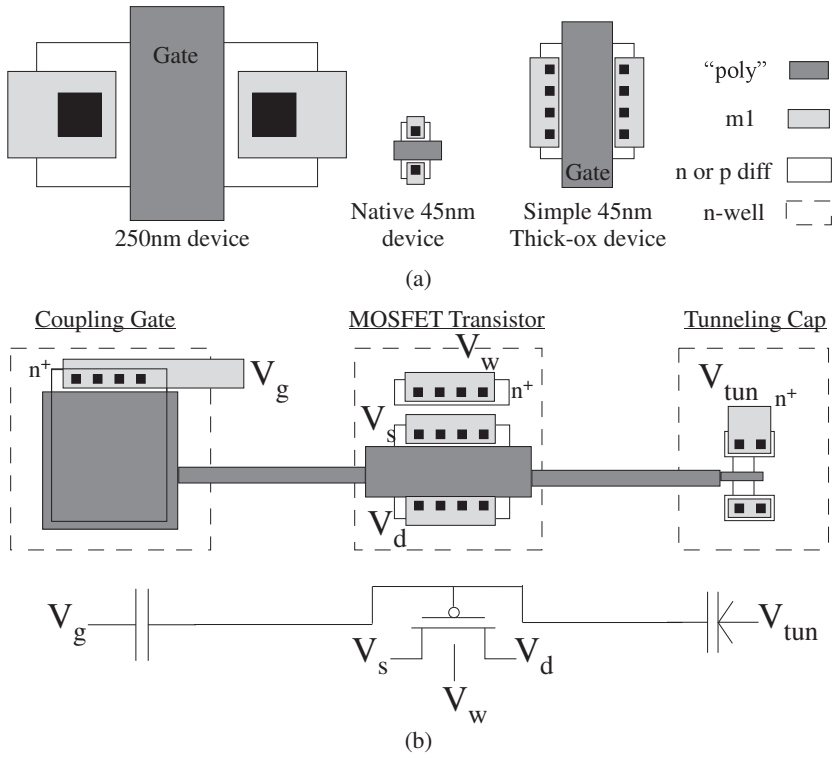


Figure 3. Scaling of FG MOSFET devices. (**a**) multiple picture of the resulting FG devices and how to look at larger insulators but with smaller parasitics. We show a typical 250 nm device, a typical 45 nm device, as well as a thicker insulator 45 nm device. The source-drain to substrate/well capacitance is significantly less in the 45 nm approach, the key parameter limiting performance for a dense FG array; and (**b**) single-poly cross-section typical for FG devices, as used for 130 nm and 45 nm measurements. Practical devices often have additional process-dependent modifications.

2.1. 130 nm FET Measurements

Figure 4 illustrates moving FG devices from 350 nm to smaller line width processes with SiO_2 gate insulator; this example shows data from a 130 nm CMOS process. FG devices are built from the larger insulator thickness, available in all processes smaller than 350 nm CMOS; the smaller insulator thickness allows significant tunneling current even with no voltage across the insulator. The insulator thickness is typically the size of a 350 nm CMOS insulator thickness (≈7 nm); we expect (and measure) similar (if not better) FG charge storage seen in the 350 nm processes [9]. Figure 4 shows measurements

of typical channel current versus gate voltage, typical tunneling current versus gate voltage measured through channel current, and typical injection current versus gate voltage (measured through channel current) and drain voltage. These measurements show typical behavior seen in 350 nm devices (e.g., Figure 2).

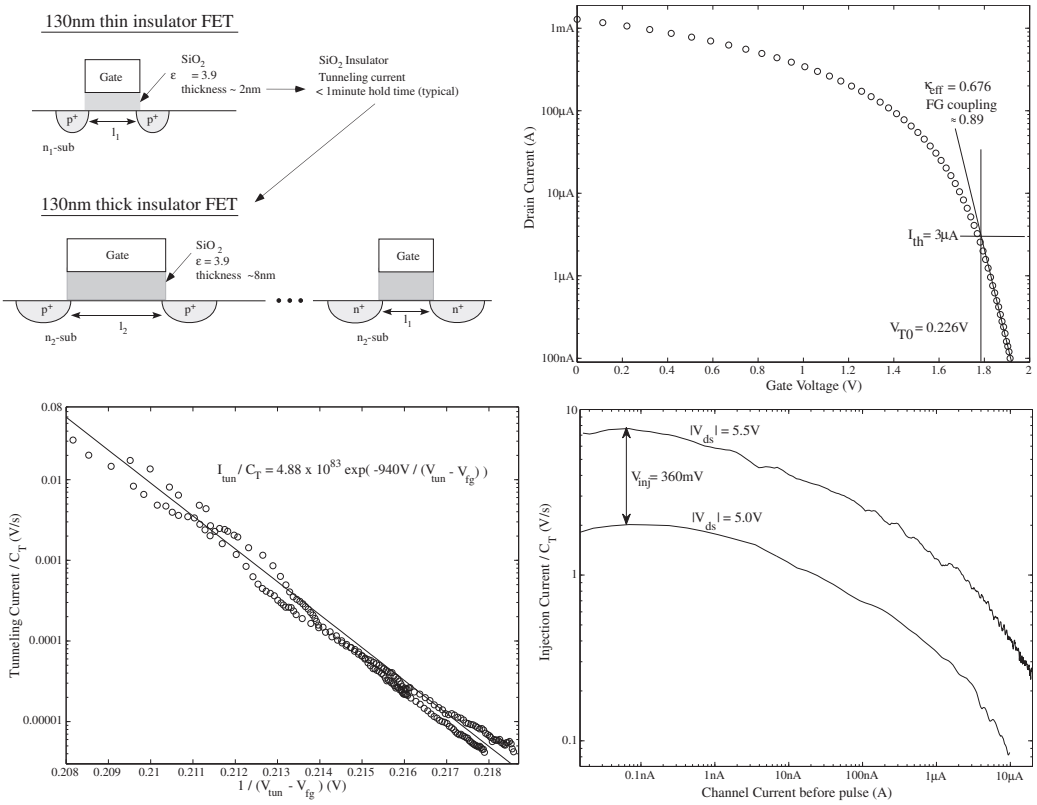


Figure 4. Moving FG devices from 350 nm to smaller line width processes with SiO_2 gate insulator; this example shows data from a 130 nm CMOS process. FG devices are built from the larger insulator thickness, available in all processes smaller than 350 nm CMOS; the smaller insulator thickness allows significant tunneling current even with no voltage across the insulator. Top Right: typical FG channel current versus gate voltage (coupling capacitively to the FG voltage), with the typical sub threshold and above-threshold regions, effective κ from the input capacitor coupling, and resulting threshold voltage ($V_{T0} = 0.226$ V) and threshold current (I_{th}); Bottom Left: typical tunneling current measured from two identical FG devices with extracted parameters; Bottom Right: typical injection current measured from a single FG device by continuous pulsing. The drain coupling for these devices creates significant current increases due to pulsing, shifting injection towards above threshold behavior for sub threshold currents.

2.2. 40 nm FG Devices

At 45 nm/40 nm (from 65 nm), one sees a major change in the resulting MOSFET device, in that we have a change in gate insulator from the time-tested SiO_2 to HfO_2 to reduce gate leakage in the thin insulator devices. Figure 5 shows a comparison of 350 nm to 40 nm FG devices, with the opportunities and changes due to a change in the gate insulator. The first question is whether these new FG devices hold charge, at least sufficiently long for testing our systems. In addition, do we get sufficiently long hold-times to expect anything close to 10 year lifetime results? Measurements to date have shown FG devices that hold charge over days with negligible change in the stored charge. To understand the effect, one looks at the square barriers between the 350 nm and 40 nm devices in Figure 5. The change in insulators do enable a thicker insulator but with a smaller barrier potential (1.4 eV [23] versus 3.0 eV [20]); therefore, for a square barrier we would expect lower leakage than the 350 nm device.

The leakage for the typical MOSFET at 40 nm, due to the larger insulator being lower than the leakage for a 90 nm/130 nm device. The FG device uses a thicker insulator to enable long (i.e., 10 year) charge storage lifetimes; this insulator thickness results in leakage levels expected in a 350 nm device.

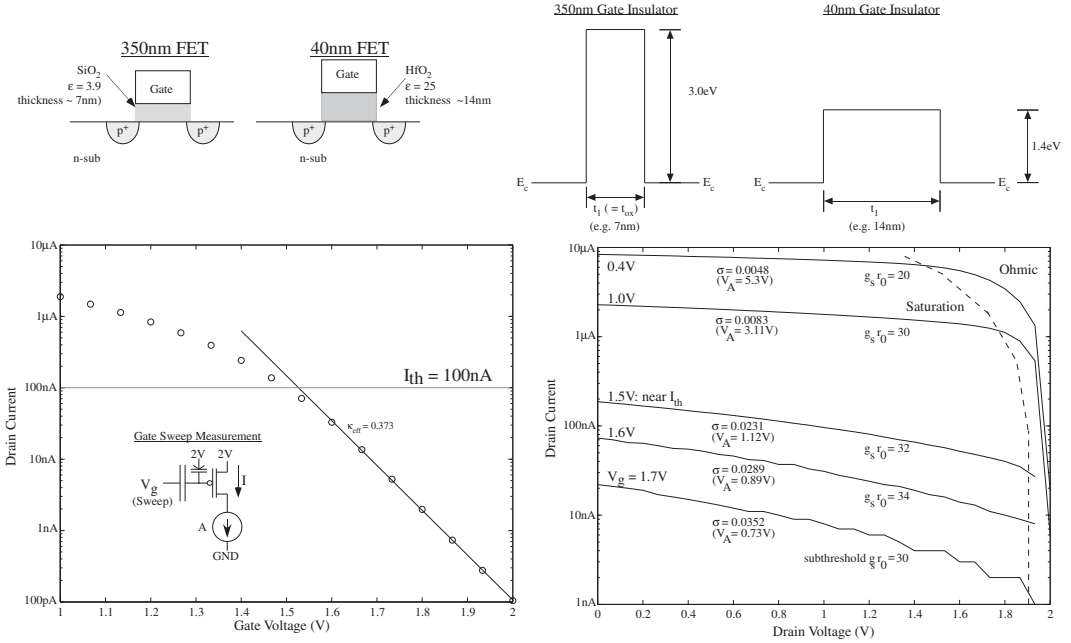


Figure 5. Illustration comparing a 350 nm FG FET and a 40 nm FG FET. We compare the typical device used for a 350 nm FET device versus a thicker insulator available 40nm FET device that could enable long-term lifetimes for FG devices. The key change in MOSFET topology at 40 nm/45 nm is the use of HfO_2 instead of SiO_2. The higher ϵ of HfO_2 (25) enables a much thicker material while enabling increased coupling capacitance into the MOSFET surface potential (Ψ). The change in insulators do enable a thicker insulator but with a smaller barrier potential (1.4 eV versus 3.0 eV); therefore, for a square barrier, we would expect lower leakage than the 350 nm device. From experimentally built FG devices in 40 nm IC process, we can measure the channel current for gate sweeps and drain sweeps, enabled by having an FG device that holds charge (currently tested to timescales of days with no degradation). From the measured drain current as a result of an FG gate sweep through the pFET subthreshold region and near threshold region, we extrapolate an effective κ of 0.373, and a threshold current of 100 nA. From the measured drain current versus swept drain voltage, we extract the resulting $g_s r_0$ of these devices that includes the effect of overlap capacitances.

Figure 6 shows the concept and measurement of electron tunneling through the HfO_2 gate insulator. In both cases, electron tunneling is described by classic Fowler–Nordheim tunneling. The modified 40 nm FG FET insulator results in higher electron tunneling current because of the smaller barrier to Si (1.4 eV) versus the classic SiO_2 to Si barrier (3.0 eV). In regressing the tunneling data, we can assume for the region used that we are in a typical MOSFET region, since these are designed to handle higher voltages. For our above-threshold current measurement versus time, we can take our model of current–voltage relationship (verified by data to be reasonable) as

$$I = \frac{\kappa^2 I_{th}}{4U_T^2}\left(V_{dd} - V_{fg} - V_{T0}\right)^2 \rightarrow V_{dd} - V_{fg} - V_{T0} = \frac{2U_T}{\kappa}\sqrt{\frac{I}{I_{th}}}, \tag{6}$$

where threshold current, I_{th}, as $2KU_T^2/\kappa$, $K = \mu C_{ox}(W/L)$, and we extracted I_{th} as 100 nA from our data on this particular device. From these measurements of V_{fg}, we can extract tunneling by writing Kirchoff's current law at the FG as

$$C_T \frac{dV_{fg}}{dt} = I_{tun}(V_{fg}). \tag{7}$$

The resulting formulation allows us to take a numerical derivative to see the resulting tunneling current, enabling the plot in Figure 6 and resulting curve fit of (3).

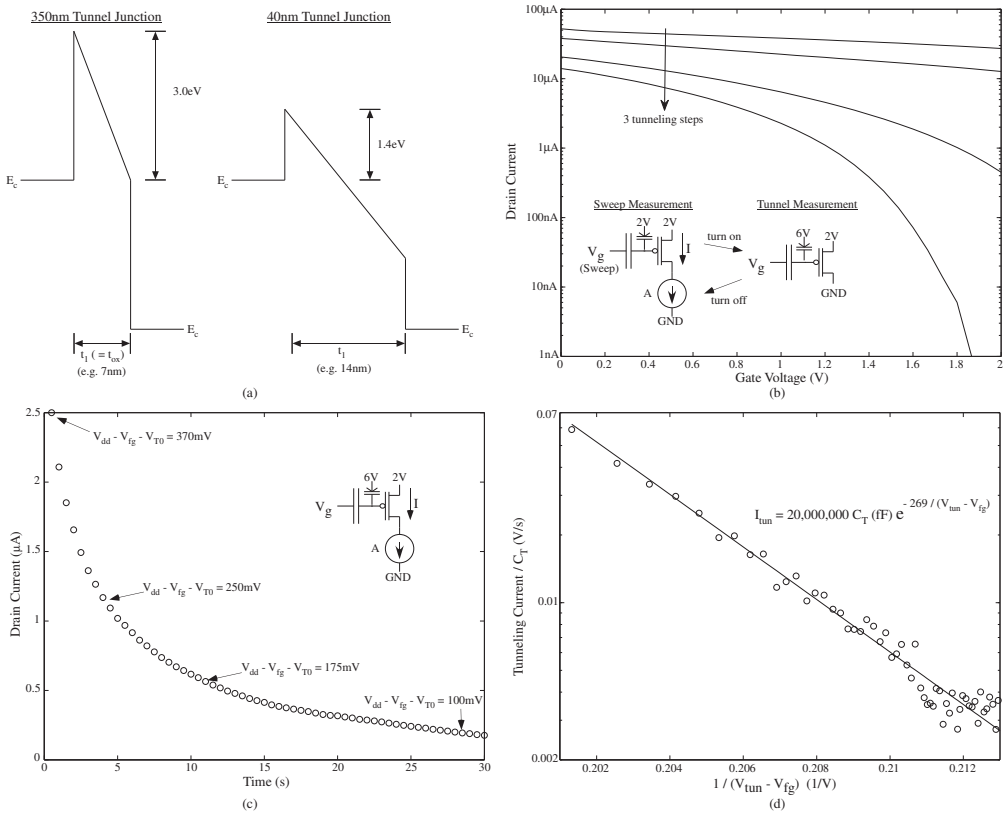


Figure 6. Measured drain current from a single 40 nm FG device demonstrating electron tunneling between sweeps. (**a**) comparison between 350 nm and 40 nm processes for electron tunneling are rooted in looking at the resulting band-diagrams; (**b**) one can take several gate sweep curves with tunneling between the curves. Tunneling occurred at 6 V supplied to V_{tun}, with delays on the order of a minute between curve sweeps (further indicating reasonable holding times from the FG devices). Curve sweeps were taken with V_{tun} at 2.0 V; (**c**) we can measure the time course of tunneling. From the resulting current (above-threshold) current measurements, we can extract floating-gate voltage ($V_{dd} - V_{fg} - V_{T0}$), enabling characterizing tunneling current versus tunneling terminal voltages ($V_{tun} - V_{fg}$); and (**d**) we regressed tunneling current per unit total floating-gate capacitance (C_T) versus $1/(V_{tun} - V_{fg})$ enabling a direct comparison of the data with the theoretical expression for Fowler–Norrdheim tunneling. We also plot a curve fit to that theoretical expression in (3).

Figure 7 shows the discussion for the hot-electron injection process. The lower energy barrier between HfO_2 impacts channel hot-electron injection by reducing the barrier for electrons injecting into the insulator.

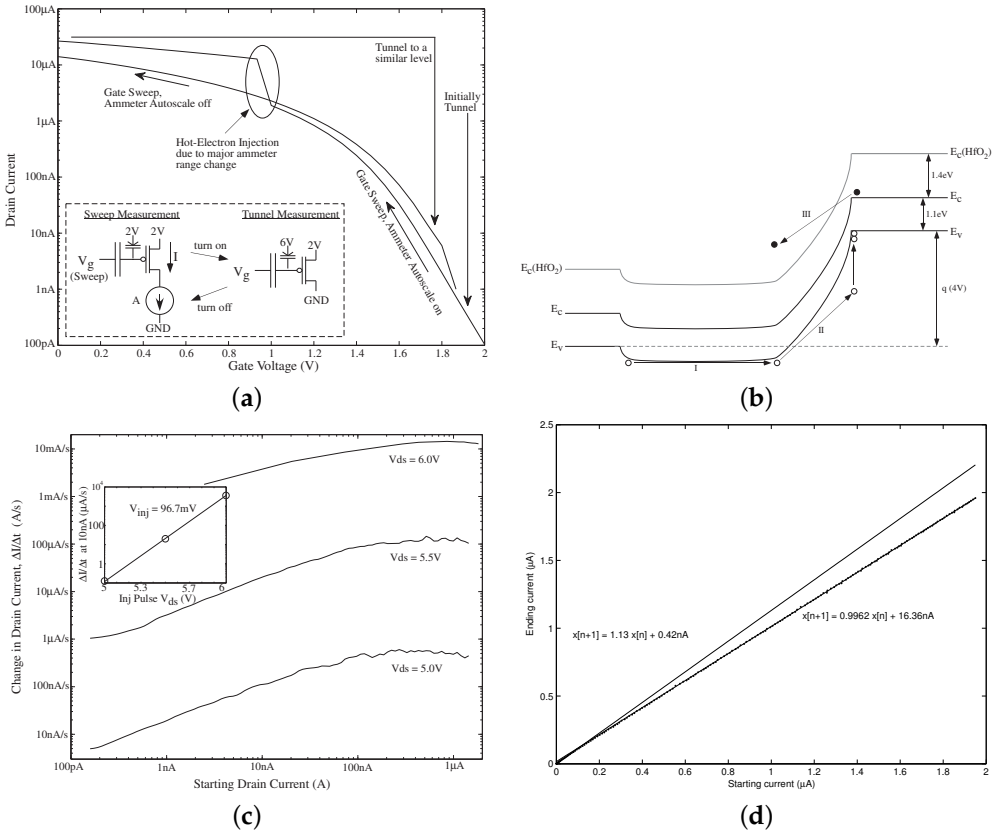


Figure 7. Initial FG hot-electron injection through measured drain current versus gate voltage sweep from a single 40 nm FG device. (**a**) initial tunneling to bring the initial curve into range. We took a gate voltage sweep that included a sharp jump in current, performed a tunneling step to return to a similar condition, and then took another gate sweep without auto ranging, eliminating the step in current during measurement. During the major ammeter range change at 2 μA, the drain voltage dropped for a short time to a voltage below ground, enabling enough field on this FG device to inject, as seen by the immediate step in current resulting from an decreased level of FG charge; (**b**) MOSFET band diagram for channel hot-electron injection for sub threshold currents in a 40 nm CMOS technology; (**c**) measured change in measured drain current for a fixed pulse width, T, versus starting drain current, measured before the pulse at low injection voltages. We show these measurements for three values of $|V_{ds}|$. We extract the resulting slope at a fixed current (i.e., 10 nA). This slope is $1/V_{inj}$; V_{inj} = 96.7 mV; and (**d**) measured resulting current after an injection event versus initial measured current.

Figure 7b shows the band diagram and the three steps required for hot-electron injection. The first step requires movement of holes through the channel region. The second step requires movement of holes through the drain-to-channel region, resulting in high energy carriers impact ionization, creating a source of electrons for the conduction band. The third step requires movement of electrons back through the drain-to-channel region, resulting in high-energy electrons that can surmount the insulator interface. For SiO_2 barriers, a wide range of the effects were limited by hole impact ionization, and we expect in these processes that the correlation will be far stronger. We expect that we will need similar voltages for injection across processes.

Qualitatively, the results are similar to hot-electron injection in larger MOS devices. Figure 7 shows a typical MOSFET injection characterization to determine parameters for FG programming [10]. Figure 7d shows an incremental increase due to injection. Often in programming algorithms, we make use of an effective linear difference equation(s) for early steps in reaching target value [10]. These approaches allow using simple fixed point functions for calculating FG injection pulses to reach an analog FG target.

2.3. FG Scaling Discussion

Figure 8 shows a progression in efficiency between the three processes for these approaches in terms of required applied voltages. The change in insulator, with its change in barrier height (3.0 eV to 1.4 eV for HfO_2), makes 40 nm significantly more efficient in FG writing capability, while still enabling 10 year lifetimes. For tunneling, we get a smaller V_o due to the lower starting barrier, and the quantity $V_o \propto t_1(E_{barrier})^{3/2}$ remains nearly constant (less than 6% change) between 350 nm and 40 nm devices. For injection, we get a significantly higher injection current and sharper slope (as seen by V_{inj}), as a combination of efficiency and higher substrate doping. We expect similar behavior scaling down to 14 nm devices give the similar insulator structure.

	350nm	130nm	40nm
$E_{barrier}$	3.0eV	3.0eV	1.4 eV
t_1	7nm	8nm	14nm
V_o	436.4V	940V (2 stage)	269V
V_{inj}	465mV	360mV	96.7mV
($\|V_{ds}\|$ center)	5.5V	5.25V	5.5V
ΔI(I = 1nA) /s	1pA/s	4pA/s	1μA/s

Figure 8. Comparison of measured device parameters. All three devices showed FG retention equivalent to less than 1mV room temperature drop over a 10 year lifetime.

Although scaling of tunneling physics is straightforward, scaling of hot-electron injection physics for these pFET devices should receive additional discussion. Hot-electron injection in pFET devices, operating at sub threshold and near threshold currents, are influenced mostly in the increased substrate doping allowed by the stronger insulator capacitor coupling, as well as the different Si-insulator barrier height. The substrate doping does not increase as fast because of the doping profile used for thicker insulator devices; when this layer is removed, one expects higher electric fields and lower impact-ionization and hot-electron injection voltages. Impact ionization always occurs significantly before any further device breakdown effects. We have two regions to consider, the hot-hole transport and resulting impact ionization that creates the resulting conduction band electrons, and the hot-electron transport and resulting electron injection efficiency.

The restoring force for the electron and hole high-field transport is primarily optical phonons, where first the carrier needs to gain more energy per unit distance than the optical phonon restoring force (E_R/λ) typically requiring a starting distance (z_{crit}) for an increasing potential, and then the average carrier trajectory includes field gained energy minus this required starting energy. The resulting distribution function around this average trajectory is a local convolution of Gaussian functions, the eigenfunction of a linear diffusion equation (e.g., heat equation). We have discussed these fundamental transport details elsewhere [21].

Electrons in an electric field will gain energy faster than holes in an electric field, as characterized by their typical mean free length for an optical phonon collision of energy E_R (≈63 meV), where electrons (λ_e) are approximately 6.5 nm [21], and holes (λ_h) are approximately 4.2 nm [24]. Although impact ionization can occur for an electron or hole with energy of 1.1 eV, requiring carriers to exceed this barrier, electron impact ionization is known to be an efficient process for electron energies above 2.3 eV [21,25], and hole impact ionization is known to be an efficient process for hole energies above 3.0 eV [24]. The resulting created electrons, sometimes starting with additional energy because of the impact ionization process, begin around the highest field region, gaining energy as they accelerate to get over the 3.0 eV (Si-SiO_2) or 1.4 eV (Si-HfO_2) barrier.

The 3.0 eV barrier level for significant hole impact ionization would be the primary limit for the hot-election injection process when using a Si-SiO_2 barrier (3.0 eV) given the significant difference between λ_h and λ_e, although some functional dependance is still possible. The 2.3 eV level for

hot-electron impact ionization means we will get some significant loss of electrons before reaching the Si-SiO_2 barrier (3.0 eV), while we have negligible loss of electrons before reaching the Si-HfO_2 barrier (1.4 eV), resulting in significantly higher injection current. For the Si-HfO_2 barrier, the electron dynamics can almost be approximated by a constant factor, and approximation often desired (but not physically correct) between injection and impact ionization currents.

3. Scaling of an FG MOSFET Used in an Array of Switches

Because we have proven that FG devices are functional throughout a wide range of CMOS IC processes, we now transition to looking at the scaling properties of an FG MOSFET acting as one of the multiple switches, say, in a small crossbar array (e.g., [7,17]). The operating speeds of an FPAA array is, to first order, limited by the FG MOSFET switch. We will analyze the high frequency behavior of a switch in an array of switches (e.g., routing fabric). For this analysis, a programmed switch is at one of two cases, when the switch is *off* and when the switch is *on*. A programmed FG voltage can be set at significantly higher or lower voltages than the power supply range. Our discussions focus on nFET and pFET devices; these dynamics are effectively interchangeable with slight changes in parameters.

3.1. Off-Switch Behavior

For the *off* switch case, we start with the channel in accumulation. Because the switch value can be programmed outside the power supply, a slightly positive value (above V_{dd}) will guarantee the device stays in accumulation for all applied voltages including the GHz frequencies that we are considering in this discussion. In accumulation, we have no appreciable depletion capacitance, and, therefore, the capacitance between the floating-gate terminal and the substrate is the oxide (or insulator 0 capacitance (C_{ox} W L). The effective conductance between source and drain is effectively zero, being the conductance of two reverse-biased diodes. Gate length has little to do with the off case (other than total gate capacitance) in accumulation.

With the zero conductance between source and drain, any potential communication between switches must happen through capacitive coupling. The gate to source-drain junction capacitance, C_{ov}, is the biggest issue in terms of signal feedthrough, which scales proportionally to other device properties. Minimizing C_{ov} decreases the amount of capacitive signal feedthrough feedthrough, which could be further decreased by opening up drain-source regions (avoiding some of the self-aligned device). C_{sb} and C_{db} p-n capacitors connected to signal ground (actually V_{dd}). Therefore, the frequency-indpendant coupling gain from source to drain voltage would be the resulting capacitive divider network as $C_{ov}^2/(C_T\ C_{db})$, where C_T is the total capacitance of the floating-gate node. Both terms result in a coupling less than 10^{-3}; with multiple switches in series, this value is nearly negligible. In a switch matrix, the resulting load capacitance is the sum of all of the capacitors on the line, further decreasing this effect. Furthermore, this effect is negligible in 350 nm FPAA devices at low frequencies, with no significant coupling measured whether in characterization or in applications.

3.2. On-Switch Conductance Behavior

For the *on* switch case, we are primarily concerned with the frequency response through the particular device. The MOSFET channel is biased far above threshold behavior, operating in the ohmic regime. The FG voltage is not constrained between the power supply rails, allowing the transistor to its maximum conductance point, the point for high gate voltage where the source-to-drain conductance of channel is approximately independent of gate voltage. This maximum conductance, or R_c, is roughly independent of process minimum channel length, where the conductance is set by velocity saturation of electrons/holes for the MOSFET channel. For a device with equal width and length, the nFET saturates around 3 kΩ and pFET near 6 kΩ.

Figure 9 presents the discussion for this maximum conductance, where an increase in gate voltage increases the number of collisions with Si–insulator barrier while carriers effectively move from source

to drain voltage. An *on* MOSFET switch typically would have a small voltage between the source (V_s) and drain (V_d) voltages, resulting in the MOSFET modeling

$$I = \mu C_1 \frac{W}{l}(V_g - V_{T0} - V_s/\kappa)(V_d - V_s), Q_d \approx Q_s = C_{ins}(V_g - V_{T0} - V_s/\kappa), \quad (8)$$

where μ is the carrier mobility in the channel, C_1 is the insulator capacitance per unit area (ϵ_1/t_1), κ is the capacitive divider between C_{ins} and the total capacitance in the channel, V_{T0} is the threshold voltage, W and l are the width and length of the MOSFET device, Q_s and Q_d are the channel charge at the source and drain edges of the channel region, respectively. The measurement in Figure 9 uses V_s = 0 V, V_d = 50 mV for continuously ohmic operation, while sweeping V_g over a wide voltage range. In this setup, V_s is set to the substrate (so V_s = 0); and pFET are measured down from their substrate held at V_{dd} (different for each technology). One would expect a conductance (G) that linearly increases, after V_{T0}, with V_g, as

$$G = \frac{I}{V_d - V_s = 50mV} = \mu C_1 \frac{W}{l}(V_g - V_{T0}). \quad (9)$$

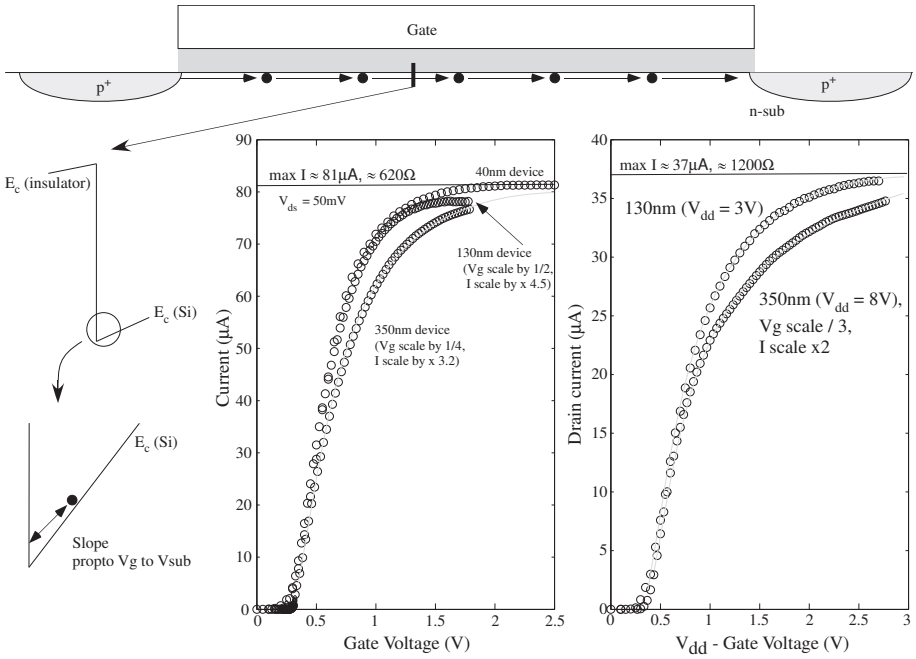


Figure 9. Maximum conductance for a MOSFET device is determined by the two-dimensional carrier behavior. The *channel* electron oscillates along the triangle barrier created by the insulator interface and MOS capacitor depletion region as it moves from source to drain regions. This oscillation increases the number of elastic collisions, decreasing the carrier mobility, as seen by the conductance saturating for higher gate voltages.

Figure 9 shows measured data illustrating this initial linear behavior, as well as deviations from it leading towards conductance saturation.

The modeling for conductance saturation investigates the change of μ with gate voltage; other terms remain nearly constant. Figure 9 shows that, although we draw carriers (e.g., electrons) moving in a straight line path through the channel from V_s to V_d, we have a field in the orthogonal direction (gate direction) that pulls these carriers towards the gate region. These carriers get pulled into the Si-insulator barrier, elastically colliding and reversing direction towards the substrate until the electric

field of the channel brings the carriers back towards equilibrium. From (8), the electric field at the Si-insulator barrier

$$\text{insulator}: \frac{V_g - V_{T0}}{t_1},$$

$$\text{Si edge}: \mathcal{E}_{Si} = \frac{\epsilon_{ins}}{\epsilon_{Si}} \frac{V_g - V_{T0}}{t_1}. \tag{10}$$

The electric field in S_i will decrease moving into the depletion region. A constant electric field (linear change in potential) is expected at the boundary layer right at the Si-insulator boundary. The additional elastic collisions will decrease the resulting carrier mobility, μ,

$$\mu = \frac{q\tau}{m^*}, \frac{1}{\tau} = \frac{1}{\tau_0} + \frac{1}{\tau_{gateE}}, \tag{11}$$

where τ is the mean free time due to collisions, m^* is the carrier effective mass, τ_0 is the mean free time due to typical restoration forces in the channel, such as acoustic phonons, elastic scattering mechanisms, as well as any effects due to some optical phonon behavior, and τ_{gateE} is the collision component due to average elastic collisions with the Si–insulator barrier. Transit time would be the ratio of average distance traveled over the average velocity of carriers. The energy of the carriers through the channel is never high, roughly at the typical kT = $q\ U_T$ average energy for a Fermi distribution (energy less than Fermi level), for an average distance traveled (due to electric field) as $U_T/\mathcal{E}_{Si}$. Velocity of carriers at lower $\mathcal{E}_{Si}$ is proportional to $\mu\mathcal{E}_{Si}$. Velocity of carriers at higher $\mathcal{E}_{Si}$ approaches velocity saturation, v_{sat}; in this region, we get

$$\tau_{gateE} = \frac{U_T}{v_{sat}} \frac{\epsilon_{Si}}{\epsilon_1} \frac{t_1}{V_g - V_{T0}}. \tag{12}$$

For increasing V_g, the transport progresses starting in the region with a constant τ_0, resulting in classical linear increase in conductance, to the region where τ is decreasing due to elastic collisions, resulting in a sub linear increase in conductance, to the region where τ is dominated by τ_{gateE} with large $\mathcal{E}_{Si}$. In this final region, where the conductance saturates at G_{max}, the current is expressed as:

$$I \to \frac{q}{m^*} \tau_{gateE} \frac{\epsilon_1}{t_1} \frac{W}{l} (V_g - V_{T0})(V_d - V_s)$$

$$= \frac{qU_T}{m^*} \frac{\epsilon_1}{t_1} \frac{\epsilon_{Si}}{\epsilon_1} \frac{t_1}{V_g - V_{T0}} \frac{1}{v_{sat}} \frac{W}{l} (V_g - V_{T0})(V_d - V_s)$$

$$= \frac{qU_T}{m^*} \frac{\epsilon_{Si}}{v_{sat}} \frac{W}{l} (V_d - V_s), \tag{13}$$

$$G_{max} = \frac{I}{V_d - V_s} = \frac{qU_T \epsilon_{Si}}{m^* v_{sat}} \frac{W}{l}, \tag{14}$$

where G_{max} is not a function of typical device parameters, except for drawn transistor width and length. G_{max} is not a function of the insulator thickness. Figure 9 shows 350 nm, 130 nm and 40 nm nFET or pFET device measurements illustrating the conductance saturation in each case.

3.3. On-Switch Capacitance Behavior

The other side of the question is the resulting capacitance to set the resulting time constant. For example, we can make wider MOSFET switches, decreasing the resulting channel resistance, but increasing the resulting capacitance found in a dense array of FG devices. Switch capacitance is primarily due to source-drain junctions, C_{sb} and C_{db}, as well as a small capacitance through the FET gate oxide, through the resulting capacitive network to other potentials. We identify the resulting capacitance as C_s. The relative size of these capacitances typically scales quadratically with scaling of process line width.

The design of such an array must discuss whether to have the well connected to signal GND, which would be a solid GND even for RF frequencies, or have a high-impedance connection. A useful high-impedance to each switch requires that each switch be placed in a separate isolated well device, as well as having a high resistance connection to the resulting well terminal.

Therefore, for a single switch, we see minimal additional coupling effects through the switch. We expect there will be some effect of the RC effect in the channel (usually modeled by a resistor and inductor) of course, around f_T (max) due to maximum conductance. The low frequency modeling directly applies to our modeling approaches.

4. FPAA Capacitance Measurement, Modeling, and Architecture Tradeoffs

Having working FG devices as well as modeling of individual switches, we move towards modeling the frequency response of an FPAA fabric, justifying Figure 1b. Figure 10 shows the basic Manhattan based routing structure used for our SoC FPAA device (i.e., [7]). The approach includes a compartment for a Computational Analog Block (CAB) or a Computational Logic Block (CLB), includes two Connection (*C*) blocks to connect these devices, and includes one routing Switch (*S*) block. Using the Manhattan style routing enables direct interactions with existing tool flows ([26]). We still hold that the routing fabric in this architecture is both useful for computation as well as switching, particularly for local CAB/CLB routing as well as in the '*C* block routing. Other FPAA architectures show similar approaches and tradeoffs, although this architecture mades these issues more explicit.

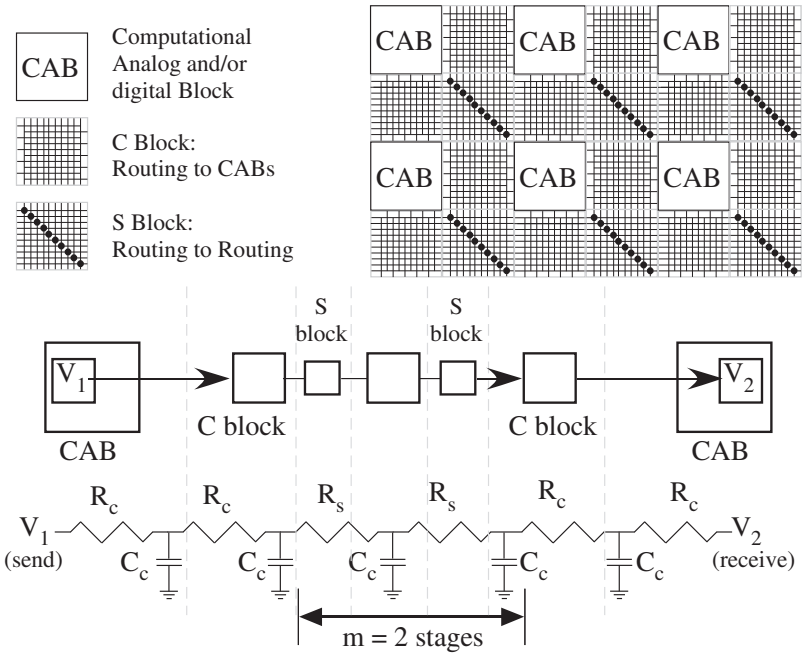


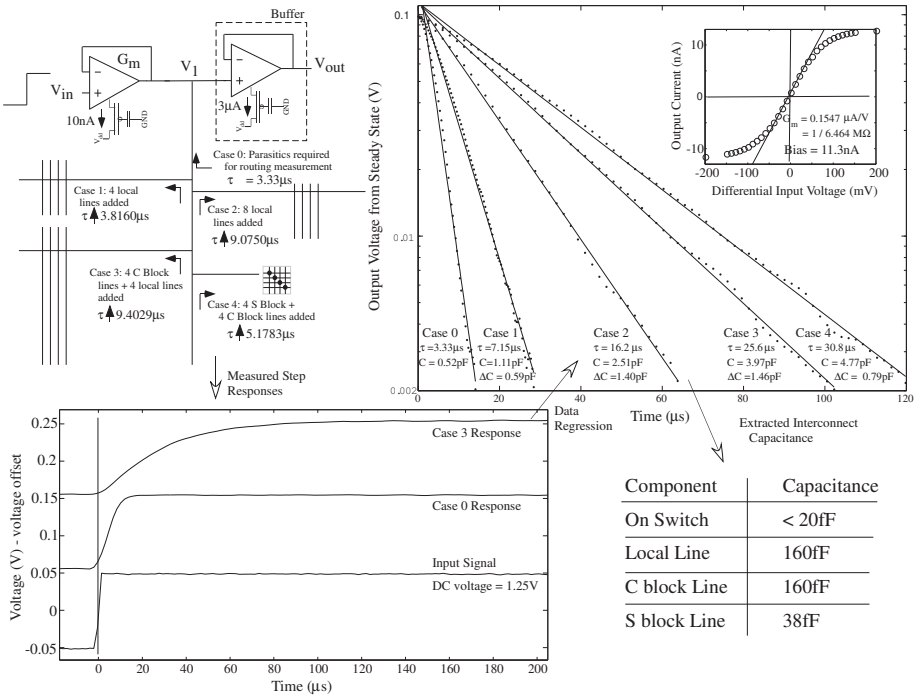
Figure 10. Manhattan FPAA architecture, including the array of computation blocks and routing, composed of Connection (*C*) and Switch (*S*) blocks. The routing infrastructure can effectively be modeled as a distributed resistance–capacitance line. The lowest figure shows a typical routing fabric assuming a single routing of *C* and *S* block switches, where C_c is the connection capacitance including the *C* block lines, R_c is the *C* block switch resistance, and R_s is the unbuffered *S* block switch resistance. m = typical number of switches needed for a connection.

From a classical FPGA approach, one considers the capability of the device to be solely in its components (CLBs, specialized blocks), and the routing fabric is simply a mechanism to interconnect these components. Minimizing the effect of the routing fabric reduces, from a circuit perspective, dead weight that can only degrade the circuit. This approach requires minimizing the number of switches, each of which adds resistance, as well as minimizing the resulting capacitance of the routing. The

routing infrastructure can effectively be modeled as a distributed RC line. The architecture looks at the relationship of the resulting switch resistances, as well as other circuit uses of the FG switch devices as a function of the number of CAB inputs and number of tracks, as well as the typical number of switches needed for a connection.

4.1. SoC FPAA Routing Fabric Characterization and Computation

The SoC FPAA [7] enables programming experiments that characterize the fundamental properties of the configurable fabric by experimental measurements on the configurable routing fabric. Figure 11 illustrates compiling (and measuring) two circuits to characterize precisely the behavior of these circuits, including load capacitance of the fabric itself. This FPAA structure facilitates the direct characterization of the resulting capacitance, coupled with the resistance of an on-switch, R_c. One can directly predict delays along each of these lines. Every experiment uses the same voltage biasing, fixing the capacitance of p-n junction devices throughout this experiment. The resulting measurements give a measurement of the resulting routing capacitance, as well as enables, through the routing fabric, a range of tunable capacitor blocks. Precise measurement of routing capacitances enables tuning, through programming switches, for precise capacitances where needed for matching. Matching of capacitances and programmability of current sources by FG techniques dramatically reduces the effect of mismatch in small cell sizes.



Component	Capacitance
On Switch	< 20fF
Local Line	160fF
C block Line	160fF
S block Line	38fF

Figure 11. FPAA characterization of routing capacitances. Initially, one first measures the current–voltage relationship for a specific Transconductance Amplifier (TA), shown in the inset, to exactly find the resulting G_m (0.1547 µA/V) of the device. That exact TA with the same programmed current is used to measure the time-constant of the step response (on a 1.2 V dc for the 2.5 V supply) for different (additive) routing combinations. From the step responses measurement shown, a linear curve (in log amplitude) fits to the time constant after removing the effect of the steady state voltage. The extracted routing capacitance values for multiple measurement configurations are summarized.

4.2. Architecture Tradeoffs for FPAA Switch Fabric

Next, we use our FG switch and network modeling to look at speed through routing fabric (Figure 10) like the SoC FPAA [7]. The architecture choices look at the relationship of the resulting switch electrical properties, as a function of the number of CAB inputs and number of tracks (n), as

well as the typical number of switches needed for a connection (m). Figure 10 shows a case when $m = 2$. We have routing in and out of the CABs: a connection in the CAB and a connection in the first C block. In this analysis, every route starts with these four connections, and might require more routes through additional S and C blocks. We consider a point-to-point communication scheme; one can extend this modeling to other connections, expecting similar scaling rules. The pure switch routing approach is often the worst case situation for capability and computation speed through a routing fabric; the situation often simplifies when using routing for computation (e.g., classifier in [7]).

Figure 12 illustrates one set of tradeoffs as well as a summary of the key parameters of switch conductance and capacitances. We consider that capacitance as C_c, which is due to multiple (n) source-drain junction capacitances (C_s), or $C_c = nC_s$. An approximation of the time constant (τ) is computed to analyze the fabric communication frequency ($= 1/(2\pi\tau)$). The first-order approximation for τ is the product of all resistances along the line times all capacitance along the line. This formulation is an overestimation of the resulting time constant, typical for an infinite (diffusive) RC line. The total resistance along the line is $4R_c + mR_s$; four switches to get signals on and off a C line with all other switches in S block. We allow for the S switch to be larger (in W) than the C switch by a factor A, resulting in $R_s = R_c A$. Active devices in the S switch elements reducing the resistive effect when required. The total capacitance along the line is $(4 + m)C_c + mAC_s$; larger S block switches have proportionally more capacitance. As a result, τ is

$$\tau = (4R_c + mR_c/A)\left((4 + m)C_c + mAC_s\right),$$

$$\tau = nR_cC_s\left(16 + 4m\left(1 + \frac{A}{n} + \frac{1}{A}\right) + m^2\left(\frac{1}{n} + \frac{1}{A}\right)\right), \tag{15}$$

for moderate levels of m. Typically $n >> 1$ and $A >> 1$. One would typically design A/n to be significantly less than 1; the optimum point seems to be $A = \sqrt{n}$. We would simultaneously want to minimize n and m. Minimizing n enables higher frequency, as well as lower power consumption. Minimizing m decreasing the resistance through the path.

Finally, we show evidence for the quadratic scaling law given in Figure 1b. At this optimal point,

$$\tau = nR_cC_s\left(16 + 4m\left(1 + \frac{2}{\sqrt{n}}\right) + m^2\left(\frac{1}{n} + \frac{1}{A}\right)\right) \approx 4nR_cC_s(4 + m). \tag{16}$$

By including the dependancies of R_c and $C_s = WLC_{s0}$, where C_{s0} is C_s per unit area, we get

$$R_cC_s = L^2\frac{m^*v_{sat}}{qU_TC_{s0}\epsilon_{Si}}, \tag{17}$$

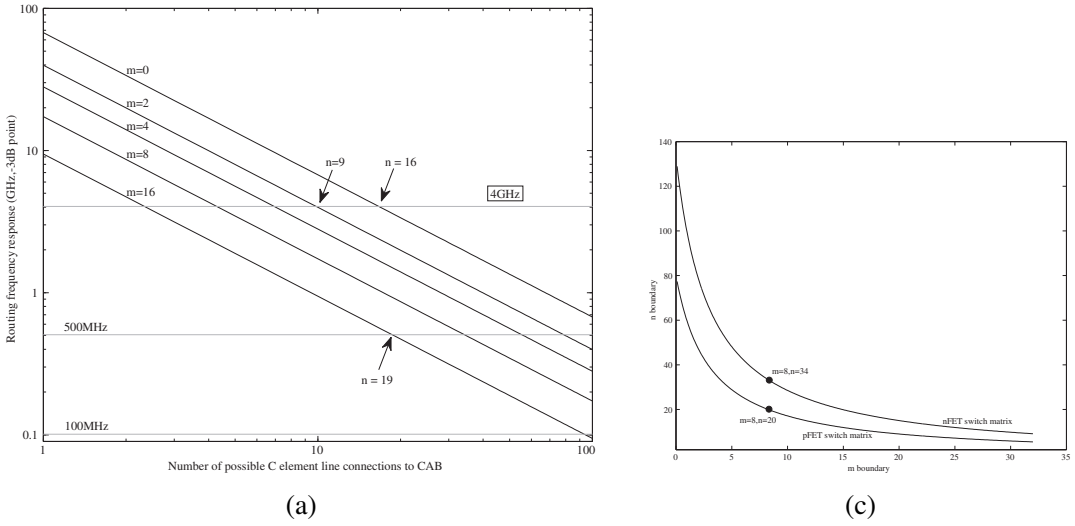
resulting in

$$\tau = L^2\frac{m^*v_{sat}}{qU_TC_{s0}\epsilon_{Si}}n\left(16 + 4m\left(1 + \frac{2}{\sqrt{n}}\right) + m^2\left(\frac{1}{n} + \frac{1}{A}\right)\right) \approx 4nR_cC_s(4 + m). \tag{18}$$

We expect a quadratic scaling of frequency with channel length for the same architecture structure, as seen in Figure 1b. We notice the size does not depend upon W, giving the designer freedom to choose W based on other constraints. The substrate doping does not change much by process starting in 130 nm process, being effectively degenerately doped. Furthermore, higher insulator FETs often use a slightly lower substrate doping for their devices. One might have seen an issue if we started FPAA development using 2 µm CMOS.

The architecture sets the frequency response. Figure 12 shows tradeoffs for a 45 nm IC CMOS process. Figure 12b solving for frequency boundary for 4 GHz frequency, and Figure 12c solving for frequency boundary for 500 MHz frequency. This process enables 4 GHz signal frequency through both routing fabric and components.

Param	Definition
R_c	maximum conductance
R_s	unbuffered *S* block switch resistance
C_s	Source-Drain capacitance to substrate or bulk
C_c	Capacitance on a CAB crossbar line or *C* block line
	$= n\ C_s$
n	MOSFETs (and s-d junctions) on a *C* block line
	number of routing tracks
m	average number of *S* block switches required



(a) (b) (c)

Figure 12. The architecture choices look at the relationship of the resulting switch resistances, or other circuit uses of the FG switch devices, as a function of the number of CAB inputs and number of tracks (*n*), as well as the typical number of switches needed for a connection (*m*). (**a**) frequency tradeoffs for FPAA architectures, both at 4 GHz, 500 MHz, and 100 MHz frequency ranges; (**b**) boundary line (4 GHz operation) between the choices for *n* and *m* for a particular frequency line for a 45 nm IC process. Boundary line for 4 GHz operation. These types of architectures could handle RF signals directly, as well as IF for very high frequency carriers (i.e., 60 GHz); and (**c**) boundary line for 500 MHz operation.

5. Conclusions

This paper presented scaling of FG devices, and the resulting implication to larger (FPAA) systems. The properties of FG circuits and systems in one technology (e.g., 350 nm CMOS) are experimentally shown to roughly translate to FG circuits in scaled down processes in a way predictable through MOSFET physics concepts. This discussion addressed the question of scaling these devices to more modern processes, in particular using the example processes of 130 nm, 40 nm CMOS, empowering moving such approaches to smaller linewidth CMOS processes. Scaling FG devices results in higher frequency response, (e.g., FPAA fabric) as well as lower parasitic capacitance and lower power consumption. An FPAA's operating frequency improves as the IC technology process is scaled down. FPAA architectures, limited to 50–100 MHz frequency ranges could be envisioned to operate at 500 MHz–1 GHz for 130 nm line widths, and operate around 4 GHz for 40 nm line widths.

Acknowledgments: The authors want to thank Brian Degnans (degs) for some of the early measurements that helped guide these discussions, as well as building a generic FG device test chip that was used for some of these measurements.

Author Contributions: Jennifer Hasler wrote the paper, did the resulting scaling device theory, FPAA scaling theory, data analysis, and was involved in data measurements for all three devices. Sihwan Kim was involved in taking the 130 nm FG data. Farhan Adil designed the 40 nm FG test cells and took a majority of the resulting data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marotta, G.G.; Macerola, A.; D'Alessandro, A.; Torsi, A.; Cerafogli, C.; Lattaro, C.; Musilli, C.; Rivers, D.; Sirizotti, E.; Paolini, F.; et al. A 3 bit/cell 32 Gb NAND flash memory at 34 nm with 6 MB/s program throughput and with dynamic 2 b/cell blocks configuration mode for a program throughput increase up to 13 MB/s. In Proceedings of the 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 19–23 February 2012.
2. Li, Y.; Lee, S.; Fong, Y.; Pan, F.; Kuo, T.C.; Park, J.; Samaddar, T.; Nguyen, H.; Mui, M.; Htoo, K.; et al. A 16 Gb 3 b/Cell NAND Flash Memory in 56 nm with 8 MB/s Write Rate. In Proceedings of the 2008 IEEE International Solid-State Circuits Conference—Digest of Technical Papers, San Francisco, CA, USA, 3–7 February 2008.
3. Harari, E. Flash Memory—The Great Disruptor! In Proceedings of the 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 19–23 February 2012.
4. Shibata, N.; Kanda, K.; Hisada, T.; Isobe, K.; Sato, M.; Shimizu, Y.; Shimizu, T.; Sugimoto, T.; Kobayashi, T.; Inuzuka, K.; et al. A 19 nm 112.8 mm^2 64 Gb Multi-Level Flash Memory with 400 Mb/s/pin 1.8 V Toggle Mode Interface. In Proceedings of the 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 19–23 February 2012.
5. Li, Y.; Lee, S.; Oowada, K.; Nguyen, H.; Nguyen, Q.; Mokhlesi, N.; Hsu, C.; Li, J.; Ramachandra, V.; Kamei, T.; et al. 128 Gb 3b/Cell NAND Flash Memory in 19 nm Technology with 18 MB/s Write Rate and 400 Mb/s Toggle Mode. In Proceedings of the 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 19–23 February 2012.
6. Lee, D.; Chang, I.J.; Yoon, S.Y.; Jang, J.; Jang, D.; Hahn, W.; Park, J.; Kim, D.; Yoon, C.; Lim, B.; et al. A 64 Gb 533 Mb/s DDR Interface MLC NAND Flash in Sub-20 nm Technology. In Proceedings of the 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 19–23 February 2012; pp. 430–431.
7. George, S.; Kim, S.; Shah, S.; Hasler, J.; Collins, M.; Adil, F.; Wunderlich, R.; Nease, S.; Ramakrishnan, S. A Programmable and Configurable Mixed-Mode FPAA SoC. *IEEE Trans. VLSI* **2016**, *24*, 2253–2261.
8. Srinivasan, V.; Serrano, G.J.; Gray, J.; Hasler, P. A precision CMOS amplifier using floating-gate transistors for offset cancellation. *IEEE J. Solid-State Circuits* **2007**, *42*, 280–291.
9. Srinivasan, V.; Serrano, G.; Twigg, C.; Hasler, P. A Floating-Gate-Based Programmable CMOS Reference. *IEEE Trans. Circuits Syst. I* **2008**, *55*, 3448–3456.
10. Kim, S.; Hasler, J.; George, S. Integrated Floating-Gate Programming Environment for System-Level ICs. *IEEE Trans. VLSI* **2016**, *24*, 2244–2252.
11. Carley, L.R. Trimming analog circuits using floating-gate analog MOS memory. *IEEE J. Solid-State Circuits* **1989**, *24*, 1569–1575.
12. Sackinger, E.; Guggenbuhl, W. An analog trimming circuit based on a floating-gate device. *IEEE J. Solid-State Circuits* **1988**, *23*, 1437–1440.
13. Bleiker, C.; Melchior, H. A four-state EEPROM using floating- gate memory cells. *IEEE J. Solid-State Circuits* **1987**, *22*, 460–463.
14. Nozama, H.; Kokyama, S. A thermionic electron emission model for charge retention in SAMOS structures. *Jpn. J. Appl. Phys.* **1992**, *21*, 111–112.
15. Hasler, P.; Minch, B.; Diorio, C. Adaptive circuits using pFET floating-gate devices. In Proceedings of the IEEE 20th Advanced Research in VLSI , Atlanta, GA, USA, 21–24 March 1999; pp. 215–229.
16. Hasler, P.; Basu, A.; Kozoil, S. Above threshold pFET injection modeling intended for programming floating-gate systems. In Proceedings of the 2007 IEEE International Symposium on Circuits and Systems, New Orleans, LA, USA, 27–30 May 2007.
17. Hasler, P.; Diorio, C.; Minch, B.A.; Mead, C.A. Single transistor learning synapses. In *Advances in Neural Information Processing Systems (NIPS) 7*; Tesauro, G., Touretzky, D.S., Leen, T.K., Eds.; MIT Press: Cambridge, MA, USA, 1995; pp. 817–824.

18. Lenzlinger, M.; Snow, E.H. Fowler–Norrdheim tunneling into thermally grown SiO_2. *J. Appl. Phys.* **1969**, *40*, 278–283.
19. Mead, C. Scaling of MOS technology to sub micrometer feature sizes. *J. VLSI Signal Process.* **1994**, *8*, 9–25.
20. Nicollian, E.H.; Brews, J.R. *MOS Physics and Technology*; Wiley Interscience: New York, NY, USA, 1982.
21. Hasler, P.; Andreou, A.; Diorio, C.; Minch, B.A.; Mead, C. Impact ionization and hot-electron injection derived consistently from Boltzman transport. *VLSI Des.* **1998**, *8*, 455–461.
22. Minch, B.; Hasler, P. A floating-gate technology for digital CMOS processes. In Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, Orlando, FL, USA, 30 May–2 June 1999.
23. Puthenkovilakam, R.; Chang, J.P. An accurate determination of barrier heights at the $HfO_2/SiHfO_2Si$ interfaces. *J. Appl. Phys.* **2004**, *96*, doi:10.1063/1.1778213.
24. Duffy, C.; Hasler, P. Scaling pFET hot-electron injection. *Int. Workshop Comput. Electron.* **2004**, *3*, 149–150.
25. Shockley, W. Problems related to p-n junctions in silicon. *Solid State Electron.* **1961**, *2*, 35–67.
26. Luu, J.; Goeders, J.; Wainberg, M.; Somerville, A.; Yu, T.; Nasartschuk, K.; Nasr, M.; Wang, S.; Liu, T.; Ahmed, N.; et al. VTR 7.0: Next Generation Architecture and CAD System for FPGAs. *ACM Trans. Reconfig. Technol. Syst.* **2014**, *7*, doi:10.1145/2617593.