Beyond-Silicon Computing: Nano-Technologies, Nano-Design, and Nano-Systems

Gage Hills

1 Introduction

For decades, humankind has enjoyed the energy efficiency benefits of scaling transistors smaller and smaller, but these benefits are waning. In a worldwide effort to continue improving computing performance, many researchers are exploring a wide range of technology alternatives, ranging from new physics (spin-, magnetic-, tunneling-, and photonic-based devices) to new nanomaterials (carbon nanotubes, two-dimensional materials, superconductors) to new devices (non-volatile embedded memories, ferroelectric-based logic and memories, q-bits) to new systems, architectures, and integration techniques (advanced die- and wafer-stacking, monolithic three-dimensional (3D) integration, on-chip photonic interconnects). However, developing new technologies from the ground up is no simple task, and requires an end-to-end approach addressing many challenges along the way. First of all, a detailed analysis of the overall potential benefits of a new technology is essential; it can take years to bring a new technology to the level of maturity required for high-volume production, and so a team of researchers must ensure upfront that they are developing the right technologies for the right applications. For example, many emerging nanotechnologies are subject to nano-scale imperfections and variations in material properties—how does one overcome these challenges at a very-large scale? Will new design techniques be required? Will circuit and system designers even use the same approaches to designing next generation systems, or would an entirely different approach offer much better results? What level of investment will be required to develop these new technologies, designs, and systems, and at the end of the day, will the outcome be worth the effort? These are just examples of the some of the major questions that are essential to consider as early as possible.

G. Hills (✉)
Harvard University, Cambridge, MA, USA
e-mail: ghills@g.harvard.edu

© Springer Nature Singapore Pte Ltd. 2023

M. M. S. Aly and A. Chattopadhyay (eds.), *Emerging Computing: From Devices to Systems*, Computer Architecture and Design Methodologies,
https://doi.org/10.1007/978-981-16-7487-7_2

To provide a concrete example of how many of these questions are being addressed in practice, in this chapter, we take a deep dive into one specific research area: using carbon nanotubes (CNTs) as the channel material of field-effect transistors, and the resulting next-generation systems that are enabled. We start by providing an overview of the potential benefits of this emerging nanotechnology, as well as the major challenges that have blocked significant progress in the field. We then describe example design techniques that have been used to overcome these challenges (i.e., nano-design techniques), and offer experimental demonstrations of larger-scale circuits that have been enabled using these techniques in practice, and that are now being developed inside commercial manufacturing facilities. Next, we illustrate how carbon nanotube field-effect transistors (CNFETs, and other nano-technologies with similar physical properties) enable entirely new types of computing systems that are impossible to build using today's silicon-based technologies, namely, monolithic three-dimensional (3D) integrated systems with multiple layers of transistors and multiple layers of memories fabricated directly on top of each other with ultra-dense vertical connectivity. We close by summarizing the potential that these 3D "nano-*systems*" have to extend progress in energy-efficient computing, and finally offer examples of how they can be combined with advances higher up the computing stack (e.g., with new computer architectures, compilers, or domain-specific programming languages) for even larger benefits. We hope that this technological journey spanning nano-technologies, nano-design, and nano-systems will motivate the reader to pursue further investigation into these cutting-edge research areas.

2 Emerging Nanotechnologies: Opportunities and Challenges

Quantifying the potential benefits of a new technology is essential to guide development of the *right* technologies for the *right* applications. Importantly, since today's systems often consist of multiple heterogenous components (including processor cores, on- and off-chip memories, power distribution, etc.), small-scale benchmarking does not capture important interactions between these various components that can limit overall system performance. For example, analyzing the delay and power consumption of a stand-alone 32-bit adder does not account for many effects present in realistic very-large-scale integrated (VLSI) systems (e.g., interconnect routing parasitics, application-level timing constraints, process variations), and does not perform technology-specific optimization of key circuit-level design parameters (e.g., total circuit area, target clock frequency). This can lead to incorrect conclusions, and thus, wasted efforts in pursuit of developing the wrong technologies. Instead, an end-to-end evaluation framework is essential, including: (a) energy-efficient circuit-/system-level techniques to overcome inherent imperfections and variations, (b) full physical design of VLSI systems, and (c) variation-aware power/timing design and

optimization, calibrated to experimental data, running real applications, and meeting circuit-level yield, test, and noise immunity constraints.

In this section, we provide an overview of such an analysis framework that compares the energy efficiency benefits of multiple promising technology candidates (shown in Fig. 1) for future technology nodes. This analysis uses complete physical designs of VLSI processor cores in order to account for the realistic circuit effects. Energy efficiency is quantified by the Energy-Delay Product metric (EDP), i.e., the product of total circuit energy consumption and the maximum propagation delay from any input to any output (the critical path delay). Importantly, this comparison leverages industry-practice VLSI designs and design flows, as well as technology parameters that calibrated to experimental data to, not only to quantify the EDP benefits of each technology, but also provide insight into the sources of their benefits.

Using the VLSI design flow shown in Fig. 2, this approach demonstrates that CNFETs offer major energy efficiency benefits for sub-10 nm node digital VLSI circuits. For additional details, we refer the reader to Hills et al. (2018), which also describes how these benefits can be maintained even in the presence of major variations in CNT processing (which we also describe in Sect. 3). While this methodology has so far been used to evaluate the benefits of CNFETs, it can also be extended to evaluate new combinations of technologies moving forward.

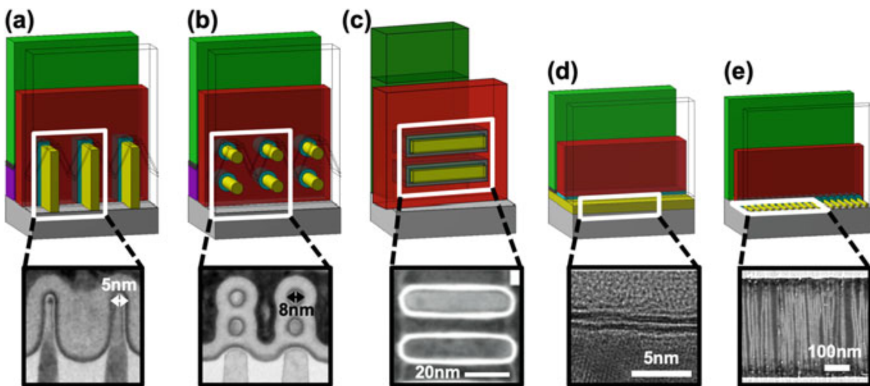


Fig. 1 Advanced technology options for field-effect transistors (FETs) for future technology nodes. For each FET, the drain contact is transparent in order to illustrate a cross-section of the transistor channel, with microscopy images provided underneath each 3D-rendered FET. **a** FinFET with multiple fins. **b** Nanowire FET with multiple nanowires both horizontally and vertically (Mertens et al. 2016). **c** Nanosheet FET with multiple nanosheets integrated vertically on top of each other (Loubet et al. 2017). **d** Two-dimensional (2D) material FET, in which the FET channel can be made of 2D materials such as MoS_2, black phosphorus, or WSe_2. **e** Carbon nanotube FET (CNFET), with multiple carbon nanotubes (CNTs) comprising the transistor FET channel, shown in the top-view scanning electron microscopy (SEM) image

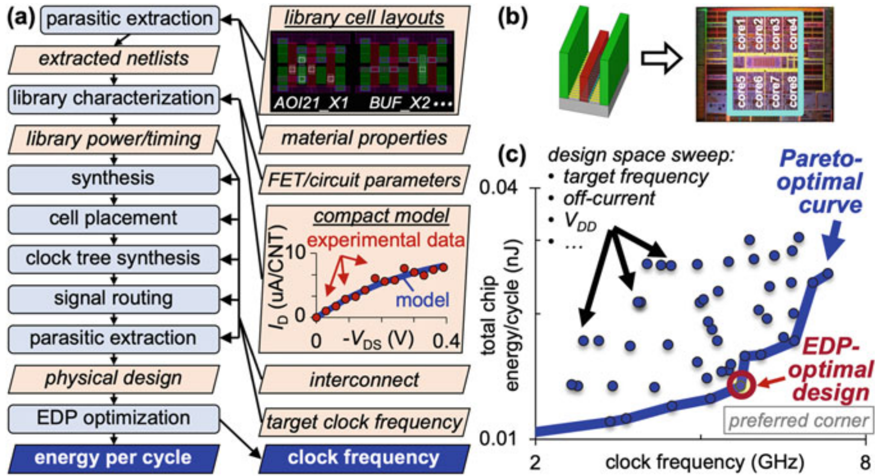


Fig. 2 End-to-end approach to quantify the benefits of emerging technologies for VLSI scale circuits. **a** VLSI design and analysis flow. Key components include: experimentally calibrated compact models (as shown for CNFET drain current vs. drain-to-source voltage: I_D vs. V_{DS}, enabling accurate circuit simulations) (Lee et al. 2015), library cell layouts (enabling extraction of parasitic resistance and capacitance elements), and circuit-level EDP optimization. **b** Example illustration of a single CNFET, which is used in conjunction with this design flow to analyze the energy and delay of the OpenSparc T2 processor core (OpenSPARC 2011) designed using CNFETs. **c** The EDP-optimal design is selected from the Pareto-optimal trade-off curve to quantify the energy efficiency of each technology

2.1 VLSI Circuit Benefits of One-Dimensional and Two-Dimensional Nanomaterials

While quantifying EDP benefits at the circuit- and system-level is certainly required to drive technology development, e.g., for motivating the use of CNFETs, it is equally important to understand where these benefits are coming from. Using the detailed analysis flow in Fig. 2 provides valuable insight into the sources of these benefits for CNFETs. In particular, a useful metric for evaluating the potential circuit-level benefits of a FET technology is the electrostatic scale length (Frank et al. 1998), which quantifies how susceptible that FET is to short-channel effects (Kuhn 2012). The scale length should be small to enable shorter gate lengths, thus reducing the energy required to charge the FET gate capacitance without degrading the FET ability to quickly turn on and off by modulating the gate voltage (quantified by the sub-threshold slope). Well-known approaches for improving the scale length include: (1) improve FET geometry (e.g., changing from FinFET to gate-all-around Nanowire FET or Nanosheet FET), and (2) reduce the semiconductor body thickness (e.g., reducing the thickness of fins in a FinFET or reducing the diameter of individual nanowires in a nanowire FET). While evolving from today's silicon–germanium (SiGe)-based FinFETs to gate-all-around nanowire FETs or nanosheet FETs reduces

FET scale length, continuing to reduce scale length requires reducing the semiconductor body thickness. Unfortunately, reducing the body thickness can have unwanted side effects. In particular, bulk materials (e.g., all Si-, Ge-, and III-V-based semiconductors) suffer from severely degraded carrier transport as the body thickness scales to sub-10 nm dimensions (Gomez et al. 2007; Hashemi et al. 2014a, b; Sasaki et al. 2015; Suk et al. 2007; Uchida et al. 2003). This degradation in carrier transport arises from increased photon scattering and surface roughness, which significantly degrades FET on-current density and thus overall circuit speed. The unfortunate result is that it has become extremely challenging for technologists to create FETs that exhibit both excellent scale length and also excellent carrier transport simultaneously.

To alleviate this challenge, many technologists have turned to alternative "low-dimensional" materials, i.e., 2D materials and one-dimensional (1D) carbon nanotubes, which inherently maintain superior carrier transport even with very thin body thickness. For example, experimental measurements quantifying carrier transport in CNTs includes hole mobility exceeding 2,500 cm^2/V s (Zhou et al. 2005) and hole velocity of 4.1×10^7 cm/s, even for CNT diameter below 2 nm. For reference, measurements of experimental Si FinFET with body thickness less than 3 nm exhibit mobility under 300 cm^2/V s. Because of these material- and device-level benefits of CNTs (and other low-dimensional materials), there are significant energy efficiency benefits at the circuit level, and we refer the reader to Hills et al. (2018) for quantified results. Key implications for gaining high-level intuition include:

- Superior CNT carrier transport enables CNFET circuits to operate with reduce supply voltage with simultaneously higher effective drive current (I_{EFF}) compared to SiGe FinFET (e.g., 20% lower V_{DD} with 25% higher I_{EFF} for the same off-state leakage current density).
- Ultra-thin body thickness (CNT diameter) results in very short scale length, enabling experimental CNFETs maintain steep sub-threshold slope (SS) with extremely-scaled gate length (e.g., SS = 70 mV/decade with 5 nm gate length, which has been shown for both PMOS and NMOS CNFETs experimentally (Qiu et al. 2017)).
- Optimized CNFET circuits exhibit lower total circuit capacitance, which reduces overall energy consumption and also contributes to higher circuit speeds (e.g., 2 × lower capacitance for projected CNFET vs. Si/SiGe FinFET (Hills et al. 2018)). This reduction in capacitance comes from multiple sources. First, shorter CNFET gate length not only reduces intrinsic gate-to-channel capacitance (due to smaller gate area), but also reduces parasitic gate-to-source/drain capacitance due to increased physical separation between the gate and the source/drain contacts. Second, high CNFET drive current enables electronic design automation (EDA) tools for logic synthesis/place-and-route to automatically select standard library cells with smaller drive strengths and still meet circuit-level timing constraints. And third, these constraints can be met even with using planar CNFETs, which have lower gate capacitance compared to three-dimensional FinFETs, nanowire FETs, and nanosheet FETs, whose channels extend vertically above the substrate to increase drive current at the cost of higher parasitic capacitance (Hills et al. 2018).

2.2 Inherent Challenges in Emerging Nanotechnologies

Despite the projected benefits of emerging nanotechnologies, such as CNFETs and FETs based on other low-dimensional nanomaterials, there are significant practical challenges that must be overcome before these benefits can be realized. Specifically, emerging nanotechnologies are inherently subject to nano-scale imperfections and process variations, and without dedicated techniques to specifically address these challenges at the fabrication-, design-, and system-levels, affected nanotechnologies may never see the light of day. As an illuminating example, key challenges that have plagued CNTs for decades include:

- *CNT aggregates*—during CNT deposition, i.e., when CNTs are deposited on the wafer substrate used for circuit fabrication, CNTs can "bundle" together forming "CNT aggregates" (an example image is shown in Fig. 4a). The presence of CNT aggregates in CNFET channel regions can lead to incorrect CNFET functionality, reducing overall CNFET circuit yield (Hills et al. 2019).
- *CNT CMOS*—today's energy-efficient digital circuits rely on having a *complementary* metal–oxide–semiconductor (CMOS) process that includes both PMOS and NMOS FETs. However, many emerging FET technologies, including CNFETs, have lacked a robust CMOS process. In particular, both PMOS and NMOS CNFETs should: (a) be air-stable, (b) have tunable electrical characteristics (e.g., threshold voltage), and (c) have limited variability (Hills et al. 2019; Lau et al. 2018).
- *Metallic CNTs*—due to a lack of precise control CNT properties (e.g., diameter and chirality), CNTs can be either semiconducting (s-CNT) or metallic (m-CNTs); m-CNTs exhibit little or no bandgap, and so their conductance cannot be effectively modulated by the CNFET gate, leading to increased off-state leakage current and potentially incorrect logic functionality in CNFET circuits (Hills et al. 2015, 2019; Zhang et al. 2009).
- *CNT variations*—in addition to metallic CNTs, CNTs exhibit additional variations in CNT density, CNT diameter, alignment, and doping (Fig. 3a). CNT variations can lead to near-zero functional yield, increase susceptibility to noise, and degrade EDP benefits of CNFET digital circuits. But a key question is: which of these variations actually matter from a *system* point-of-view, which is ultimately what we care about? Without a systematic methodology to evaluate the system-level impact of CNT variations, one might blindly pursue difficult CNT processing paths with diminishing returns, while overlooking other CNT process advances that enable far larger yield and performance benefits overall. This challenge is further exacerbated by the fact that CNT processing advances can also be combined with CNFET circuit design techniques to reduce the impact of CNT variations (e.g., selective transistor upsizing), which lead to massive design spaces that can be intractable to explore (Fig. 3b); for example, existing approaches rely on trial-and-error-based ad hoc techniques that can be prohibitively time consuming (e.g., requiring computation runtimes exceeding 1.5 months (Hills et al. 2015)).

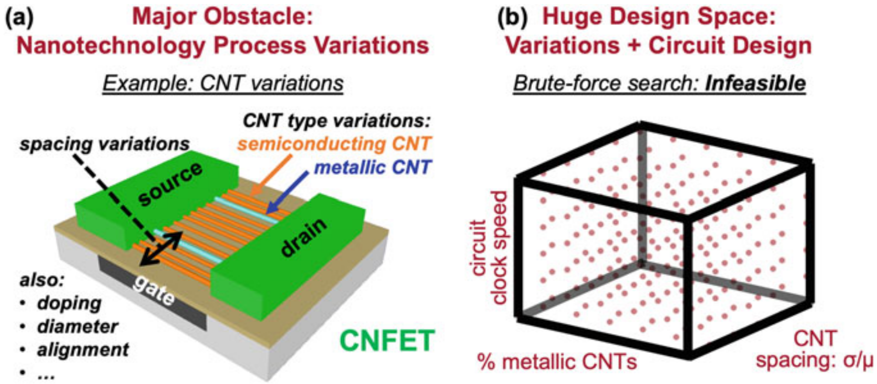


Fig. 3 Nanotechnology challenges. **a** CNT variations, including variations in CNT type, density, diameter, alignment, and doping (Hills et al. 2015). **b** Subset of the massive design space to explore in order to co-optimize CNT process improvements (e.g., in the percentage of metallic CNTs, or variations in CNT spacing) and CNFET circuit design parameters (e.g., target clock speed). Note that, these three dimensions only represent a small subset of the entire design space; e.g., for each one of these points, we can also use circuit-level techniques (such as selective transistor upsizing) to reduce the impact of CNT variations at the cost of increased energy consumption

3 Overcoming Challenges: Coordinated Nano-Fabrication + Nano-Design

Isolated improvements in processing or design are insufficient for overcoming the challenges in Sect. 2. Instead, in this section, we start by describing the essential interplay between advances in nano-fabrication and nano-design that are essential for overcoming CNT aggregates, CNT CMOS, metallic CNTs, and CNT variations in an energy-efficient and computationally-efficient manner (Sects. 3.1, 3.2 and 3.3). Section 3.4 presents experimental demonstrations of larger-scale CNFET circuits that have now been realized, showing that these techniques work in practice. In Sect. 3.5, we highlight that many of these techniques are now being transferred to multiple high-volume commercial manufacturing facilities.

3.1 VLSI CNFET Nano-Fabrication

Figure 4 illustrates two nanofabrication techniques that are used to address two of the key challenges described in Sect. 2, i.e., removing CNT aggregates and realizing a robust CMOS process for CNFETs. Specifically, *RINSE* (**R**emoval of **I**ncubated **N**anotubes through **S**elective **E**xfoliation) reduces the number of CNT aggregates per unit area by 250× (Hills et al. 2019), and *MIXED* (**M**etal **I**nterface engineering **crossed** with **E**lectrostatic **D**oping) realizes a VLSI-compatible CMOS process for

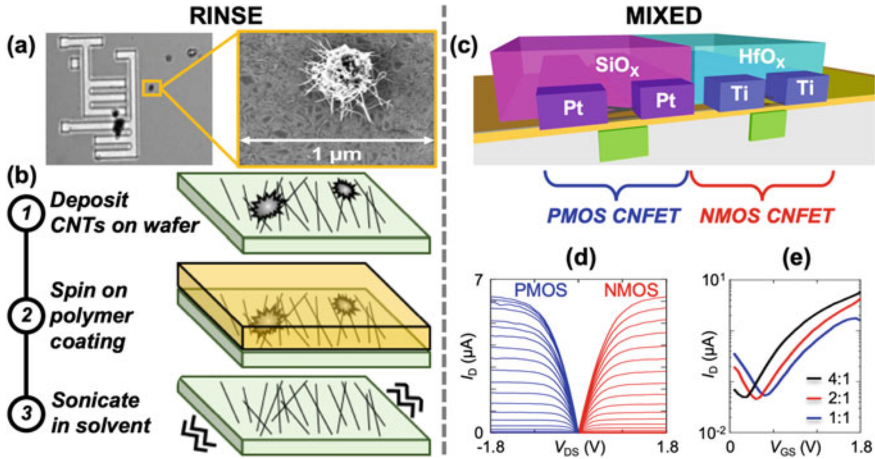


Fig. 4 Summary of RINSE and MIXED to overcome CNT aggregates and to enable CNT CMOS. **a** Scanning electron microscopy (SEM) images of a CNT aggregate on the wafer. **b** 3-step RINSE process to remove CNT aggregates, resulting in >250× reduction in the number of CNT aggregates per unit area. **c** Schematic illustration of a PMOS CNFET and an NMOS CNFET using the MIXED process flow. Here, PMOS CNFETs have Platinum source/drain contacts and SiO_X doping oxide, and NMOS CNFETs have Titanium source/drain contacts and HfO_X doping oxide. **d** Experimentally measured drain current (I_D) vs. drain-to-source voltage (V_{DS}) characteristics from fabricated CNFETs indicating similar drive current for PMOS and NMOS CNFETs. For NMOS (shown in red), the upper-most curve is measured with gate-to-source voltage (V_{GS}) of 1.8 V with V_{GS} decreasing in steps of 0.1 V for each subsequent curve. For PMOS (shown in blue), the upper-most curve is measured with $V_{GS} = -1.8$ and increasing in steps of 0.1 V for each subsequent curve. **e** I_D vs. V_{GS} (with $V_{DS} = 1.8$ V) for an NMOS CNFET, showing the ability change the threshold voltage (i.e., to horizontally shift the I_D vs. V_{GS} curve) by controlling the stoichiometry of the doping oxide. The ratios shown in the legend ("4:1", "2:1", and "1:1") indicate the relative number of Hafnium (Hf) pulses to Oxygen (O) pulses during HfO_X deposition to control the stoichiometry of the oxide (Lau et al. 2018)

CNFETs that is air-stable, electrically tunable, and robust (Hills et al. 2019). Details for RINSE and MIXED are provided below:

- *RINSE*—To enable CNFET circuit fabrication, CNTs must be uniformly deposited across the entire wafer. This can be achieved via solution processing, in which (150 mm) wafers are submerged in solutions that contain dispersed CNTs. Unfortunately, this CNT deposition technique can result in CNT aggregates deposited randomly across the wafer, which are considered manufacturing defects that act as particle contamination and thus reduce die yield. Existing techniques that attempt to remove CNT aggregates in the solution, i.e., *before* deposition, such as high-power sonication, centrifugation or excessive filtering prior to deposition, are insufficient to meet strict yield requirements for large-scale systems or to remove CNT aggregates without damaging CNTs, thus degrading CNFET performance (e.g., CNFET on-current density). Instead, by applying RINSE, we are able to

selectively remove CNT aggregates *after* deposition, without damaging the non-aggregated CNTs. Figure 4b illustrates the 3-step process for RINSE, including: Step 1: deposit CNTs on the wafer (by submerging wafers pretreated with a CNT adhesion promoter in a pre-dispersed CNT solution. Step 2: spin-coat a standard photoresist (polymethylglutarimide) onto the wafer and curing it at ~200 °C. Step 3: place the wafer in a solvent (N-methylpyrrolidone) for sonication. Hills et al. (2019) experimentally demonstrates that RINSE reduces CNT aggregate density (i.e., the number of CNT aggregates per unit area) by >250×, without damaging CNTs or affecting CNFET performance.
- *MIXED*—Energy-efficient CMOS logic circuits using CNFETs relies on the ability to fabricate both PMOS and NMOS CNFETs that are air-stable, robust, and have tunable electrical characteristics (e.g., controlling the threshold voltage to trade-off higher CNFET circuit speed vs. lower CNFET leakage power). Existing techniques for CNT CMOS are insufficient, since they either: have large CNFET-to-CNFET variability, use materials that are not air-stable, silicon CMOS compatible, or are not robust. In order to address all of these challenges simultaneously, MIXED uses a combined doping approach that engineers both the oxide deposited over the CNTs to encapsulate the CNFET, as well as optimizing the metal source/drain contacts to CNTs by using a lower workfunction metal (e.g., Titanium) for NMOS CNFETs and higher workfunction metal (e.g., Platinum) for PMOS CNFETs. MIXED leverages only air-stable and silicon-CMOS compatible materials, and also allows for precise threshold voltage tuning by controlling the stoichiometry of robust atomic layer deposition (ALD) oxides deposited over the CNTs. MIXED also leverages workfunction engineering of the metal-CNT contacts in order to increase drive current for both PMOS CNFETs and NMOS CNFETs (Hills et al. 2019).

3.2 VLSI CNFET Nano-Design

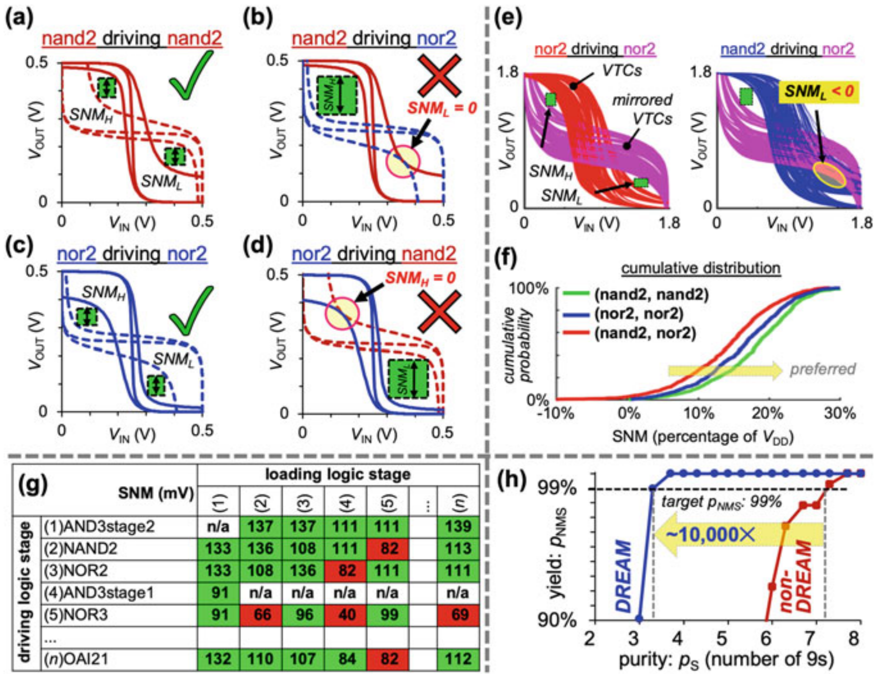
While RINSE and MIXED address CNT aggregates and CNT CMOS, metallic CNTs (m-CNTs) remain an outstanding challenge that have not been overcome by isolated advances in nano-fabrication. M-CNTs increase leakage power in VLSI CNFET circuits (degrading EDP benefits) and also degrade noise resilience of connected logic stages for digital VLSI, which can lead to incorrect logic functionality (Hills et al. 2019). To quantify the circuit-level impact of m-CNTs, we consider the noise resilience of a pair of connected logic stages (comprising a *driving* logic stage and a *loading* logic stage, e.g., two cascaded inverters to form a CMOS buffer); a useful metric for quantifying noise resilience is the *static noise margin* (SNM). SNM is defined using the **V**oltage **T**ransfer **C**haracteristics (VTCs: which defines the output voltage, V_{OUT}, as a function of the input voltage: V_{IN}, for each input of a logic stage) of the driving and loading logic stages. Using the static noise margin to quantify the noise resilience of digital VLSI circuits, one can then define the probability that all pairs of connected logic stages have SNM exceeding a target minimum required SNM,

i.e., SNM_R (SNM_R is chosen by the designer to meet circuit-level noise resilience requirements, and is typically a fraction of the circuit supply voltage: V_{DD}, e.g., $SNM_R = V_{DD}/5$). Then the **P**robability that all static **N**oise **M**argin requirements are **S**atisfied is p_{NMS}, where p_{NMS} is the probability that SNM(G_i, G_j) $\geq SNM_R$ for all logic stages in the circuit, and SNM(G_i, G_j) is the SNM for driving logic stage G_i and loading logic stage G_j (with $i \neq j$).

M-CNTs can lead to near-zero p_{NMS}, which is not acceptable for VLSI circuits. To quantify the relationship between p_{NMS} and the fraction of m-CNTs on the wafer substrate, we also define p_S as the probability that a given CNT is a semiconducting CNT (s-CNT) instead of a metallic CNT (m-CNT). Figure 5h illustrates the relationship between p_{NMS} and p_S (shown for $SNM_R = V_{DD}/5$ for a circuit consisting of approximately one million logic gates); to achieve $p_{NMS} = 99\%$, p_S must satisfy $p_S \geq 99.999,9\,9\%$, which corresponding to 1 m-CNT in 10–100 million CNTs. Despite many efforts to remove m-CNTs (Arnold et al. 2006; Patil et al. 2009; Shulaker et al. 2013a, 2015), the highest-purity results achieve $p_S \sim 99.99\%$ (1 m-CNT in 10,000 CNTs), i.e., 3–4 orders of magnitude off the target in terms of purity.

This is where the benefit of nano-design techniques comes into play. Specifically, Hills et al. (2019) describes a circuit design technique called *DREAM* ("**D**esigning **RE**silience **A**gainst **M**etallic CNTs"), which overcomes the presence of m-CNTs entirely through circuit design, and enables VLSI CNFET circuits to meet p_{NMS} requirements with $p_S = 99.99\%$ CNT purity (e.g., $p_{NMS} \geq 99\%$ for CNFET circuits with one million logic gates: Fig. 5h). Importantly, $p_S = 99.99\%$ which has already been achieved today, and can achieved through multiple techniques, e.g., solution-based CNT processing using the RINSE process. The key insight for DREAM is that m-CNTs affect the VTCs of different logic stages differently depending on how each logic stage is implemented (including both its schematic and physical layout). Thus, m-CNTs affect the SNM of different *pairs* of logic stages depending on which driving logic stage and which loading logic stages is. In particular, the SNM between a pair of connected logic stages, SNM (G_i, G_j), is more susceptible for specific *combinations* of logic stages (G_i, G_j). DREAM first quantifies SNM for all possible combinations of logic stages in a standard cell library, and then applies a logic transformation during logic synthesis (e.g., using Synopsys Design Compiler® or Cadence Genus®) to preferentially avoid the specific combinations of logic stages whose SNM is most susceptible to m-CNTs, while preferring combinations of logic stages whose SNM is more robust to m-CNTs. Importantly, the same overall circuit logic functionality is maintained, since there can be multiple configurations of logic gates that can achieve the same overall logic function in a digital circuit. Figure 5g quantifies the SNM in the presence of m-CNTs of different pairs of connected logic stages in an example standard cell library (derived from Clark et al. 2016), and an algorithm for implementing DREAM using standard electronic design automation (EDA) tools for logic synthesis is provided in Hills et al. (2019).

As an illustrative example, Fig. 5a–d illustrates the SNM for the four combinations of connected logic stage pairs using a 2-input not-and gate ("nand2") and a 2-input not-or gate ("nor2") in the presence of m-CNTs. Note that, a single m-CNT can affect multiple CNFETs simultaneously, since the length of an m-CNT can be much longer



SNM (mV)	(1)	(2)	(3)	(4)	(5)	...	(n)
(1)AND3stage2	n/a	137	137	111	111		139
(2)NAND2	133	136	108	111	82		113
(3)NOR2	133	108	136	82	111		111
(4)AND3stage1	91	n/a	n/a	n/a	n/a		n/a
(5)NOR3	91	66	96	40	99		69
...							
(n)OAI21	132	110	107	84	82		112

Fig. 5 DREAM overview. **a**–**d** Static Noise Margin (SNM) illustrated for four different pairs of connected logic stages using nand2 and nor2 logic stages. Simulation results are derived using a compact model for CNFETs (Lee et al. 2015) with parameters defined in Hills et al. (2019), in conjunction with library cells derived from Clark et al. (2016). Note that, SNM is the minimum of the "high" SNM (SNM_H) and the "low" SNM (SNM_L), i.e., $SNM = \min(SNM_H, SNM_L)$ (Hills et al. 2015). **e** SNM illustration from analyzing experimentally measured CNFETs. Current–voltage (I-V) characteristics are measured from 1,000 NMOS CNFETs and 1,000 PMOS CNFETs, and then are used to solve the VTCs of nand2 and nor2 logic stages. Despite using the exact same CNFETs, (nor2, nor2) has better (higher) SNM than (nand2, nor2). See Hills et al. (2019) for details. **f** Cumulative distribution of SNM for one million combinations of nand2 and nor2 logic stages, solved using the same method as in (**e**). **g** Minimum SNM for combinations logic stages for a projected 7 nm node CNFET technology (Hills et al. 2019). **h** p_{NMS} vs. p_S shown with and without DREAM (for SNM_R = $V_{DD}/5$, for CNFET circuits with one million logic gates), illustrating that DREAM can relax p_S requirements by 10,000× (Hills et al. 2019)

than the length of the CNFET channel, and so a single m-CNT can comprise part of the channel for multiple different CNFETs depending on their relative physical locations. Importantly, (nand2, nand2), (nor2, nor2) have better (higher) SNM compared to (nand2, nor2), (nor2, nand2) despite using the exact same VTCs. Thus, in this case, DREAM can be used to prefer (nand2, nand2), (nor2, nor2) while avoiding (nand2, nor2), (nor2, nand2), and still permitting use of *both* nand2 and nor2.

DREAM is one technique that emphasizes the essential interplay between emerging nano***technologies*** and emerging nano***design***. For example, achieving 99.999, 999% CNT purity is currently impossible using material synthesis alone,

but VLSI systems can still be demonstrated today by using DREAM to overcome inherent technology challenges.

3.3 Rapid Co-optimization of Processing and Design to Overcome Nanotechnology Variations

While the above nano-fabrication and nano-design techniques can be combined to overcome CNT aggregates, CNT CMOS, and metallic CNTs, another challenge remains: CNT variations. CNT variations can lead to near-zero functional yield, increase susceptibility to noise (quantified by p_{NMS} in the previous section), and degrade energy efficiency benefits of CNFET digital circuits (quantified by EDP). To overcome CNT variations, joint exploration and optimization of CNT processing parameters (to be improved during CNFET fabrication) and CNFET digital circuit design are required. However, existing approaches for such exploration and optimization rely on trial-and-error-based ad hoc techniques resulting in very long computation runtimes. Thus, how can a designer efficiently explore the large design space of CNT process improvements and CNFET circuit design, to overcome CNT variations in an energy efficient manner?

In this section, we present a new approach that achieves fast runtimes (e.g., 30 min for a processor core design vs. a month using existing approaches). This approach can be used to derive multiple design points (each representing a combination of parameters for CNT processing and CNFET circuit design) to overcome CNT variations. These design points preserve 90% of the projected EDP benefits of CNFET digital circuits (despite CNT variations), while simultaneously meeting circuit-level yield and noise margin constraints. The derived design points directly influence experimental research on CNFETs, and are thus essential to guide the allocation of valuable research time in developing new technologies.

An existing approach to overcome CNT variations is based on brute-force trial-and-error (Zhang et al. 2011): a designer iterates over many design points (example design points are illustrated in Fig. 3b, e.g., each one represents a combination of values for CNT processing parameters and CNFET circuit design parameters, e.g., the percentage of metallic CNTs, standard deviation of the spacing between CNTs, the target clock frequency, or values to parameterize how many CNFETs are selectively upsized), analyzing each one until a design point that satisfies a target clock frequency and target p_{NMS} with small energy cost is found (e.g., energy cost: $\Delta E < 5\%$). Furthermore, this approach utilizes highly accurate yet computationally expensive models to calculate delay penalties and *PNMV*. It suffers from two significant bottlenecks.

(1) The computational runtime required to calculate CNFET circuit delay and p_{NMS} in the presence of CNT variations limits the number of design points that can be explored.

(2) The number of required simulations can be exponential in the number of CNT processing and CNFET design parameters.

The approach described in Hills et al. (2015) overcomes these bottlenecks as follows:

(1) Degradations in CNFET circuit delay and p_{NMS} (induced by CNT variations) are computed 100× faster than the previous approach by using linearized circuit models. This speed-up enables exploration of many more design points while maintaining sufficient accuracy to make correct design decisions (details in Hills et al. (2015)).
(2) An efficient gradient descent search algorithm, which based on delay and p_{NMS} sensitivity information with respect to the processing parameters, is used to systematically guide the exploration of design points (details in Hills et al. (2015)).

Figure 6 illustrates that the combination of these techniques can exponentially reduce the required simulation time. Specifically, by leveraging *linearized* models for variations in circuit delay, energy, and noise, a designer can easily combine these models with high-level optimization techniques, such as a gradient descent search algorithm. Then by using gradient descent search, each time the designer takes a step with the gradient, they are able to *incrementally* compute the impact of variations, leveraging computation results from the previous design point, instead of starting from scratch. Thus, the *combination* of all these techniques together provides an exponential speed-up compared to brute force, e.g., reducing the required computational runtime from 1.5 months to 30 min for the "fgu" module of OpenSparc T2 (Fig. 6b).

Importantly, all of the techniques described in Sect. 3 can be integrated into standard VLSI processing and design flows, using industry-practice electronic design

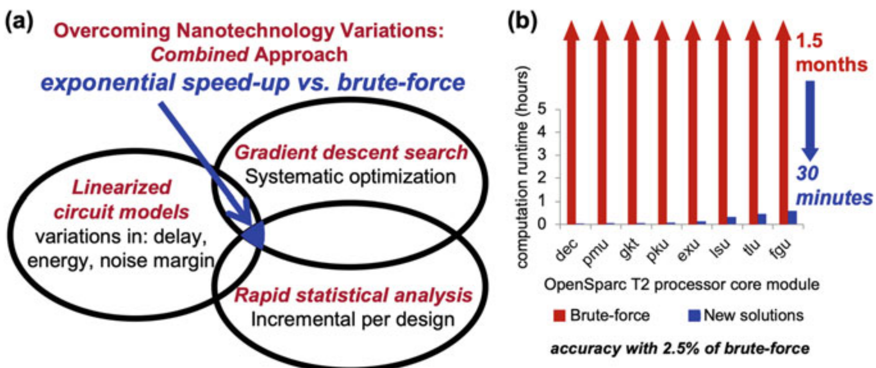


Fig. 6 Rapid co-optimization of CNT process improvements and CNFET circuit design. **a** Combined approach, leveraging linearized circuit models, gradient descent search, and rapid statistical analysis. **b** Resulting speed-up in computational runtime, shown for modules from the processor core of OpenSparc T2 (Hills et al. 2015)

automation (EDA) tools, which is a critical component of accelerating the adoption of a new technology into the mainstream. For example, Fig. 7 illustrates a reference design flow integrating RINSE, MIXED, DREAM, and the rapid co-optimization of CNT processing and CNFET circuit design described here to overcome CNT variations. The experimental CNFET circuit demonstrations described in the next section have leveraged this flow.

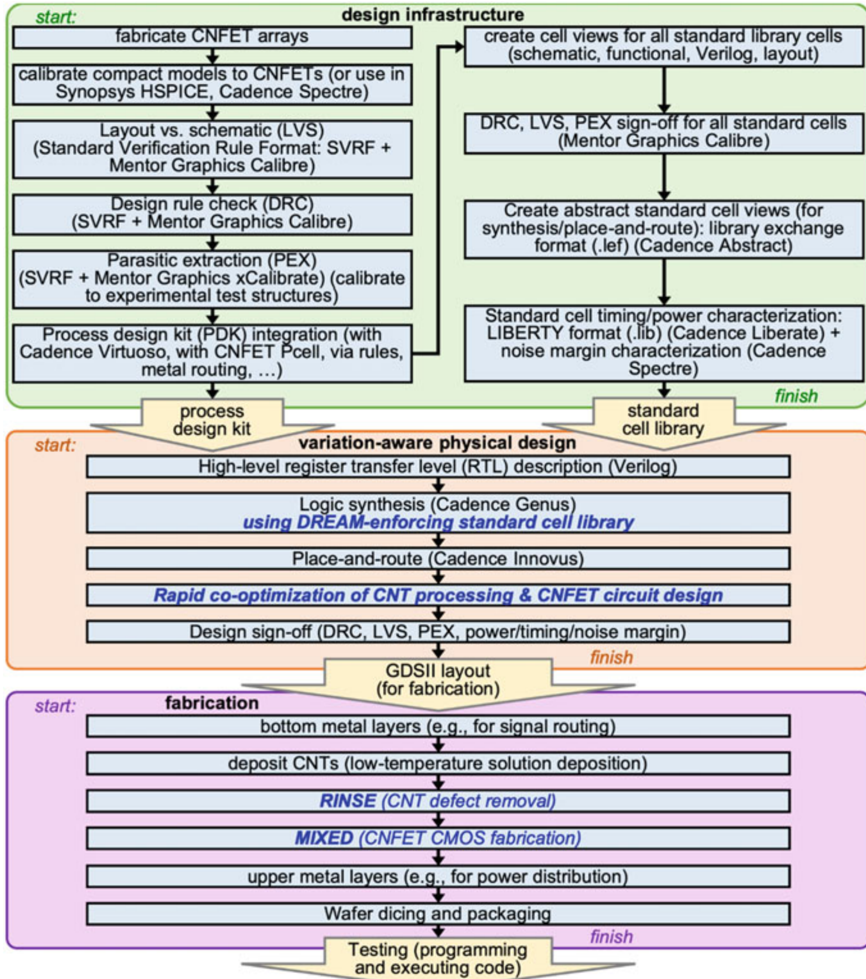


Fig. 7 VLSI CNFET Nano-Fabrication + Nano-Design flow, including RINSE, MIXED, DREAM, and the rapid co-optimization of CNT processing and CNFET circuit design to overcome CNT variations (each of these steps is highlighted in blue). Details of the "DREAM-enforcing standard cell library" can be found in Hills et al. (2019)

3.4 Experimental CNFET Circuit Demonstrations

This sub-section summarizes how the combined nano-fabrication and nano-design techniques have enabled experimental realizations of larger-scale CNFET circuits. These demonstrations include CNT CMOS analog and mixed-signal circuits (Amer et al. 2019; Ho et al. 2019), static random-access memory (SRAM) arrays using CNFETs (Kanhaiya et al. 2019a, b), and a RISC-V microprocessor built using CNFETs (Hills et al. 2019). Analog, digital, and memory circuits have become essential parts of VLSI computing systems today, and so the ability to yield these types of circuits is an important aspect of technology development for the wide range of new technologies being considered for next-generation computing systems. We refer the reader to the respective references for more details of these experimental CNFET circuit demonstrations.

- *CNT CMOS analog and mixed-signal circuits*—While CNFET digital logic can maintain correct logic functionality in the presence of m-CNTs (e.g., leveraging DREAM, although increased leakage current can degrade overall EDP and SNM metrics), m-CNTs can result in catastrophic failure mechanisms for analog CNFET circuits. For example, m-CNTs can severely attenuate amplifier gain, resulting in incorrect operation of mixed-signal circuit building blocks, including digital-to-analog converters (DACs) and analog-to-digital converters (ADCs). To overcome the challenge of m-CNTs for analog and mixed signal circuits, Amer et al. (2019) provides an overview of a combined processing and design technique called SHARC (**S**elf-**H**ealing **A**nalog with **R**RAM and **C**NFETs). SHARC leverages programmable **R**esistive **R**andom-**A**ccess **M**emory (RRAM) elements, which are configured in series with CNFETs, to automatically "self-heal" analog circuits to operate correctly despite the presence of m-CNTs. SHARC enabled the first analog and mixed-signal CNT CMOS circuits that are robust to m-CNTs, including 4-bit DACs and successive approximation register (SAR) ADCs (Amer et al. 2019; Ho et al. 2019). Additional CNFET analog circuit demonstrations are described in Ho et al. (2019) and are shown in Fig. 8.
- *CNT CMOS SRAM arrays*—Fig. 9 summarizes experimental demonstrations and measurements of 1 kilobit (32 × 32) 6-transistor (6 T) CNFET SRAM arrays, each comprising 6,144 CNFETs (both PMOS and NMOS), with all 1,024 cells functioning correctly while being connected within the same circuit (with shared wordlines and shared bitlines) (Kanhaiya et al. 2019a, b). Additional demonstrations in Kanhaiya et al. (2019a, b) include the first 10-transistor (10 T) SRAM cells, which exhibit relatively higher read- and write-margins (Calhoun and Chandrakasan 2007), and which operate at highly scaled voltages down to V_{DD} = 300 mV (Kanhaiya et al. 2019b). Because CNFETs can be fabricated at lo processing temperatures, CNFET SRAM cells can be fabricated directly on top of interconnect routing (additional details in Sect. 4). This enables new circuit-/system-level opportunities for CNFET SRAM, including: (1) fabricating SRAM directly on top of processor cores (Shulaker et al. 2014, 2017), and (2) utilizing back-end-of-line (BEOL) metal routing both above *and below* CNFETs (e.g., buried power rails

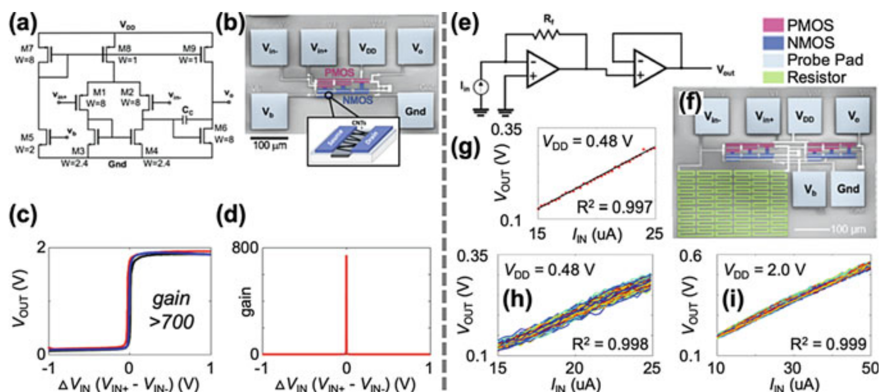


Fig. 8 CNT CMOS Analog Circuits (Ho et al. 2019), including 2-stage operational amplifier (op-amp) in (**a**)–(**d**), and implementation of CNFET op-amp in a current-sensing analog sub-system. **a** 2-stage op-amp schematic. Annotated CNFET widths are multiples of a CNFET with width $W = 5\ \mu$m and length $L = 3\ \mu$m. **b** Scanning electron microscopy (SEM) image of one fabricated 2-stage op-amp, false-colored to show the PMOS and NMOS CNFETs in the circuit (large squares are probe pads). **c** Three overlaid measured waveforms from the 2-stage op-amp, showing output voltage (V_{OUT}) as a function of differential input voltage ($\Delta V_{IN} = V_{IN+} - V_{IN-}$). **d** Corresponding gain for the same measurements in (**c**), with gain $= \Delta V_{OUT}/\Delta V_{IN}$, where $\Delta V_{OUT} = V_{OUT}(V_{IN+}) - V_{OUT}(V_{IN-})$ (additional figures of merit are provided in Ho et al. (2019)). **e–f** Schematic and SEM of the current-sensing analog sub-system with external current source. **g** Measured linear response of the sub-system, converting input current to output voltage (with supply voltage $V_{DD} = 0.48$ V). **h–i** 100 repeated measurement cycles, illustrating minimum drift over time, for $V_{DD} =$ 0.48 V (in (h)) and $V_{DD} = 2.0$ V (in (i)), demonstrating functionality and linearity over a range of supply voltages

(Chava et al. 2018)) to potentially improve SRAM cell density (Kanhaiya et al. 2019b).

- *RV16X-NANO*—Fig. 10 illustrates a recent demonstration of a microprocessor built entirely from CNFETs, which is based on the RISC-V instruction set (https://riscv.org/specifications), runs standard 32-bit instructions on 16-bit data and addresses, comprises >14,700 CMOS CNFETs (both PMOS and NMOS), and can execute compiled programs while interfacing with memory (Hills et al. 2019). Importantly, it leverages substantial existing infrastructure for both VLSI processing and design, which can more easily facilitate its adoption into high-volume commercial foundries (Sect. 3.5). As alluded to in Fig. 9, since CNFETs can be fabricated on top of back-end-of-line (BEOL) metal interconnects, RV16X-NANO also implements a new physical design architecture with BEOL metal routing both above and below the active CNFET layers. Such routing architectures can help to reduce overall routing congestion, e.g., as standard library cells continue to scale to extreme dimensions; for RV16X-NANO, metal layers above CNFETs are primarily used for power distribution, while metal layers underneath CNFETs are primarily used for signal routing, all of which has been designed

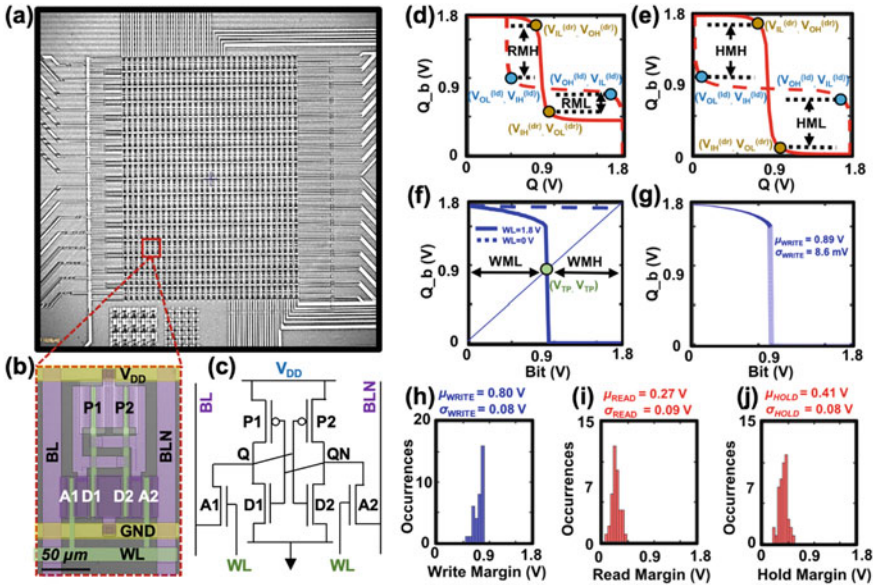


Fig. 9 CNT SRAM. **a** SEM image of 1 kilobit CNFET 6 T SRAM memory array. **b** SEM of individual SRAM cell, which is false-colored to highlight the power rails (V_{DD} and GND), pull-up CNFETs (P1 and P2), pull-down CNFETs (D1 and D2), access CNFETs (A1 and A2), wordline (WL) and bitlines (BL and BLN). Relative CNFET sizing is 2.25:1.5:1 for D1/D2:A1/A2:P1/P2. **c** Corresponding schematic for each 6 T SRAM cell. **d**–**j** 6 T SRAM cell characterization, including: **d** read margin, **e** hold margin, **f** write margin, all measured from a typical CNFET CMOS 6 T SRAM cell. **g** 1,000 overlaid measurements for a single CNFET SRAM cell. Statistical distributions from 40 CNFET SRAM cells are shown for: (h) write margin, **i** read margin, and **j** hold margin, with summary statistics μ_{WRITE}, μ_{READ}, μ_{HOLD} to denote the average values and σ_{WRITE}, σ_{READ}, and σ_{HOLD} to denote the standard deviations. Additional details are provided in Kanhaiya et al. (2019b)

using standard electronic design automation (EDA) tools for physical placement-and-routing (Cadence Innovus®). We refer the reader to Hills et al. (2019) for extensive details of the architecture, programs, standard cell libraries, and process design kits (PDKs) used to design and fabricate RV16X-NANO.

Additional experimental demonstrations, which established many of the foundations for the larger-scale demonstrations discussed here, include CEDRIC: a Turing-complete microprocessor built using CNFETs (Shulaker et al. 2013a), and Sacha: the Stanford Carbon Nanotube-based Hand-Shaking Robot, which operated based on a phase-lock loop (PLL) circuit fabricated using CNFETs (Shulaker et al. 2013b).

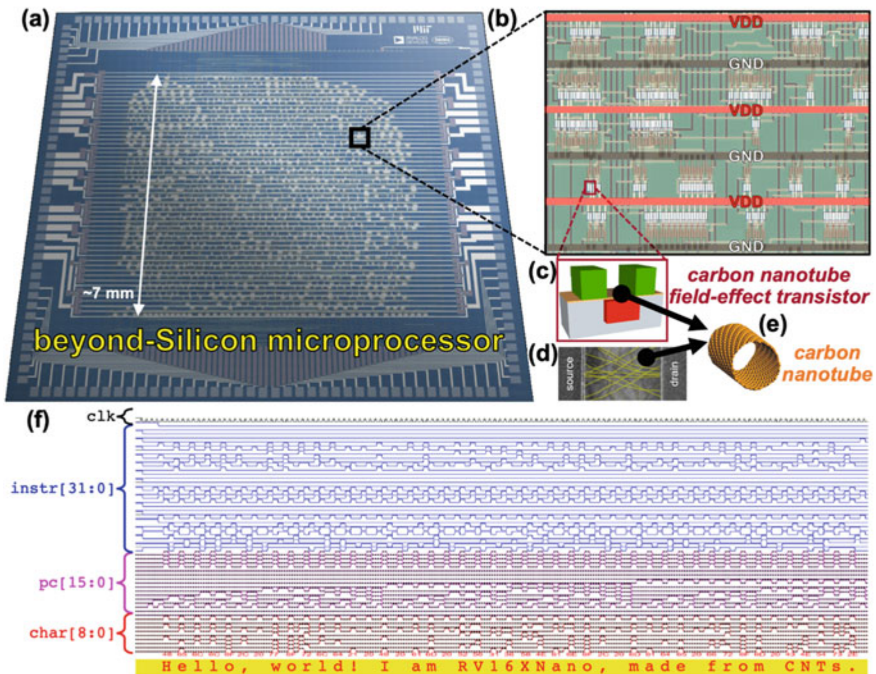


Fig. 10 RV16X-NANO. 16-bit microprocessor designed entirely using CNFETs (RV16X-NANO) (Hills et al. 2019). **a** Die photograph, including standard library cells comprising the processor core in the center (~7 mm by 7 mm), power rails extending horizontally, and probe pads around the perimeter (core inputs are primarily located toward the top, outputs are primarily located toward the bottom, and power pads are located on the left and right edges). **b** Zoomed-in photograph illustrating five rows of CNFET library cells with alternating supply voltage (VDD) and ground (GND) power rails (PMOS CNFETs are adjacent to VDD and NMOS CNFETs are adjacent to GND). **c** Schematic of a CNFET, with source/drain contacts shown in green, and gate contact in red underneath the CNFET channel (for back-gate CNFET geometries). **d** Top-view scanning electric microscopy (SEM) image of a CNFET channel with false-colored CNTs. **e** CNT rendering illustrating the location of CNTs in the CNFET channel. **f** Measured waveforms from the canonical "Hello, world" program, with input 32-bit instructions shown in blue and character output shown in red; the message translated from the ascii-valued 8-bit char[7:0] (which is valid when char[8] is high) is highlighted at the bottom

3.5 CNFET Technology Transfer to High Volume Commercial Manufacturing Facilities

Despite progress described in previous sections for developing CNFET technologies, existing demonstrations of CNFETs have been limited to academic institutions and research laboratories. While technology transfer into commercial manufacturing facilities is a necessary step for high-volume proliferation of CNFET technologies, significant obstacles must be overcome beforehand. Among others, one of these major challenges is that all of the materials and processes used to fabricate CNFETs

must meet the strict compatibility requirements of silicon-based commercial fabrication facilities. In this section, we provide an overview of recent efforts to address a specific aspect of these material- and process-based challenges: how to develop a suitable method for depositing CNTs uniformly over industry-standard large-area substrates.

To facilitate high-volume, low-cost manufacturing of CNFETs, such a deposition method for CNTs needs to be manufacturable, compatible with today's silicon-based technologies, and provide a path to achieving systems with energy efficiency benefits over silicon. Bishop et al. (2020) provides a method that meets these requirements, using a solution-based CNT deposition technique, called *incubation*, a substrate is submerged within a CNT solution, allowing CNTs to adhere to its surface (Hills et al. 2019; Zhong et al. 2017). The CNT incubation technique described in Bishop et al. (2020) offers the following key advantages that are particularly useful for the initial adoption of CNFETs in commercial manufacturing facilities:

(1) *Low barrier for integration*—uniform CNT deposition across 200 mm substrates has been experimentally demonstrated using equipment that is *already being used* for silicon CMOS fabrication within these facilities, which accelerates adoption of CNTs by leveraging existing infrastructure.
(2) *Large quantity production*—solution-based CNTs can be synthesized in large quantities for high-volume production while meeting CNT material-level requirements for realizing digital VLSI circuits, e.g., with high semiconducting CNT purity exceeding 99.99%, which meets the requirements described in Sect. 3.2) (Cao et al. 2013; Ding et al. 2015; Green and Hersam 2007; Hills et al. 2019). The creation of these highly purified semiconducting CNT solutions that meet the stringent chemical and particulate contamination requirements is also a key enabler for including CNFETs in commercial facilities (Baltzinger and Delahaye 1999).
(3) *Improved throughput and a path for energy efficiency*—while various incubation techniques have been demonstrated to be practical and effective (Hills et al. 2019; Kanhaiya et al. 2019b; Srimani et al. 2019), Bishop et al. (2020) offers characterization of the fundamental aspects of CNT incubation, with respect to manufacturability, compatibility and the resulting CNFET performance that can be achieved. This insight has resulted in both increased throughput (accelerating the time required to perform CNT incubation from 48 h to 150 s), and also VLSI circuit-level power/performance-based analysis demonstrating that incubation enables a path for CNFET circuits to compete with and eventually surpass the energy efficiency of silicon-based circuits at comparable technology nodes.

The advances in CNT incubation described in Bishop et al. (2020), together with the co-optimized nano-fabrication and nano-design techniques described previously in this section, have enabled CNFET fabrication within two distinct industry manufacturing facilities: a commercial silicon manufacturing facility (Analog Devices, Inc.) and a high-volume manufacturing semiconductor foundry (SkyWater Technology Foundry). At each of these facilities, CNFETs are fabricated using the same

equipment currently being used to fabricate silicon product wafers, explicitly demonstrating that CNFET fabrication can leverage existing infrastructure and is silicon-CMOS-compatible. Figures 11 and 12 illustrate some of the first experimental data from these commercial facilities fabricating CNFETs and CNFET-based circuits, including uniform and reproducible CNFET fabrication across industry-standard 200 mm wafers, with 14,400/14,400 CNFETs distributed across multiple wafers and across 200 mm substrates (Fig. 11) (Bishop et al. 2020), and electrical measurements of the first CNFET-based standard library cells fabricated at SkyWater at a 130 nm technology node (Fig. 12) (Srimani et al. 2020).

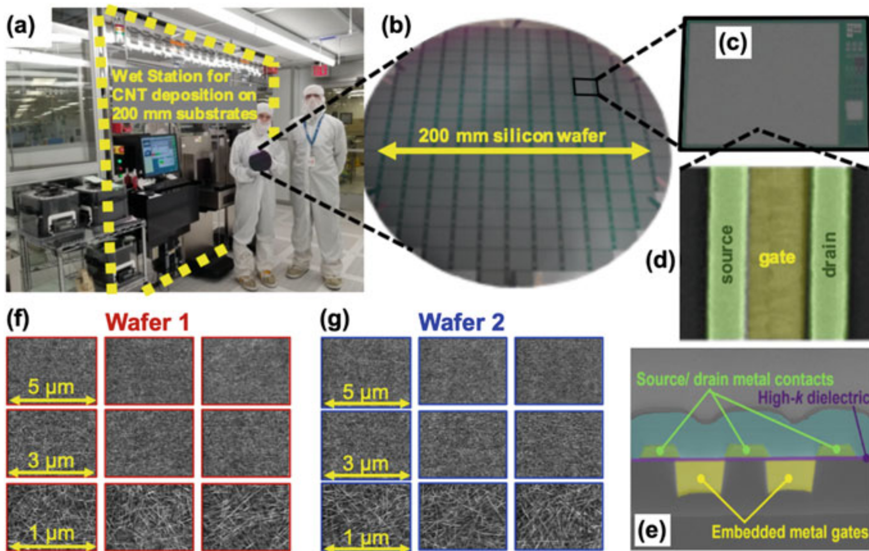


Fig. 11 Wafer-scale integration of CNFETs across 200 mm wafers within a commercial silicon foundry. **a** Processing station within the foundry for performing CNT incubation (details in Bishop et al. (2020)). **b**–**d** Images from the foundry, including: **b** 200 mm wafer with CNFETs, **c** individual die, and **d** top-view SEM of a single CNFET (note that, the CNTs are not visible in this image due to the CNFET gate embedded underneath the channel region. **e** Cross-sectional SEM of two CNFETs connect in series (sharing a source/drain contact), with false-colored source/drain metal contacts, high-*k* gate dielectric, and embedded metal gates (leveraging a back-gate CNFET geometry). The sum of the channel length (~285 nm) and the contact length (~265 nm) sets the contacted gate pitch (CGP) of ~550 nm, suitable for a ~130 nm technology node. **f**–**g** SEM images from multiple points across 200 mm wafers, which illustrate CNT deposition (after CNT incubation), and which are used for characterizing CNT density, uniformity, and reproducibility (see Bishop et al. 2020)

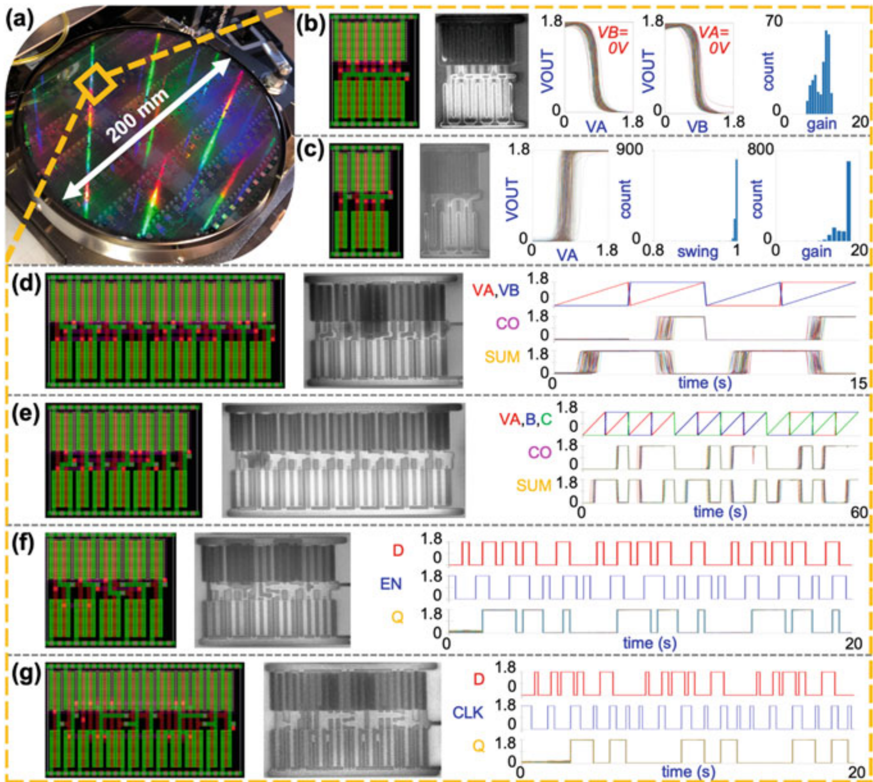


Fig. 12 Experimental measurements of CNFET standard library cells fabricated at SkyWater Technology Foundry (schematics for all library cells follow design guidelines for static CMOS logic families, i.e., with PMOS CNFETs comprising the pull-up network and NMOS CNFETs comprising the pull-down network, see (Hills et al. 2019) for schematics). **a** 200 mm wafer with CNFET circuits, including multiple standard library cells shown in (**b**)–(**g**). Each unique entry in (**b**)–(**g**) corresponds to a unique standard library cell, with the physical layout shown on the left (image from Cadence Virtuoso®), an SEM image shown in the center, and experimentally measured waveforms from multiple instantiations of that cell shown on the right (with supply voltage $V_{DD} = 1.8$ V). Waveforms include overlaid measurements from at least 100 instantiations of each logic gate. **b** 2-input not-or "NOR2" logic gate, with logical function "OUT = !(A + B)". The relationships V_{OUT} vs. V_A and V_{OUT} vs. V_B are the voltage transfer curves for multiple instantiations of NOR2 logic gates. Gain is the maximum is the maximum value of $\Delta V_{OUT}/\Delta V_{IN}$ for V_{IN} in the range of [0, V_{DD}] (where V_{IN} corresponds to either of the A or B inputs being swept, i.e., either V_A or V_B). **c** 2-stage buffer "BUF" logic gate. Swing is difference between the maximum and minimum value of V_{OUT} (as a fraction of V_{DD}) as V_A is swept over the range [0, V_{DD}]; thus, for static CMOS logic, swing should approach 1.0 (i.e., V_{OUT} is "rail-to-rail"). **d** Half-adder logic gates, with "SUM = XOR(A, B)" and "CO = A*B". Measured waveforms show input voltages (V_A and V_B) and output voltages (SUM and CO) as functions of time. **e** Full-adder logic gate, with "SUM = XOR(A, B, C)" and "CO = A*B + B*C + A*C". Sequential logic elements include D-Latches (shown in (**f**), with data input "D" and enable input "EN") and D-Flip-Flops (shown in (**g**), with data input "D" and clock input "CLK"), both of which have output "Q" to indicate the state. Additional library cell functions realized (not shown here) include: D-flip-flops with asynchronous reset, D-flip-flops with scan, clock-gating cells, multiplexors, exclusive-or, exclusive-nor, fill cells (to connect power rails during place-and-route), and "decap" cells (to increase capacitance between power supply rails)

4 Next-Generation Nano-Systems

While the sections so far have highlighted CNFETs as an example of an end-to-end approach for developing one specific nanotechnology, finding the “best” transistor or memory technologies *alone* is insufficient to satisfy future application demands. Instead, heterogeneous integration of multiple technologies simultaneously, which can be combined to create entirely new computing systems, can result in far-larger benefits overall. This is because systems today (including general purpose processors and domain-specific accelerators) are often limited by *system-level* inefficiencies; for example, the “memory wall,” refers to the vast majority of execution time and energy that wasted passing data back and forth between processing elements and off-chip memory (e.g., off-chip DRAM).

To overcome these outstanding challenges, the device-level benefits of new nanotechnologies must be combined with the novel systems architectures that they naturally enable. For example, many nanotechnologies can be fabricated at low processing temperatures (<400 °C, compared to >1,000 °C for silicon CMOS), which is a key property that enables the development of monolithic 3D nanosystems. For example, it is projected that monolithic 3D systems, with multiple layers of computation and multiple layers of memory densely integrated directly on top of each other, can improve energy efficiency by over two orders of magnitude compared to systems today (quantified by Energy-Delay Product (EDP)) (Aly et al. 2018). Alternative approaches to 3D integration include “2.5-dimensional” integration (integrating multiple chips on interposers) or 3D chip stacking, but the relatively large pitch of vertical interconnects (such as Through-Silicon Vias: TSVs) limits the density of vertical interconnects. Monolithic 3D systems, on the other hand, leverage standard metal routing vias from the back-end-of-line (BEOL), which can be much denser (e.g., over 2 orders of magnitude denser than TSVs (Aly et al. 2015)), which can translate to massive increase in system-level performance metrics such as processor-to-memory bandwidth (Aly et al. 2015, 2018).

In this section, we provide a summary of the progress and prospect of developing such monolithic 3D “nanosystems”. Figure 13 illustrates that nanosystems are naturally enabled by low-temperature fabrication of emerging nanotechnologies for both logic and memory. Using these technologies, nanosystems offer radically new opportunities to improve energy efficiency, e.g., with separate circuit tiers optimized for processor cores, caches, power delivery, heat removal, etc.), and an example 3D nanosystem is shown in Fig. 14. Section 4.1 presents experimental demonstrations of 3D nanosystem prototypes that have been developed in academic institutions; Sect. 4.2 presents progress toward the development of 3D nanosystems at commercial manufacturing facilities, including advances in both processing and design infrastructures.

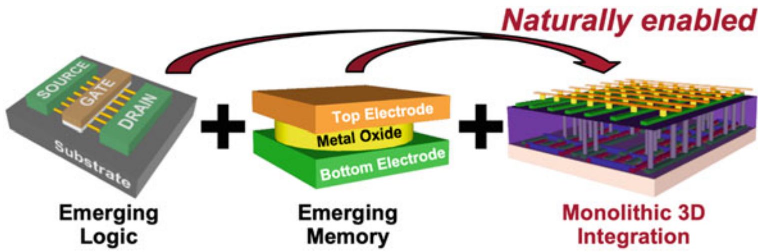


Fig. 13 Monolithic 3D integration is naturally enabled by emerging nanotechnologies that can be fabricated at low processing temperatures. For example, for logic, one can use CNFETs (using all the techniques described in the previous sections) or various 2D materials (black phosphorus, MoS_2, WSe_2), and for memory, there is a wide range of technologies to choose from, including RRAM, spin-transfer torque magnetic RAM (STT-MRAM), conductive bridge RAM (CBRAM), and more, and a designer can choose the technology with characteristics best-suited for a particular application

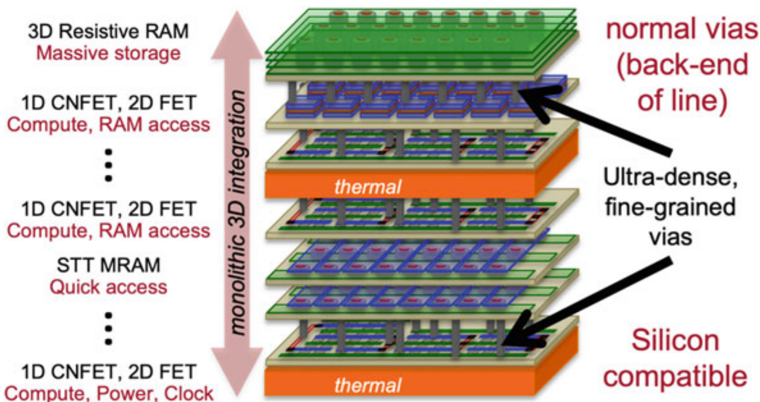


Fig. 14 Example 3D nanosystem, enabled by the low-temperature fabrication of emerging nanotechnologies. Such nanosystems combine advances from across the computing stack, including nanomaterials such as CNTs for high-performance and energy-efficient transistors, high-density on-chip non-volatile memories, fine-grained 3D integration of logic and memory with ultra-dense connectivity, new 3D architectures for computation immersed in memory, and integration of new materials technologies for efficient heat removal solutions. Resulting nanosystems offer radically new opportunities for computing architectures, e.g., with separate circuit tiers optimized for processor cores, caches, power delivery, heat removal, etc., as shown here

4.1 Experimental 3D Nano-System Demonstrations

Just as an end-to-end approach for evaluating the potential benefits of CNFETs for 2-D circuits, quantifying the benefits of 3D nanosystems is a necessary step before investing in resources for their fabrication and experimental development. For extensive analysis on various system configurations and potential paths for continuing

to improve 3D nanosystems, we refer to the reader to Aly et al. (2018), which serves to motivate the experimental nanosystem demonstrations described in this section. To demonstrate that experimental nanosystems are now becoming a reality, we summarize two representative system-level demonstrations, which not only show that monolithic 3D integration of multiple nanotechnologies is achievable in practice, but also demonstrate some of the application domains enabled by monolithic 3D integrated systems.

- *3D nanosystem integrating layers of computation, memory, and sensing*—Fig. 15a illustrates a prototype of 3D nanosystem that comprises over two million CNFETs, one megabit of RRAM, all of which are fabricated sequentially over a bottom tier of silicon FETs (Shulaker et al. 2017). In particular, the CNFETs occupy two unique vertical circuit tiers: the top tier, in which the CNFETs are exposed to the environment and function as gas sensors and write their captured data directly into the tier of RRAM memory underneath (the "1-transistor 1-resistor" or "1T-1R" memory cells use the bottom layer of silicon FETs for the access transistor). Another tier of CNFET circuits is then used to implement a classification accelerator that extracts features from the data stored in the RRAM memory. Since each CNFET sensor writes directly into its own dedicated memory cell, without

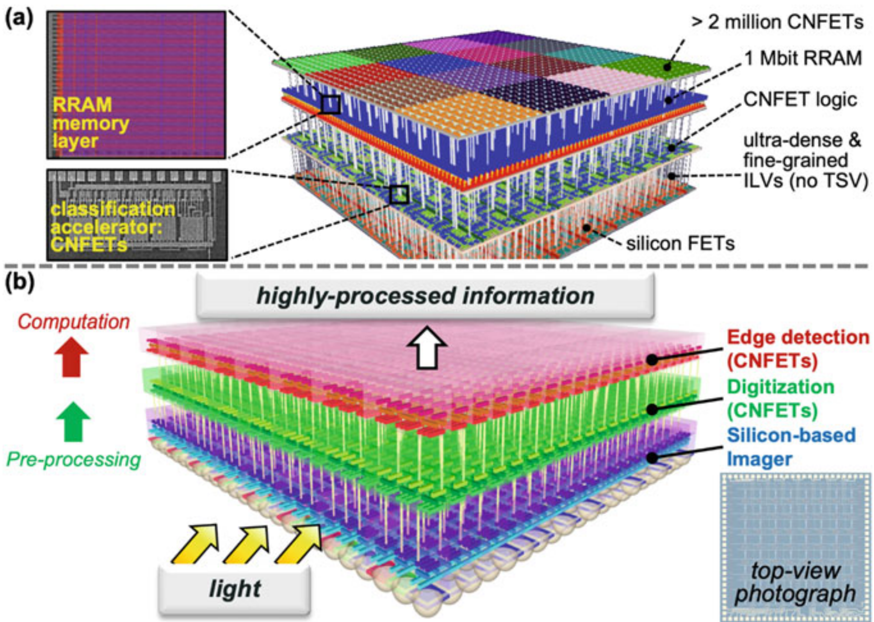


Fig. 15 Experimental demonstrations of 3D nanosystems. **a** 4-tier nanosystem comprising two tiers of CNFETs, one tier of RRAM, and one tier of silicon FETs (example applications include high-throughput characterization of ambient gases) (Shulaker et al. 2017). **b** Monolithic 3D imaging system, with CNFET-based edge detection circuitry fabricated directly on top of silicon-based imaging pixels (Srimani et al. 2019)

the need to be serialized through a memory port interface, this 3D nanosystem can capture massive amounts of data every second and process it on-chip, so that the overall chip output is highly-processed information instead of raw CNFET sensor data. As a demonstration, Shulaker et al. (2017) shows how this system is used to classify ambient gates. Furthermore, the fact that the layers are fabricated on top of silicon circuits experimentally demonstrates that 3D nanosystems are silicon-CMOS compatible, i.e., emerging nanotechnologies can be fabricated on top of existing silicon-based technologies.

- *Monolithic 3D imaging system*—Fig. 15b illustrates an experimentally fabricated and tested 3D nanosystem comprising three vertical circuit tiers: silicon-based imaging pixels on the bottom tier, followed by CNFET circuits on the tier above (tier 2) to perform pre-processing on the image data, and then CNFET circuits on the third tier for executing algorithms. Srimani et al. (2019) offers an demonstration of how this system is used to perform in-situ edge detection. Levering the ultra-dense vertical connectivity enabled by monolithic 3D integration, every pixel in parallel sends data vertically through the chip to the upper layers for subsequent processing, instead of having to read out data from each pixel serially, store the raw pixel values in memory, and then compute on the data in memory (e.g., in a conventional 2D system). Thus, the output of this 3D camera system is able to output highly-processed information instead of the raw pixel data. This system-level approach can enable high-throughput and low-latency image classification systems that would otherwise be impossible to build using today's silicon-based technologies.

Additional 3D nanosystem demonstrations, not described here but that we refer the interested reader to, include Wu et al. (2018, 2019).

4.2 Three-Dimensional Nano-Systems in Commercial Foundries

With the ongoing adoption of emerging nanotechnologies in commercial foundries, e.g., as described in Sect. 3.5, the subsequent development of 3D nanosystems is a natural progression to fully capitalize on the benefits that new technologies have to offer. For 3D nanosystems leveraging CNFETs, each individual CNFET circuit tier follows similar processing as described for 2D CNFET systems, and so all of the techniques for overcoming inherent CNT imperfections and variations (described above) can be used lock, stock, and barrel for the development of 3D systems. This approach has been taken by SkyWater Technology Foundry, who has demonstrated in Srimani et al. (2020) that they are developing processes for CNFETs and RRAM that can be integrated directly into the back-end-of-line (BEOL), enabling the recent demonstration of a monolithic 3D systems being developed at SkyWater with multiple layers of CNFETs and multiple layers of RRAM at a 130 nm technology node (https://spectrum.ieee.org/nanoclast/semiconductors/

devices/first-3d-nanotube-and-rram-ics-come-out-of-foundry). In this section, we provide an overview of this technology that is currently being developed, including infrastructure for both VLSI processing and design.

Figure 16 illustrates the initial foundry process, which is implemented across industry-standard 200 mm substrates. The full monolithic 3D stack, which integrates four tiers of active devices distributed throughout the BEOL metal layers, offers 15 metal layers on 13 different physical layers, using 42 mask layers. These active device tiers include two tiers of CMOS CNFETs and two tiers of RRAM, all of which are fabricated using low-temperature and BEOL-compatible process flows. All vertical layers are fabricated sequentially over the same starting substrate, using the same BEOL inter-layer vias that are used to connect standard metal layers (such as the vias connecting “metal 1” and “metal 2”). Due to *monolithic* 3D integration, the vertical connectivity between tiers can exceeds 11 million vertical interconnects per mm^2 (with via pitch of ~300 nm at the ~130 nm technology node). All fabrication is wafer-scale without any per-unit customization, leveraging existing silicon CMOS high-volume manufacturing processing and infrastructure. As an example of circuits spanning multiple tiers, electrical current–voltage characteristics for 1T-1R memory cells are shown in Fig. 16g, for all four combinations of: either NMOS or PMOS CNFET on tier 3 (for the “1T” element), and RRAM on either tier 1 or tier 2 (for the “1R” element). Electrical characteristics for CNFETs are similar to those shown in Fig. 12, since the same process is used for each CNFET tier.

In addition to the monolithic 3D process infrastructure, this process is accompanied by a complete design infrastructure, so that a designer would have everything they need to tape-out a monolithic 3D system using this process. An essential component of this 3D design infrastructure has been the development of a *monolithic* 3D process design kit (PDK), which provides 3D support for: Design Rule Check (DRC), LVS, Parasitic Extraction (PEX), circuit simulation, electromigration/voltage drop analysis (EM/IR), logic synthesis, place-and-route, metal fill, and optical proximity correction (OPC) for final photomask generation. Alternatively, many of today’s existing efforts to design 3D systems rely on (manually) stitching together separate circuit tiers each designed using conventional PDKs; however, this approach can neglect critical effects such as inter-tier parasitics (affecting timing closure), and also can prevent teams from verifying that their designs are correct (lacking tools such as Layout Vs. Schematic (LVS) for full 3D systems). Thus, while these alternative approaches may suffice for academic exercises, they can be insufficient for analyzing, verifying, and taping out 3D systems.

Figure 17 summarizes the industry-practice VLSI design flow described in Srimani et al. (2020), which corresponds to the process in Fig. 16. In addition to the 3D design tools described above, this design flow also incorporates compact models for CNFETs and RRAM on all circuit tiers that are compatible with standard circuit simulations (e.g., Synopsys HSPICE® and Cadence Spectre®), as well as standard cell libraries with 906 total standard cells, including high-density, high-speed, and low-leakage standard cell variants. Importantly, the design flow leverages existing commercial tools and performs all steps required to transform high-level hardware descriptions into standard layout formats for generating final reticles for fabrication.

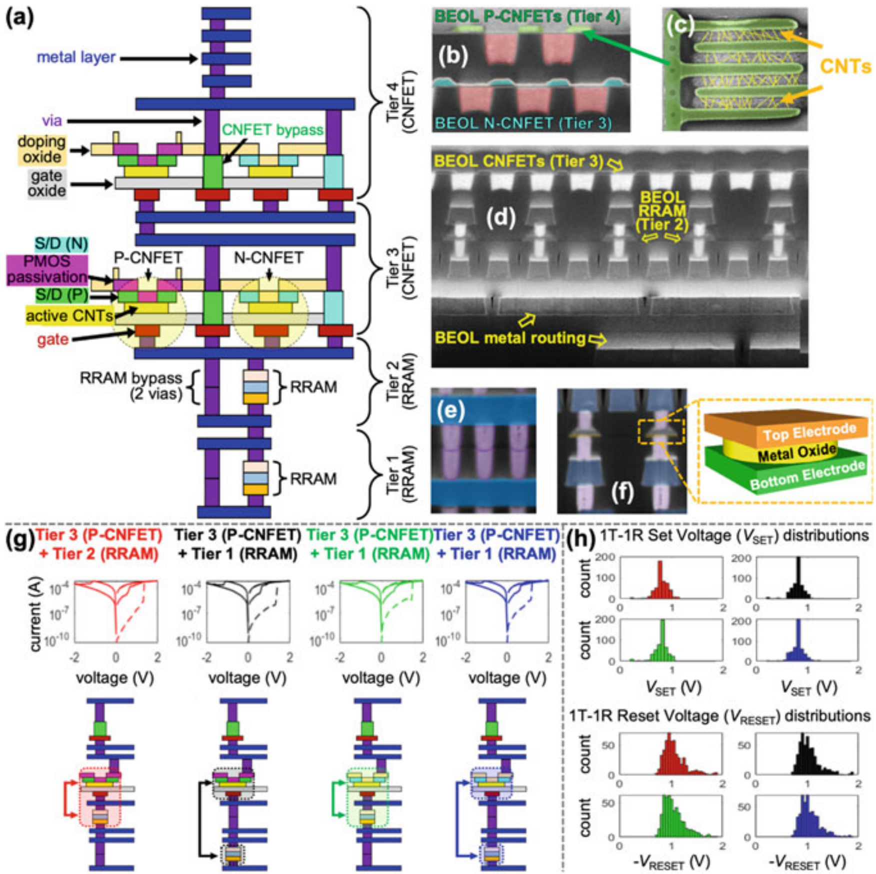


Fig. 16 Multi-Tier CNFET and RRAM process in a commercial foundry. **a** Schematic illustration of the process cross-section established within the foundry. This initial process includes 4 device tiers: RRAM memory for tier 1 and tier 2, and CNFET CMOS for tier 3 and tier 4, all of which are fabricated in the BEOL. There are 15 metal layers total layers on implemented on 13 physical layers (since source/drain deposition for NMOS CNFETs and PMOS CNFETs use separate metal depositions but occupy the same physical layer). **b** Cross section SEM images of NMOS CNFETs (top) and PMOS CNFETs (bottom), highlighting the MIXED CMOS process (Sect. 3.1). **c** Top-view SEM of a CNFET with multiple fingers, with false coloring to indicate the CNTs in the channel. **d** Cross-section SEM image showing CNFETs fabricated directly over RRAM memory cells, with routing above and below. Here, bottom metal layers show dummy metal fill (automatically performed using standard electronic design automation tools). **e** RRAM bypass vias through an RRAM device layer, illustrating the option of using RRAM tier 1 for additional routing resources. **f** Zoomed-in view of tight-pitched RRAM with corresponding schematic. Colors for (**b**–**f**) correspond to coloring in (**a**). **g** Typical I-V characteristics of 1T-1R memory cells for different combinations of NMOS/PMOS CNFETs and RRAM tiers, showing the form ("F"), set ("S"), and reset ("R") events of the RRAM cell through the CNFET select transistor. (c) Measured distributions of Set Voltage (V_{SET}) and Reset Voltage (V_{RESET}) for 512-bit RRAM arrays fabricated across different BEOL layers in the monolithic 3D integrated circuit

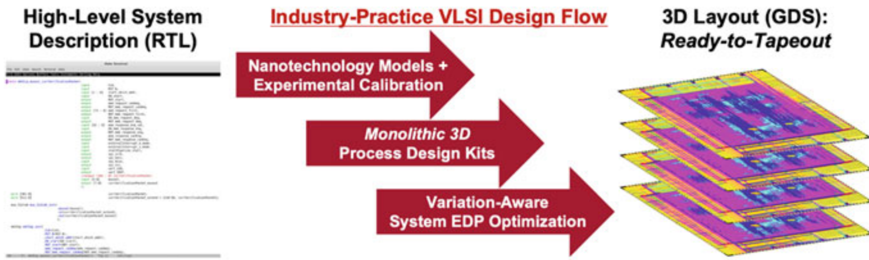


Fig. 17 Design flow for creating monolithic 3D nanosystems, using the process in Fig. 16. This flow leverages a monolithic 3D PDK, standard cell libraries, and standard EDA tools, so that designers can transform a high-level description of a system (e.g., a register transfer level (RTL) description in Verilog) into a 3D layout (e.g., in standard graphic database system (GDS) format) for taping out monolithic 3D nanosystems (Srimani et al. 2020)

5 Outlook

We hope that this chapter has given the reader a taste of the end-to-end approach required for developing new technologies to address growing system-level bottlenecks in today's computing systems. While we have focused on CNFETs and the monolithic 3D nanosystems that they enable, many of the principles described here can and should be applied to any emerging technology, including the wide range of new materials, devices, systems, architectures, and integration that are currently being investigated today (including those described in the introduction), and which are at varying levels of maturity. Of course, just as 3D nanosystems are not constrained by the properties of today's silicon technologies, futuristic systems may evolve to become increasingly dissimilar to systems today, both from a physical perspective, and from an architectural perspective. Solutions may require diverse design abstractions, design methodologies adapted for different fields, or statistical methods to model complex system interactions with dynamic environments. No matter how system development continues to progress, we are confident that it will require tight-knit coordination among interdisciplinary researchers in both academia and industry, and we hope that this chapter may spark the reader to start thinking about new revolutions in the development of next-generation electronic systems.

References

M.M.S. Aly et al., Energy-efficient abundant-data computing: the N3XT 1,000 x. IEEE Comput. **48**(12), 24–33 (2015)

M.M.S. Aly, T.F. Wu, A. Bartolo, Y.H. Malviya, W. Hwang, G. Hills, I. Markov et al., The N3XT approach to energy-efficient abundant-data computing. Proc. IEEE **107**(1), 19–48 (2018)

A.G. Amer, R. Ho, G. Hills, A.P. Chandrakasan, M.M. Shulaker, 29.8 SHARC: self-healing analog with RRAM and CNFETs, in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, 2019), pp. 470–472

M.S. Arnold, A.A. Green, J.F. Hulvat, S.I. Stupp, M.C. Hersam, Sorting carbon nanotubes by electronic structure using density differentiation. Nat. Nanotechnol. **1**(1), 60–65 (2006)

J.-L. Baltzinger, B. Delahaye, *Semiconductor Technologies*, ed. by J. Grym (IntechOpen, 1999), Chap. 4, pp. 57–78

M.D. Bishop, G. Hills, T. Srimani, C. Lau, D. Murphy, S. Fuller, J. Humes, A. Ratkovich, M. Nelson, M.M. Shulaker, Fabrication of carbon nanotube field-effect transistors in commercial silicon manufacturing facilities. Nat. Electron. 1–10 (2020)

B.H. Calhoun, A.P. Chandrakasan, A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation. IEEE J. Solid-State Circuits **42**(3), 680–688 (2007)

Q. Cao et al., Arrays of single-walled carbon nanotubes with full surface coverage for high-performance electronics. Nat. Nanotechnol. **8**, 180–186 (2013)

B. Chava, J. Ryckaert, L. Mattii, S.M.Y. Sherazi, P. Debacker, A. Spessot, D. Verkest, DTCO exploration for efficient standard cell power rails, in *Design-Process-Technology Co-optimization for Manufacturability XII*. International Society for Optics and Photonics, vol. 10588 (2018), p. 105880B

L.T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy, G. Yeric, ASAP7: a 7-nm finFET predictive process design kit. Microelectron. J. **53**, 105–115 (2016)

J. Ding et al., A hybrid enrichment process combining conjugated polymer extraction and silica gel adsorption for high purity semiconducting single-walled carbon nanotubes. Nanoscale **7**, 15741–15747 (2015)

D.J. Frank, Y. Taur, H.-S. Philip Wong, Generalized scale length for two-dimensional effects in MOSFETs. IEEE Electron Device Lett. **19**(10), 385–387 (1998)

L. Gomez, I. Aberg, J.L. Hoyt, Electron transport in strained-silicon directly on insulator ultrathin-body n-MOSFETs with body thickness ranging from 2 to 25 nm. IEEE Electron Device Lett. **28**(4), 285–287 (2007)

A.A. Green, M.C. Hersam, Ultracentrifugation of single-walled nanotubes. Mater. Today **10**, 59–60 (2007)

P. Hashemi et al., Strained Si_{1-x} Ge_x-on-insulator PMOS FinFETs with excellent sub-threshold leakage, extremely-high short-channel performance and source injection velocity for 10 nm node and beyond, in *Proceedings of the Symposium on VLSI Technology (VLSI-Technology), Digest Technical Papers* (2014a), pp. 1–2

P. Hashemi et al., First demonstration of high-Ge-content strained- Si_{1-x} $Ge_x(x = 0.5)$ on insulator PMOS FinFETs with high hole mobility and aggressively scaled fin dimensions and gate lengths for high- performance applications, in *Proceedings of the IEEE International Electron Devices Meeting* (2014b), pp. 16.1.1–16.1.4

G. Hills, J. Zhang, M.M. Shulaker, H. Wei, C.-S. Lee, A. Balasingam, H.-S. Philip Wong, S. Mitra, Rapid co-optimization of processing and circuit design to overcome carbon nanotube variations. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **34**(7), 1082–1095 (2015)

G. Hills, M.G. Bardon, G. Doornbos, D. Yakimets, P. Schuddinck, R. Baert, D. Jang et al. Understanding energy efficiency benefits of carbon nanotube field-effect transistors for digital VLSI. IEEE Trans. Nanotechnol. **17**(6), 1259–1269 (2018)

G. Hills, C. Lau, A. Wright, S. Fuller, M.D. Bishop, T. Srimani, P. Kanhaiya et al., Modern microprocessor built from complementary carbon nanotube transistors. Nature **572**(7771), 595–602 (2019)

R. Ho, C. Lau, G. Hills, M.M. Shulaker, Carbon nanotube CMOS analog circuitry. IEEE Trans. Nanotechnol. **18**, 845–848 (2019)

P.S. Kanhaiya, C. Lau, G. Hills, M. Bishop, M.M. Shulaker, 1 Kbit 6T SRAM arrays in carbon nanotube FET CMOS, in *2019 Symposium on VLSI Technology* (IEEE, 2019a), pp. T54–T55

P.S. Kanhaiya, C. Lau, G. Hills, M.D. Bishop, M.M. Shulaker, Carbon nanotube-based CMOS SRAM: 1 kbit 6T SRAM arrays and 10T SRAM cells. IEEE Trans. Electron Devices **66**(12), 5375–5380 (2019b)

K.J. Kuhn, Considerations for ultimate CMOS scaling. IEEE Trans. Electron Devices **59**(7), 1813–1828 (2012)

C. Lau, T. Srimani, M.D. Bishop, G. Hills, M.M. Shulaker, Tunable n-type doping of carbon nanotubes through engineered atomic layer deposition HfOX films. ACS Nano **12**(11), 10924–10931 (2018)
C.-S. Lee, E. Pop, A.D. Franklin, W. Haensch, H.-S. Philip Wong, A compact virtual-source model for carbon nanotube FETs in the sub-10-nm regime—Part I: Intrinsic elements. IEEE Trans. Electron Devices **62**(9), 3061–3069 (2015)
N. Loubet et al., Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET, in *Proceedings of the Symposium on VLSI Technology* (2017), pp. T230–T231
H. Mertens et al., Gate-all-around MOSFETs based on vertically stacked horizontal Si nanowires in a replacement metal gate process on bulk Si substrates, in *Proceedings of the IEEE Symposium on VLSI Technology* (2016), pp. 1–2
OpenSPARC (Dec. 2011), http://www.opensparc.net/opensparc-t2
N. Patil, A. Lin, J. Zhang, H. Wei, K. Anderson, H.-S. Philip Wong, S. Mitra, VMR: VLSI-compatible metallic carbon nanotube removal for imperfection-immune cascaded multi-stage digital logic circuits using carbon nanotube FETs, in *2009 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2009), pp. 1–4
C. Qiu, Z. Zhang, M. Xiao, Y. Yang, D. Zhong, L.-M. Peng, Scaling carbon nanotube complementary transistors to 5-nm gate lengths. Science **355**(6322), 271–276 (2017)
Y. Sasaki et al., Novel junction design for NMOS Si Bulk-FinFETs with extension doping by PEALD phosphorus doped silicate glass, in *Proceedings of the IEEE International Electron Devices Meeting* (2015), pp. 21–28
M.M. Shulaker, G. Hills, N. Patil, H. Wei, H.-Y. Chen, H.-S. Philip Wong, S. Mitra. Carbon nanotube computer. Nature **501**(7468), 526–530 (2013a)
M.M. Shulaker, J.V. Rethy, G. Hills, H. Wei, H.-Y. Chen, G. Gielen, H.-S. Philip Wong, S. Mitra, Sensor-to-digital interface built entirely with carbon nanotube FETs. IEEE J. Solid-State Circuits **49**(1), 190–201 (2013b)
M.M. Shulaker, K. Saraswat, H.-S. Philip Wong, S. Mitra. Monolithic three-dimensional integration of carbon nanotube FETs with silicon CMOS, in *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers* (IEEE, 2014), pp. 1–2
M.M. Shulaker, G. Hills, T.F. Wu, Z. Bao, H.-S. Philip Wong, S. Mitra. Efficient metallic carbon nanotube removal for highly-scaled technologies, in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), pp. 32–34
M.M. Shulaker, G. Hills, R.S. Park, R.T. Howe, K. Saraswat, H.-S. Philip Wong, S. Mitra. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. Nature **547**(7661), 74–78 (2017)
T. Srimani, G. Hills, C. Lau, M. Shulaker, Monolithic three-dimensional imaging system: carbon nanotube computing circuitry integrated directly over silicon imager, in *2019 Symposium on VLSI Technology* (IEEE, 2019), pp. T24–T25
T. Srimani et al., Heterogeneous integration of BEOL logic and memory in a commercial foundry: multi-tier complementary carbon nanotube logic and resistive RAM at a 130 nm node, in *VLSI* (2020)
S.D. Suk et al., Investigation of nanowire size dependency on TSNWFET, in *Proceedings of the IEEE International Electron Devices Meeting* (2007), pp. 891–894
K. Uchida, J. Koga, S.-I. Takagi, Experimental study on carrier transport mechanisms in double- and single-gate ultrathin-body MOSFETs-Coulomb scattering, volume inversion, and/spl T_{SOI}-induced scattering, in *Proceedings of the IEEE International Electron Devices Meeting Technical Digest* (2003), pp. 33–35
T.F. Wu, H. Li, P.-C. Huang, A. Rahimi, G. Hills, B. Hodson, W. Hwang et al., Hyperdimensional computing exploiting carbon nanotube FETs, resistive RAM, and their monolithic 3D integration. IEEE J. Solid-State Circuits **53**(11), 3183–3196 (2018)
T.F. Wu, B.Q. Le, R. Radway, A. Bartolo, W. Hwang, S. Jeong, H. Li et al., 14.3 A 43 pJ/cycle non-volatile microcontroller with 4.7 μs shutdown/wake-up integrating 2.3-bit/cell resistive RAM

and resilience techniques, in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)* (IEEE, 2019), pp. 226–228

J. Zhang, N.P. Patil, S. Mitra, Probabilistic analysis and design of metallic-carbon-nanotube-tolerant digital logic circuits. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **28**(9), 1307–1320 (2009)

J. Zhang, N. Patil, H.-S. Philip Wong, S. Mitra, Overcoming car- bon nanotube variations through co-optimized technology and circuit design, in *Proceedings of the InternationalElectron Devices Meeting (IEDM)*, Washington, DC, USA, 2011, pp. 4.6.1–4.6.4

D. Zhong, M. Xiao, Z. Zhang, L.M. Peng, Solution-processed carbon nanotubes based transistors with current density of 1.7 mA/μm and peak transconductance of 0.8 mS/μm, in *Proceedings of the 2017 IEEE International Electron Devices Meeting* (*IEDM*) (IEEE, 2017), pp. 5–6

X. Zhou, J.-Y. Park, S. Huang, J. Liu, P.L. McEuen, Band structure, phonon scattering, and the performance limit of single-walled carbon nanotube transistors. Phys. Rev. Lett. **95**(14) (2005), Art. no. 146805