GDOT: A Graphene-Based Nanofunction for Dot-Product Computation

Ning C. Wang*, Sujan K. Gonugondla†, Ihab Nahlus†, Naresh R. Shanbhag†, Eric Pop*
*Electrical Engineering, Stanford Univ., Stanford, CA 94305, USA. E-mail: ningwang@stanford.edu
†Electrical & Computer Engineering, Univ. Illinois Urbana-Champaign, Urbana IL, 61801, USA.

Abstract

Though much excitement surrounds two-dimensional (2D) beyond CMOS fabrics like graphene and MoS_2, most efforts have focused on individual devices, with few high-level implementations. Here we present the first graphene-based dot-product nanofunction (GDOT) using a mixed-signal architecture. Dot product kernels are essential for emerging image processing and neuromorphic computing applications, where energy efficiency is prioritized. SPICE simulations of GDOT implementing a Gaussian blur show up to ~10^4 greater signal-to-noise ratio (SNR) over CMOS based implementations — a direct result of higher graphene mobility in a circuit tolerant to low on/off ratios. Energy consumption is nearly equivalent, implying the GDOT can operate faster at higher SNR than CMOS counterparts while preserving energy benefits over digital implementations. We implement a prototype 2-input GDOT on a wafer-scale 4" process, with measured results confirming dot-product operation and lower than expected computation error.

Introduction

Graphene has garnered much interest due to a combination of unique electrical, mechanical, and thermal attributes. Yet, its use in circuits has proven elusive as the absence of a band gap precludes digital application and low gain restricts RF use [1]. As novel computing methods emerge, new circuits can benefit from graphene. A potential candidate is the CMOS dot product kernel suggested by [2], which uses a switched-analog circuit (SAC) to convert time- and voltage-domain inputs into a dot product output voltage. Here we show that such mixed-signal nanofunctions benefit from features unique to graphene, offering higher performance compared to a CMOS implementation and energy savings compared to purely digital dot products.

GDOT Operation

Fig. 1 outlines basic GDOT operation with graphene FETs (GFETs); a full description of SAC dot-product is found in [2]. The GDOT kernel has N identically sized GFETs connected at the source. Weights are encoded as input pulses of varying duty cycle ($p_i = T_{ON,i}/T$), while vector inputs are encoded as DC voltage values. As the circuit cycles through the GDOT branches, the varying weights determine charge contribution as a function of time [**Fig. 1b**]. The *R'C'* output serves both as low-pass filter and charge collector for the dot-product computation. The output voltage V_{out} is proportional to the dot-product so long as $\tau = R'C' \gg T$ and $\tau \gg \tau_{GFET}$ (the device delay). Both error (E) and noise are proportional to: 1) T/τ, 2) τ_{GFET}/τ, and 3) $R_{HL} = R_{HIGH}/R_{LOW}$ (GFET high and low resistance ratio). These relationships represent a fundamental tradeoff in computational accuracy vs. speed, with use of GFETs impacting the latter two.

Simulation Results

We compare the GDOT vs. a CMOS dot-product, simulating a Gaussian blur filter [$\sigma^2 = 0.85$] in SPICE using equivalently sized [$L = 180$ nm, $W = 1$ μm] transistors. For GFETs, we use the MIT Virtual Source model [3] fitted to our better experimental devices (described below). Weight generation is handled using a 16-stage ring counter operating at $T = 2$ ns. $C' = 2$ pF and R' is varied from 10-100 kΩ, yielding τ = 20-200 ns. Simulation results of V_{out} vs. time [**Fig. 2**] show comparable output for GDOT and CMOS at $\tau = 200$ ns [**Fig. 2a**], validating GDOT operation. The CMOS circuit starts to exhibit increased noise and error, shown in **Fig. 2b** (τ = 20 ns), **Figs. 3** and **4** where %Error = $(V_{desired} - V_{out})/V_{desired}$ and SNR = $(V_{out}/\sigma)^2$ vs. τ. For CMOS, both figures increase rapidly when $\tau \ll 100$ ns, suggesting the CMOS pass-transistors begin to dominate the circuit. In contrast, GDOT figures remain relatively stable even at the shortest τ, albeit at %Error larger than CMOS unless the input range [$\beta = \max(V_i)/\min(V_i)$] is reduced (a direct consequence of a low R_{HL}). The resiliency of GDOT noise figures to increased circuit speed is due to: 1) higher mobility and 2) low R_{HL}, which suppresses the output voltage swing and subsequently the noise. **Fig. 5** shows estimated energy/operation = $P_{avg}\tau_{op}$ for the dot-product kernel, where $\tau_{op} = 5\tau$. The GDOT has higher efficiency, likely due to reduced charging/discharging activity from lower R_{HL}. Note energy/op. decreases with τ, implying the simulated values are not fully optimized.

Experimental Results

We fabricate a prototype 2-input GDOT using 4" wafer-scale graphene grown by chemical vapor deposition (CVD), optical lithography, and top-gated GFETs [**Fig. 6**]. Pulsed measurements show closely matched devices [**Fig. 7**] with extracted parameters listed in **Table 1**. Pulsed measurements limit GFET hysteresis [4], thus GDOT operation benefits from graphene advantages despite non-ideal dielectrics. Measured results for $\tau = 170$ μs and $T = 10$ μs [**Fig. 8**] validate experimental GDOT operation. We intentionally chose long measurement times to circumvent measurement parasitics and generate stable dot-product output. We sweep both: 1) p_1, holding V_1 and V_2 fixed [**Fig. 9**] and 2) V_1 and V_2 while holding $p_1 = 0.7$ and $p_2 = 0.3$ [**Fig. 10**]. Increasing pulsed amplitude (V_{pp}) yields lower error [**Fig. 9b**], but exhibits diminishing returns for $V_{pp} = 1.8$ V due to GFET contact resistance which is not fully optimized here. SPICE simulations actually overestimate circuit error [**Fig. 10b**], and instead we develop a behavioral model [**Fig. 10c**] to more accurately predict circuit output [dashed line in **Fig. 10a**]. Note error is kept below 12% for $\beta < 2$ [**Fig. 10b**], experimentally demonstrating the restriction on operation from a low R_{HL}.

Conclusion

The GDOT nanofunction represents a new application for graphene with performance exceeding CMOS, as demonstrated by simulation results that show superior noise and speed in exchange for a small error increase. The GDOT also takes advantage of properties that make graphene special (e.g. higher mobility) while tolerating its drawbacks (e.g. lack of band gap). We experimentally implemented a 2-input GDOT exhibiting surprisingly low error, suggesting greater than anticipated benefits from further GFET scaling. The GDOT is a significant first step towards realizing a dot-product network based on 2D beyond CMOS nanomaterials, which could also leverage their native transparency for image processing applications. This work has been supported by the STARNet SONIC Center.

References

[1] R. Grassi, *et al.*, *Solid. State. Electron.* **100**, pp. 54–60 (2014).
[2] I. Nahlus, E. Kim, N. Shanbhag, D. Blaauw, *ISLPED 2014*.
[3] S. Rakheja, *et al.*, *IEEE Trans. Nano*. **13**, pp. 1–23 (2014).
[4] E. Carrion, *et al.*, *IEEE TED*, **61**, pp. 1583–1589 (2014).

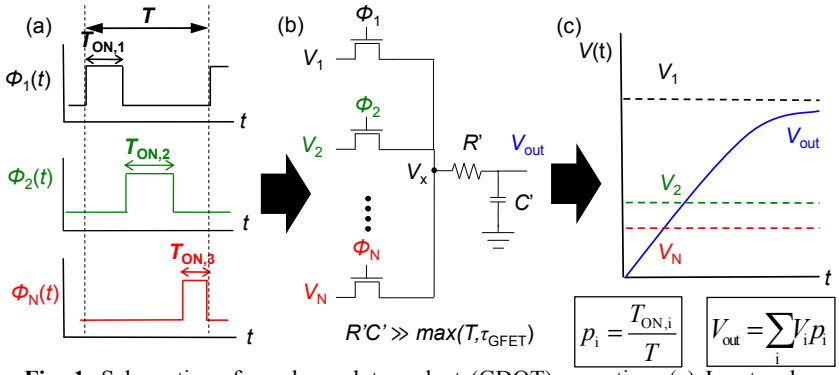


Fig. 1. Schematics of graphene dot product (GDOT) operation. (a) Input pulse weights. (b) Dot-product kernel transistor-level schematic. (c) Ideal output voltage.

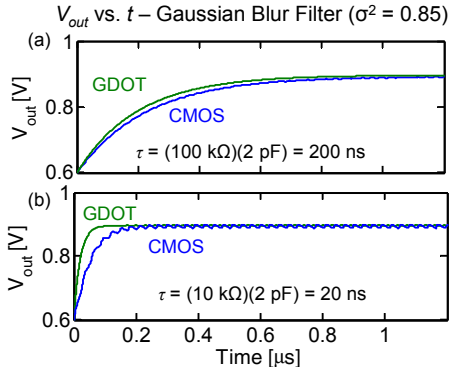


Fig. 2. GDOT vs. CMOS SAC simulation (L= 180 nm, W=1 μm) for (a) τ = 200 ns and (b) 20 ns.

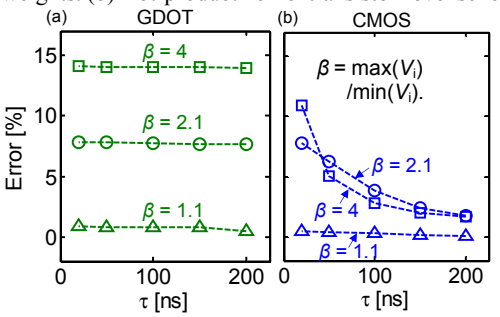


Fig. 3. Simulated % Error vs. τ for different input voltage ranges for (a) GDOT and (b) CMOS dot product implementation.

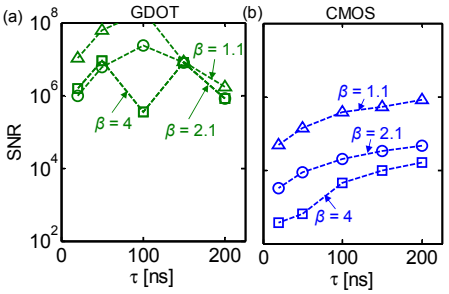


Fig. 4. Signal-to-Noise Ratio [SNR=($V_{desired}/\sigma$)²] vs. τ for (a) GDOT and (b) CMOS dot product implementation.

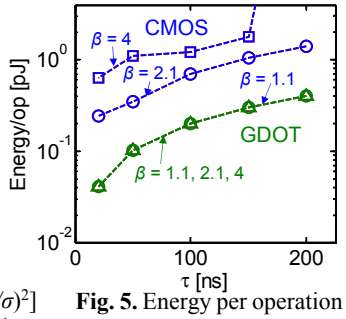


Fig. 5. Energy per operation [$\tau_{op} = 5\tau$] vs. τ.

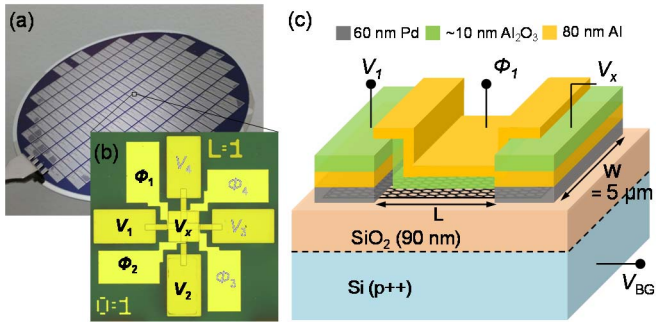


Fig. 6. (a) Wafer-scale (4") GDOT fabrication. (b) Optical image of prototype 4-input L = 1 μm GDOT. Only 2-input (bold) used due to test equipment limitations. (c) Representative GFET cross section [4].

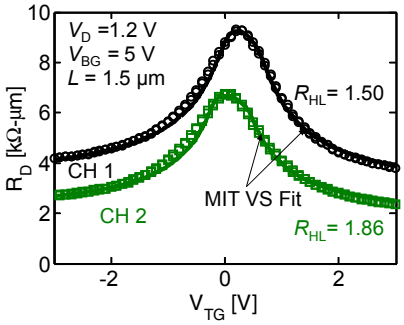
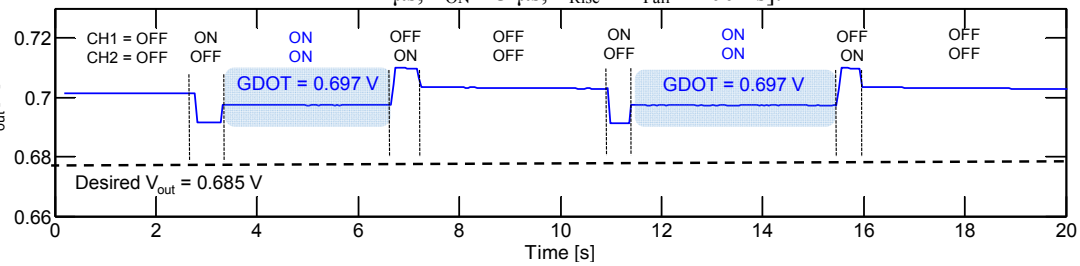


Fig. 7. Pulsed device resistance [R_D] vs. top-gate bias [V_{TG}] (symbols) and model fit (lines) for fabricated GFETs [T = 10 μs, T_{ON} = 3 μs, T_{Rise} = T_{Fall} = 200 ns].

Table 1. Values extracted from VS Model fit [3] for GFETs connected to IN1 (i.e. p_1, V_1) and IN2 (i.e. p_2, V_2).

GFET	IN1	IN2
μ [$cm^2V^{-1}s^{-1}$]	960	1370
$R_{C,elec}$ [kΩ-μm]	1.3	0.8
$R_{C,hole}$ [kΩ-μm]	1.6	0.9

Fig. 8. Measured V_{out} vs. time (V_1 = 0.665 V, p_1 = 0.7, V_2 = 0.7321 V, p_2 = 0.3). CH 1 and CH 2 refer to manually triggered pulse generator outputs.



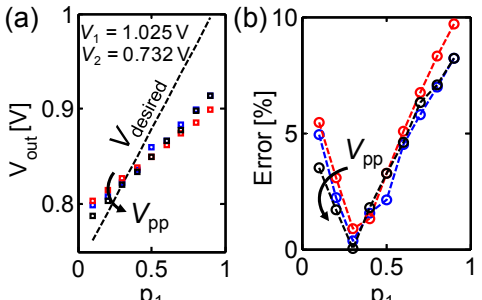


Fig. 9. Measured (a) V_{out} and (b) % Error vs. input-weight, with V_{pp} = 0.6, 1.2, 1.8 V.

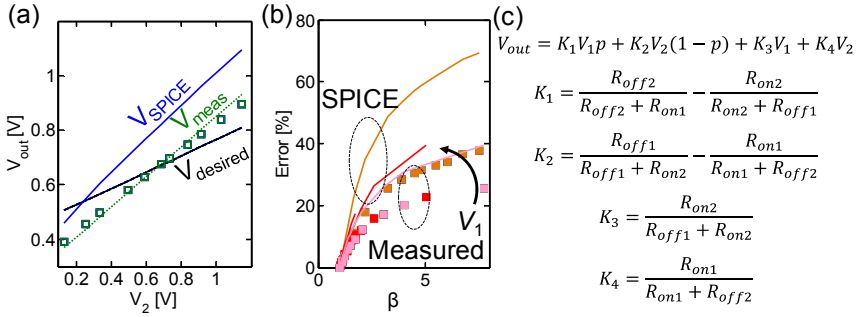


Fig. 10. (a) V_{out} vs. V_2 with V_1 = 0.665 V. (b) % error vs. β. (c) Behavioral model describing 2-input GDOT output with $p = p_1$.