An Evaluation Framework for Nanotransfer Printing Based Feature-Level Heterogeneous Integration in VLSI Circuits

Greg Leung, Shaodi Wang, Andrew Pan, Puneet Gupta, and Chi On Chui

***Abstract*—We develop an evaluation framework to assess the potential benefits of feature-level heterogeneous integration (HGI) in nanoscale VLSI circuits. We study, for the first time, the impact of HGI on circuit delay, layout area, and power by comparing the integration of 15nm InGaAs and Ge FinFETs via nanotransfer printing with baseline Si-only FinFET technology. To properly account for the performance, power, and area tradeoffs, we perform comprehensive evaluations including synthesis, placement, and routing of digital circuit benchmarks. We show circuits designed with HGI exhibit lower delay and power due to improved device performance at the cost of larger area induced by misalignment errors. We also demonstrate that HGI misalignment area penalties can be drastically reduced using post-transfer fin trimming. Our findings provide substantial motivation for industry to explore HGI as a technology route for the post-Si era.**

***Index Terms*—Design rule, FinFET, germanium, heterogeneous integration, indium gallium arsenide, nanotransfer printing, non-equilibrium Green's functions, overlay.**

I. Introduction

Heterogeneous integration (HGI) of silicon, germanium, and/or III-V semiconductor devices on a single platform can open alternative pathways to improving the performance and functionality of nanoscale integrated circuits. In contrast with homogeneous (all-Si) designs, HGI combines the advantages of disparate materials to optimize the complex requirements and tradeoffs faced in circuit design. Unfortunately, the challenges of processing dissimilar materials on a single platform have so far hindered the development of truly heterogeneous systems, especially in digital applications where feature-level (transistor-to-transistor) co-integration is desired.

Existing HGI methods typically fall under one of three categories: 1) wafer/die bonding, 2) heteroepitaxy, or 3) nanotransfer printing. Of these three technologies, wafer/die bonding is the most mature and has been the primary choice for developing 3-D ICs [1], [2] as well as circuits [3] where each heterogeneous interconnection is made by a large via typically several μm wide and deep. The simplicity of the back-end bonding process and its wide area coverage make it a popular choice for chip-to-chip heterogeneity, but the large via size precludes feature-level HGI connections [2]. An alternative technique is to use direct bonding of heterogeneous substrate stacks such as InGaAs/InP with silicon-on-insulator (SOI) as a vehicle for blanket film transfer of vertically stacked active layers which can be subsequently defined into neighboring *n*- and *p*-type field-effect transistors (FETs) for HGI circuits [4], [5]. While this method circumvents the bottleneck on achievable HGI density due to wafer bonding overlay accuracy, it unfortunately presents added challenges related to co-processing of dissimilar material technologies on a shared platform at the front-end-of-line (FEOL).

Direct heteroepitaxial growth, for instance of Ge-on-Si [6], [7], may enable feature-level HGI, but is burdened by lattice mismatch issues, thermal budget limitations, poor epitaxial film quality (unless μm-thick buffer layers are used), and throughput. Techniques such as aspect ratio trapping [8] and epitaxial lateral overgrowth [9] have been used to avoid buffer layers, but other process-related challenges still remain. Intriguingly, bonding and heteroepitaxy have both been used to co-integrate high speed/power III-V amplifiers with Si CMOS circuits in analog/mixed-signal applications with some success [3], [10].

Nanotransfer printing (NTP) involves physically transferring patterned structures from one substrate to another using an elastomeric stamp. The process is extremely versatile and has been used to co-integrate a wide assortment of materials including 3-D semiconductors, metals, quasi 2-D sheets, 1-D nanotubes/nanowires, and 0-D quantum dots on both rigid and flexible, transparent substrates for a myriad of applications [11]–[20]. Transfer printing is capable of large-area coverage, is a room temperature process, is unburdened by lattice constant mismatch, and can yield feature-level HGI so long as the transfer overlay accuracy is sufficiently high. The combination of these properties places NTP in a unique position to simultaneously deliver true feature-level HGI with minimal impact on processing requirements and materials selection compared to bonding or heteroepitaxy-based approaches. The major concerns for NTP are transfer yield and overlay accuracy compared with industry standard lithography tools; both of these issues will be discussed in further detail in Section II.

Despite much ongoing research in developing HGI processes, there is little to no understanding of the overall impact feature-level HGI would have on near-future digital circuit generations. As a preliminary study, in Fig. 1 we show the benefits of InGaAs- and Ge-based HGI over an all-Si design within a realistic processor architecture, as projected using the

The authors are with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: gleung1@ucla.edu).

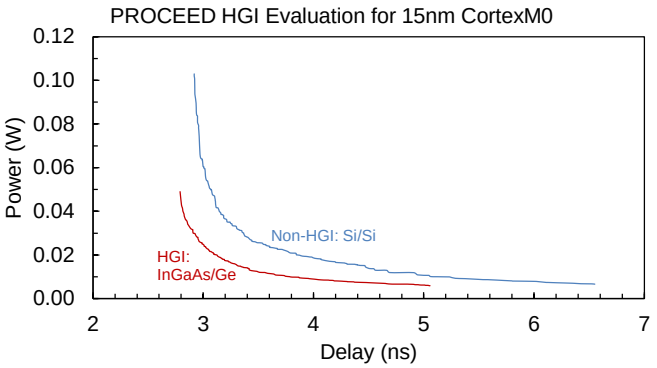
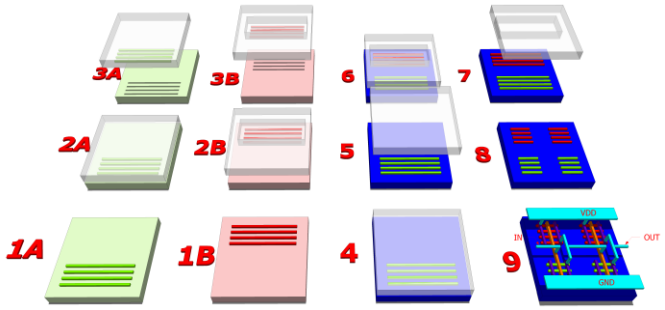


Fig. 1. Power-delay tradeoff for 15nm InGaAs/Ge and Si/Si built CortexM0 generated by PROCEED [21] [22].



1A/1B: Pattern source wafers
2A/2B: Pick-up N/P fins on stamp
3A/3B: Release stamp
4: Transfer N fins to receiving wafer
5: Release stamp
6: Align/transfer P fins to receiving wafer
7: Release stamp
8: Trim (etch) fins between different FETs
9: Gate stack formation, doping/annealing interconnect formation, metallization

Fig. 2. Process flow sequence for NTP-based HGI.

PROCEED device evaluation framework [21], [22]. Such assessments demonstrate that feature-level HGI can offer significant power savings at a given operating speed, although they do not consider layout area penalties that arise from HGI processes.

Here, we present for the first time a quantitative cross-layer study on the impact of NTP-based HGI versus Si-only technology on digital circuit performance and layout density. In Section II, we discuss the HGI NTP process in detail, including some of our preliminary experimental work in this area, and discuss current technological challenges that must be addressed for commercial usage. In Section III, we describe our HGI evaluation framework, considering achievable process capabilities (e.g., NTP overlay accuracy), intrinsic device performance, and circuit layout options. In Section IV we present the inverter- and block-level results of our cross-layer evaluation, using the specific case of VLSI circuits in 15nm FinFET technology to compare the use of all-Si FinFETs with HGI of InGaAs and Ge as the NFET and PFET channel materials, respectively. We explicitly map the technological conditions in which this HGI technology holds an advantage over Si CMOS. The results of our simple and versatile framework provide a tangible rationale for industry to seriously pursue HGI as a technology option in coming years, as we conclude in Section V.

II. HGI Process Description

A. Nanotransfer Printing Method

An illustration of the NTP process is shown in Fig. 2 where a simple FinFET buffer is implemented using different materials for the NFET and PFET devices. The active layers are first patterned into a discrete number of fins on their respective source wafers (Step 1). The fins are then undercut by a selective etching step which partly removes the underlying sacrificial layer, possibly even suspending the fins. After undercutting, an elastomeric stamp is pressed on the source substrate, causing the fins to adhere to the stamp surface (Step 2). The stamp is then delaminated from the source wafer, picking up the fins because of stronger interfacial adhesion between the fins and the stamp compared to the source substrate (Step 3). Note that the same sequence of steps is performed for each source material to be transferred. Once the stamp has picked up the NFET fins (Step 3A), it is pressed against the receiving wafer (Step 4) transferring the fins to the wafer. Unlike the pickup step, the transfer step relies on stronger interfacial adhesion between the fins and the receiving substrate compared to the stamp. Upon stamp release (Step 5), the NFET fins are successfully transferred while, in principle, preserving their original pitch, size, and number.

After the NFET fins transfer, another stamp containing PFET fins (Step 3B) is then carefully aligned and transferred to the receiving wafer (Steps 6 and 7) in a similar fashion. The alignment step is critical because it directly sets the HGI proximity and determines whether feature-level integration is possible without a significant area or yield penalty due to overlay errors. After the PFET transfer, a trim mask is used to etch away the NFET and PFET fin regions that bridge different transistors or logic gates in the circuit layout (Step 8). The use of large-area fin transfer followed by trimming has a significant benefit over small-area fin transfer for reasons to be discussed in Section III.B. Finally, remaining process steps such as transistor gate stack formation, doping and annealing, local interconnect formation, and metallization are performed as needed and can be tailored to the process requirements for the actual integrated materials.

In general, co-integration of different materials such as Si, Ge, and InGaAs may entail different thermal budget restrictions in downstream process steps. For example, the traditionally high temperatures ($T \geq 1000$°C) reached during rapid thermal annealing (RTA) in Si processing may approach or even exceed the melting point for other semiconductors like InGaAs ($T_m \cong 1100$°C) and Ge ($T_m \cong 938$°C), while a lower temperature anneal may result in sub-optimal dopant activation for inversion-mode FinFETs. There is evidence that Si^+ implanted *n*-InGaAs can reach near 100% activation for a 10 sec RTA between 750–850°C for electron sheet densities up to 5×10^{14} cm^{-2} [23], while B^+ implanted *p*-Ge can be fully activated even without any post-implant annealing for hole sheet densities up to 10^{14} cm^{-3} and BF_2^+ implanted *p*-Ge can be fully activated after a 30 min. low temperature anneal of 350°C [24]. Experimental demonstrations have also shown successful use of sub-800°C RTAs for post-implant dopant activation in InGaAs FETs [25]–[27] and sub-400°C fabrication of entire Ge PFETs [28]. These findings suggest that simultaneous HGI processing of InGaAs and Ge may be possible for inversion-mode devices requiring precise junction definition. On the other hand, co-integration of Ge or InGaAs with Si may be

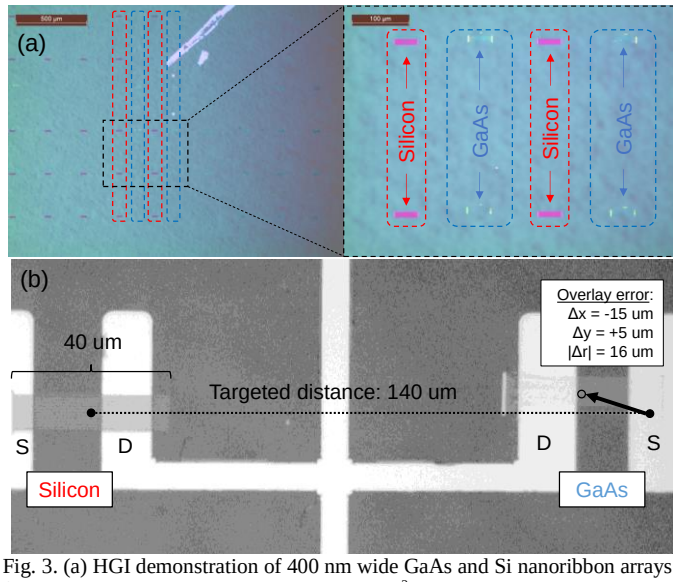


Fig. 3. (a) HGI demonstration of 400 nm wide GaAs and Si nanoribbon arrays formed by NTP on SiO_2 substrate with mm^2 area coverage. (b) Measured overlay error (16 μm) after aligned transfer and source/drain electrode formation using optical lithography. The dotted line in (b) illustrates the as-designed separation between the Si and GaAs NR arrays after transfer, while the solid arrow indicates the vector of transfer misalignment which quantifies the overlay error in (x,y). The overlay error magnitude is ~16 μm.

more problematic because of the much higher anneal temperatures required for dopant activation in Si.

Alternatively, uniformly doped junctionless FETs (JLFETs) [28] are particularly suitable for HGI because of their relaxed thermal budget requirements. Since the channel materials can be doped *in situ* during growth on the source wafers, it is possible to circumvent subsequent high temperature processing such as post-implant RTA. However, intra-device variability from line edge roughness and random dopant fluctuation has been shown to be more significant in JLFETs compared to equivalent inversion-mode FETs [30]–[32]; it remains to be seen whether heightened variability will pose a significant obstacle to commercial adoption of JLFET technology in HGI or conventional settings. For these reasons, we focus our analysis on co-integration of InGaAs/Ge inversion-mode FinFETs in this work.

B. *Transfer Alignment Accuracy*

As discussed earlier, the accuracy of the aligned transfer step is a critical factor for realizing NTP-based heterogeneous circuits. The primary bottlenecks to alignment accuracy are the limited resolution of the optical systems (e.g., contact or stepper aligners) used to perform the alignment, the precision of the (x, y, and θ axis) stage movement, and the topography of the stamp and receiving wafer over large areas. Academic research efforts have demonstrated alignment and transfer of heterogeneous structures with overlay errors on the order of μm to tens of μm [11]–[20]. Fig. 3(a) depicts one of our efforts [11] to transfer aligned periodic arrays of 400 nm wide GaAs nanoribbons (NRs) next to Si NRs at predefined locations over a large area using a Karl Suss MA6 contact aligner. The misalignment vector in Fig. 3(b) gives us an estimated overlay error of ~16 μm for the setup, which may be improved with better tools.

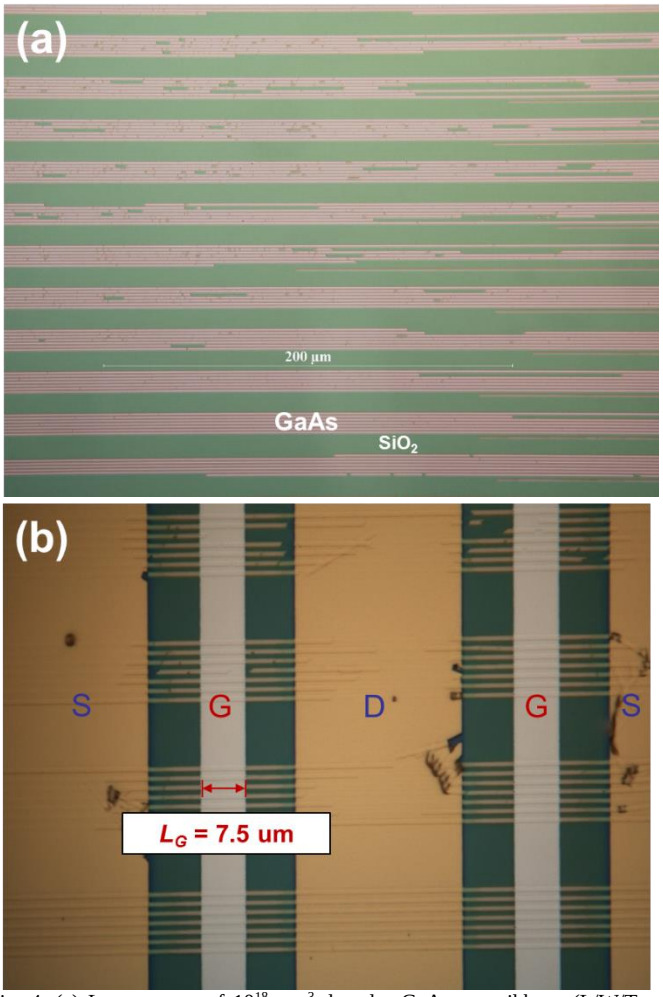


Fig. 4. (a) Large arrays of 10^{18} cm^{-3} doped *n*-GaAs nanoribbons (L/W/T = 400/0.75/0.03 μm) transferred to SiO_2/Si. (b) Example of a L_G = 7.5 μm GaAs junctionless MOSFET fabricated on SiO_2/Si. Discontinuities along the nanoribbons indicate broken segments resulting in <100% yield.

Clearly, such overlay errors are too high for nanoscale technologies where the proximity between NFET and PFET fins may be below 100 nm. Commercial steppers with overlay errors of less than 10 nm [33] may provide the needed alignment accuracy if they can also be modified to perform the transfer process, although the overlay tolerance may still exceed several tens of nm due to more severe topography issues and mechanical properties of the stamp (which is usually quite soft and flexible). Since there is no consensus on what σ values can be obtained (or will be needed) from NTP for use in future technologies, we surmise expected values of σ from 3 nm up to 50 nm for use in our framework, representing possible field-size "step and transfer" scenarios using state-of-the-art tools [34] derived from nanoimprint lithography (NIL). Recent NIL demonstrations [35] have shown minimum overlay errors of $3\sigma \cong 10$ nm for templates up to 2 × 3 cm^2 fields, so our projected σ values for NTP in this work should be reasonable—if not conservative—based on the similarities between NIL and NTP.

C. *Similarities between NIL and NTP*

Both NIL and NTP are contact processes which use physical contact to either form or transfer patterned features to a

substrate. In NIL, a mold containing the feature to be printed on the receiving substrate is physically pressed onto a UV-curable liquid resist layer on the substrate which results in displacement of the resist to conform to the mold's patterned shape. After the mold and resist are contacted, the resist is cured in light to solidify it and the mold is removed. Because NIL is a contact process, the overlay tolerance must be well controlled since features are printed at a 1:1 ratio with no magnification. The overlay accuracy depends on the precision of the stage movement, uniformity of the resist layer and the flatness of the mold and substrate surfaces [36], with the best demonstrations to date reaching $3\sigma \cong$ 10–15 nm [35] as mentioned in Section II.B. Using a "step and imprint" technique [36], [37] the mold template can cover an entire field to balance throughput and accuracy over a large area.

In NTP, a soft adhesive stamp is used to pick up patterned structures from one substrate and transfer them to another. Unlike NIL, there is no actual lithography during the transfer process. Like NIL, however, the printing process relies on physical contact between two surfaces meaning overlay accuracy will depend on flatness of the receiving substrate and the stamp containing patterned structures. In our experiments, the polydimethylsiloxane (PDMS) stamp could vary in thickness by several hundred μm over a region of several cm^2. This can severely affect the alignment process due to limited depth of field in the equipment optics; when coupled with deformation of the stamp during contact transfer, the achievable overlay accuracy over large areas may be substantially limited compared to what is theoretically possible based on the (x, y, θ) precision of the stage movement. Because of this, it is reasonable to expect that NTP overlay accuracies based on our current experimentation capability may not yet reach those of the best NIL demonstrations to date. The reader should bear in mind, however, that engineering of the stamp properties may substantially reduce the severity of these issues, especially compared to what has/can be demonstrated in academic laboratories.

D. Transfer Yield and Performance Loss Considerations

Besides overlay accuracy, the transfer yield must be high enough to ensure that the process is manufacturable (i.e., repeatable) for commercial use. Since NTP as an HGI enabler is still in the early stages of research and development, reliable data about transfer yield is currently sparse. The authors in [20] claimed 87%, 95%, and 99% transfer yield in their experiments for GaN, GaAs, and Si microribbons transferred to plastic substrates, indicating promise for this technology. Fig. 4 shows our experimental results for 10^{18} cm^{-3} doped *n*-GaAs nanoribbons (NRs) transferred to SiO_2/Si substrate. The nominal length, width, and thickness of individual NRs is 400, 0.75, and 0.03 μm. Some of the NRs show broken or missing segments, indicating yield loss either during the undercutting, pickup, or transfer stages. Currently, we estimate <10% transfer yield for our process. The length-to-width aspect ratio (AR) of our transferred NRs in Fig. 4 is ~533:1, which is among the highest reported values to date and, to our knowledge, the highest result for sub-1 μm wide features. A deeper investigation of the yield loss mechanisms and potential routes for improvement thereof are subjects of ongoing research, the results of which are expected to give more insight into what HGI circuit layout methodologies should be selected to enable more robust designs.

Potential reasons for lackluster yield include microscopic variations in undercutting rates, bending stresses, poor adhesion strength between the stamp and semiconductor surface, and AR constraints resulting from the limited structural integrity of nanoscale features during undercutting (and possibly suspension), pickup, and transfer. The probability of successful pickup hinges on the adhesion differential between the stamp-feature and feature-substrate interfaces: the former must exceed the latter in order for active features to be lifted off from the source substrate by the stamp. In the case of PDMS stamps with peel-rate-dependent adhesion strength (a consequence of viscoelasticity), this is normally accomplished using a fast peel-back speed (e.g., 10 cm/s or higher) after the features have been sufficiently undercut [16]–[18]. More complete undercuts increase the adhesion strength differential to increase the chance of pickup, but excessive undercuts also make stiction-induced collapse from capillary forces during drying [38] more likely to occur when selective wet (isotropic) etching is used, especially for ultrathin, fragile features such as the NRs shown in Fig. 4 and possibly those with extremely long ARs. Because of the random nature of wet etching, significant variations in undercutting were commonly observed in our experiments, leading to unpredictable (and often poor) pickup yield. To circumvent these issues, isotropic dry etching could be used to fully suspend active features which would avoid the problems of stiction collapse and variable undercutting, thereby resulting in more predictable pickup yields.

During pickup and transfer, the stamp undergoes elastic deformation as it is directionally peeled off of a substrate. This can lead to bending and possible fracture of active features which result in further yield loss. The speed and direction of peeling can have a significant influence on the transfer pattern and the ultimate yield. Generally, a slow peel-back speed (e.g., 1 cm/s or lower) is desirable for transferring features to the receiving substrate [16]–[18].

Even if 100% transfer yield can be achieved, the quality of transferred materials may be degraded after the stamping process. For example, the backside interface between the transferred fins and the receiving substrate could exhibit a higher density of interface traps due to poor bonding quality between the different materials, resulting in higher leakage current and parasitic capacitance. Since the amount of degradation will very likely be material- and process-dependent, it is difficult to quantify these effects without detailed experimental analysis. Some evidence suggests that NTP does not appreciably degrade the front-side interface between the channel and gate dielectric in terms of measured subthreshold characteristics from InAs-on-insulator FETs fabricated through a similar process [39], but more extensive studies will be needed to support this finding, especially regarding the backside interface properties. These topics remain the subject of ongoing research on our part.

III. HGI Evaluation Framework

To project the ultimate effects of HGI on future digital systems, we have developed a general evaluation methodology which we apply to the specific case of NTP-based integration.

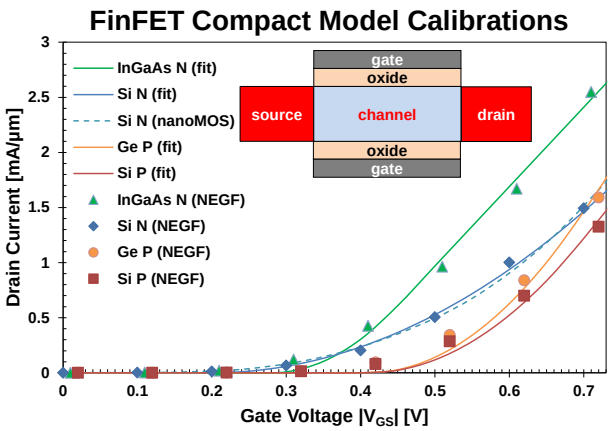


Fig. 5. NEGF (symbols) and model fit (lines) $|I_D|$-$|V_{GS}|$ curves for Si, Ge, and InGaAs double-gate FinFETs. The dashed line represents the NEGF Si NFET simulation using nanoMOS [45]. All simulations are with drain bias V_{DS} = 0.73 V. Inset: double-gate structure used for simulations.

Our framework is divided into three stages, 1) device simulation, 2) compact model calibration, and 3) circuit analysis, which are used to predict the benefits of an HGI (over non-HGI) implementation given the following data: 1) device specifications for a desired technology node, 2) an HGI process to implement the technology, and 3) representative circuit layouts for the technology, which will be used for benchmark comparisons. The following sections explain each part of the framework in more detail.

A. *Device Modeling*

Because experimental data on scaled III-V MOSFETs is sparse, we use simulations to project *I–V* and *C–V* device performance at the 15nm node studied here. For maximal accuracy, we use non-equilibrium Green's functions (NEGF) [40] to perform quantum mechanical device calculations and capture important phenomena like ballistic transport and tunneling that cannot be fully modeled by conventional technology computer-aided design (TCAD). We employ our own 2-D NEGF code [41] to simulate 15nm ITRS [42] Si and $In_{0.53}Ga_{0.47}As$ FinFETs.

Our device structure is shown in the inset to Fig. 5, with physical gate length of 12.8 nm, channel thickness of 8.5 nm, oxide thickness of 0.68 nm, and supply voltage V_{DD} = 0.73 V. For all devices, gate work functions are adjusted to set the leakage current to 100 nA/μm. These values are taken from the ITRS projections for 15nm multigate devices [42]. Our simulations assume ballistic transport, i.e., no scattering, which represents the upper bound of performance. Experiments show that devices are indeed approaching this limit as they scale, albeit more quickly for III-V compared to Si [43]. For *n*-type FinFETs, we perform effective mass simulations for Si and $In_{0.53}Ga_{0.47}As$ to extract device characteristics. We use three band k-p to simulate the Si PFETs; due to computational complications with the Ge band structure, we approximate Ge PFET devices by scaling the Si characteristics by 20%, in accordance with other ballistic studies that show this enhancement ratio [44].

Lastly, we fit standard compact models to the simulated *I–V* curves for circuit delay calculations (Section III.C) by adjusting parameters like mobility and saturation velocity. To validate our device simulations, we also compare our *n*-Si simulation with that performed using the standardized NEGF simulator nanoMOS [45] and observe close agreement. The characteristics and fits are shown in Fig. 5. We also extract the averaged off- and on-state capacitance for each device; for Si NFET and PFET and Ge PFET, this value is about 0.42 fF/μm, whereas it is about 0.27 fF/μm for InGaAs NFET. The reduced capacitance for III-V *n*-type devices is a well-known effect due to the conduction band density of states of such materials [46].

B. *HGI Impact on Circuit Layout and Design Rules*

All HGI circuit designs face two new complications: a potential loss in intrinsic device performance (a likely problem for heteroepitaxy-based HGI due to crystal defects) and a reduction in layout density (particularly important for transfer-based HGI due to overlay accuracy limitations). The former effect can be accounted for by adjusting the device models presented in Section III.A, but this is not easy to predict without extensive experimental data on HGI process-induced degradation of device characteristics. On the other hand, density loss can be easily accounted for by adjusting layout design rules, given some knowledge of the NTP overlay accuracy. Since we are mainly concerned with transfer-based HGI in this study, we will assume an ideal case where no loss in device performance occurs and focus on the layout area penalty from the NTP process. For this study, we assume that the NTP process occurs with 100% transfer yield; that is, no fins are missed or broken during the pickup and transfer steps and the channel quality is not degraded in any way by the transfer. This is certainly optimistic, but it allows us to set an upper limit for the foreseeable gains from HGI. Certainly, a more realistic projection of NTP-HGI technology demands the inclusion of non-ideal transfer yield and the possibility of material degradation from the transfer process (e.g., higher interface states, structural defects, etc.), but these effects are poorly understood at the moment. We believe that further research into this area is desperately needed to pinpoint the critical issues related to NTP and whether or not any systematically-dependent yield loss from transfer printing can be mitigated through smarter layout strategies. We do allow for misalignment of fins, however, resulting in "alignment yield" < 100%. We will not explicitly consider any rotational (θ) misalignment in this study, though in principle its effects can be absorbed into additional *x* and *y* translational misalignments, which would get progressively worse at locations further from the point-of-alignment. A conservative approach to solve this problem is to modify the global design rules to accommodate for the worst case local misalignment.

Another possible issue may arise from deformation of the stamp during pickup and transfer due to shear forces which may cause the fin-to-fin spacing to stretch beyond its designed value. This could result in local, nonuniform variations in overlay error which differ from the global (systematic) overlay error. Assuming the amount of shear deformation is quantifiable, in theory we could also include its effects in σ when generating the design rules and cell layouts.

A simple FinFET inverter layout is shown in Fig. 6, where the PFET fins have been transferred with a one-sigma overlay error of $\pm\sigma$ and must satisfy three conditions: 1) the PFET fins

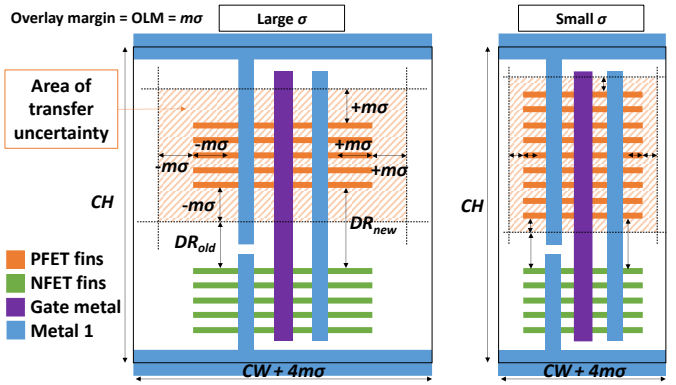


Fig. 6. Schematic layouts for heterogeneous FinFET inverters from NTP without fin trimming. The area of transfer uncertainty indicates the region where PFET fins can land due to misalignment.

must not land too close to the NFET fins, 2) all fins must be contacted by the drawn source, drain, and gate lines, and 3) all fins must lie within the cell boundaries. Each of these conditions imposes extra constraints on the appropriate design rules.

Separation of the PFET and NFET fins ensures that they do not overlap during or after transfer, causing device failure. In non-HGI layouts, a design rule setting the minimum distance DR_{old} between NFET and PFET fins exists due to the masked diffusion or implantation steps for the two devices; however, this minimum distance is not too large (~35 nm for the 15nm node) since it is set by lithography. In HGI layouts, however, the new minimum separation DR_{new} is increased by some multiple m of the transfer overlay accuracy σ, which may be significantly larger (~10 to 100+ nm). In other words, $DR_{new} = DR_{old} + m\sigma$. Determination of m is not straightforward and directly impacts the resulting alignment yield and area penalty at the cell level, as we will see later. Our approach for choosing m is detailed in Section IV.A. Conditions 2 and 3 impact the HGI layout area penalty differently depending on the presence of a trim step after fin transfer (Step 8 in Fig. 2), meriting a separate discussion.

1) HGI without Fin Trimming

The requirement that all fins be properly contacted has two consequences. First, the fin length must be extended by $m\sigma$ on each end, meaning the minimum fin length increases by $2m\sigma$ in order to guarantee proper electrical contact when a $\pm m\sigma$ horizontal HGI misalignment occurs. This also means the minimum cell width (CW) must increase by $2m\sigma$ to accommodate the longer fins when HGI is used. Second, to absorb any vertical HGI misalignment, the maximum number of transferable PFET fins per cell is reduced to a value dictated by the fin pitch, the minimum fin-to-metal 1 (M1) overhang, and the minimum M1-to-M1 separation. The end result is that fewer PFET fins can be transferred within a minimum size cell when HGI overlay accuracy is poor compared to the non-HGI case; a "stronger" PFET will require more transistor folding and consume a larger cell area.

Finally, to enforce the cell boundaries and account for any vertical misalignment, the minimum distance from the PFET fins to the top of the cell becomes $m\sigma$. This sets another limit on the number of PFET fins that can be transferred within a minimum sized cell. More catastrophically, the cell boundary

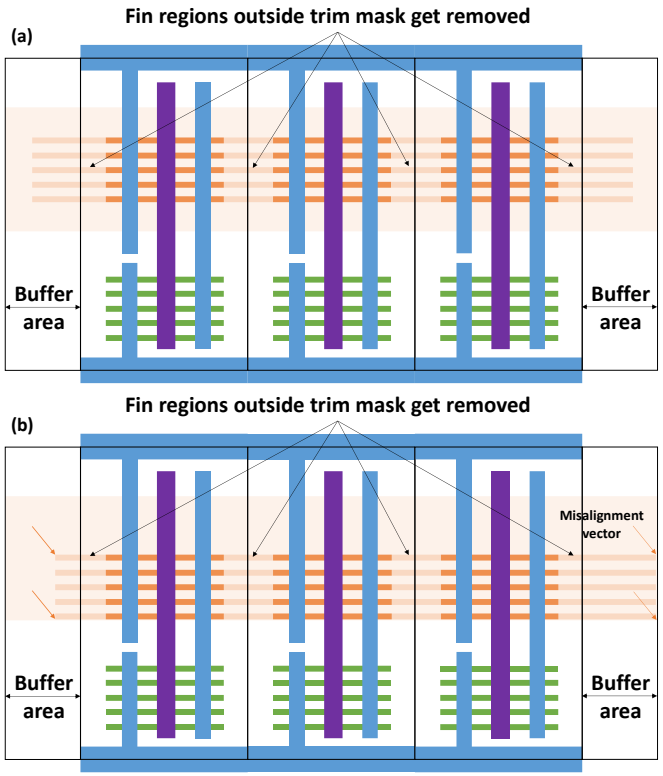


Fig. 7. (a) Schematic layout for a row of heterogeneous FinFET inverters made with NTP and fin trimming. (b) The effect of transfer misalignment with fin trimming is now absent within each cell except at the buffer areas on ends of a row.

condition also forces the cell width to increase by an additional $m\sigma$ on each side for a net increase of $2m\sigma$. Adding this to the $2m\sigma$ penalty from using longer fins means the width of *every* cell must increase by a total of $4m\sigma$, absent fin trimming. For instance, if σ = 50 nm and m = 2 (for 95% alignment yield), every cell would widen by 400 nm, thereby increasing cell area by more than 5× over a 15nm non-HGI design.

2) HGI with Fin Trimming

Fin trimming (see Step 8 in Fig. 2) effectively removes the impact of lateral misalignment on layout except for the cells at the ends of a row. This is because lateral misalignment will only appear at the left and right fin ends as shown in Fig. 7, which will inevitably be removed after the trim. Within each row there is no need for the fins to be longer than normal to guarantee electrical contact, nor is there a need for extra room in the $\pm x$ direction to keep neighboring transistors isolated since the trim step guarantees it. Thus, the cell width does not increase (discounting transistor folding) to accommodate HGI overlay.

The only area penalty incurred is the addition of two dedicated empty regions (at least $2m\sigma$ in length) which absorb the misalignment penalty at the very ends of each transferred fin. Since the empty regions can sandwich many active cells within a row, this area penalty is amortized across the cells, mitigating the per cell penalty, and reduces as the transferred fin length increases. Most likely, however, arbitrarily long fins cannot be transferred with good yield due to complications from microscopically variable undercutting rates before fin

pick-up, peeling forces during transfer, and stamp surface topography. We speculate that transfer yield may be correlated to the fin length/width AR, limiting the transferrable fin length and per-cell penalty reduction. Unfortunately, exact constraints on the AR are not clear at this point due to limited experimental evidence; this will be revisited later in Section IV.C.

Ultimately, compared to non-HGI circuits of equal performance, circuits using HGI will incur a layout density hit that is dependent on σ as well as the number of fins in each cell (i.e., the cell strength). As an example, for a given cell height, a minimum size inverter with just one NFET and PFET fin can tolerate a larger misalignment due to the large amount of empty space in the cell, whereas a cell containing more NFET and PFET fins can only tolerate a small misalignment before design rule violations occur. Consequently, for a given HGI process (i.e., a given value of σ), only some circuit cells will incur a layout area increase.

Finally, the reader may note that all area penalties mentioned originate only from the PFET transfer. The reason is that the NFET fins are transferred to the receiving wafer before any other patterns are formed, so they serve as the reference to which all other features (PFET fins, gate/interconnect lines, etc.) are aligned. As such, the alignment-related penalties discussed here only apply to PFETs.

C. Circuit Level Evaluation

In our framework, we use UCLA Design Rule Evaluator (DRE), a free online tool [47], to generate 15nm FinFET circuit layouts using modified design rules to account for the HGI-related penalties. For simplicity, the 15nm design rules are first obtained from a scaled version of an existing 45nm [48] planar process where all dimensional quantities are scaled by 15/45 = 33%. Once a nominal set of rules is obtained, a subset is modified to account for the different methods of FET formation: physical transfer in the HGI process, and standard lithography plus etch for the non-HGI process. The actual rule values used in our study will be discussed later in Section IV.A.

With the design rules in place, we synthesize a 15nm cell library using Nangate Open Cell Library [48] as a template[1] and scale all transistor sizes to match the 15nm node. All FinFETs have gate length L_g = 13 nm and effective width $2N \times H_{fin}$, where N is the number of fins per transistor and H_{fin} = 17 nm is the fin height. After the cell layouts are generated, switching delays are estimated for each cell using the fitted compact model parameters discussed in Section III.A. Using this simple model, we can rapidly compare the cell-dependent impact of different HGI technology and design rule scenarios without brute force circuit simulations over an entire library. Once the cell library is characterized for each type of process (HGI and non-HGI), we compare the relative delay–area and delay–power impact across a few benchmark designs for a full chip-level HGI evaluation.

[1] While the 15nm library used is not derived from an actual commercial FinFET library (bearing in mind that no such library has been made publicly available), we believe our findings should still be useful to the design community even if the reported results are based on projected inputs.

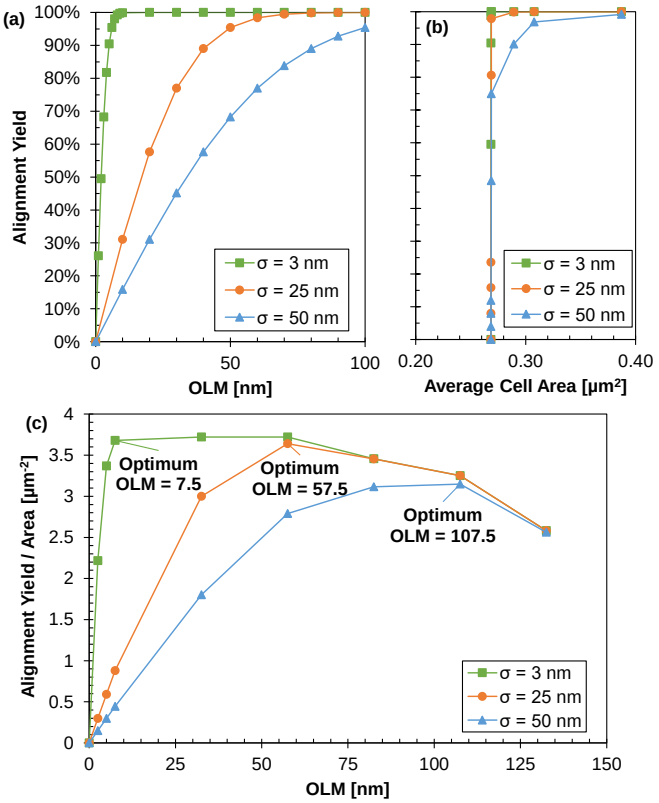


Fig. 8. (a) Probability of successful fin placement as a function of transfer misalignment and allotted overlay margin. (b) Alignment yield versus average cell area in reduced MIPS processor. (c) Optimal OLM value search to maximize alignment yield per cell area.

IV. Projected HGI Benefits

A. Setting the HGI Design Rules

In Section III.B, we noted that DR_{new} must exceed DR_{old} by at least $m\sigma$ to ensure good alignment yield (AY) without consuming excessive area. Assuming a Gaussian distribution for the overlay error, we have alignment yield AY(OLM) = erf($m/\sqrt{2}$) where erf is the error function and OLM ("overlay margin") = $m\sigma$. This is plotted in Fig. 8(a) for σ = 3, 25, and 50 nm, representing expected NTP capabilities as discussed in Section II.B. As shown in Fig. 6, OLM essentially represents the extra space needed on all sides of the PFET fins to account for transfer misalignment. Note that an increase in OLM only results in larger cell area if transistor folding becomes necessary for a given cell height (CH) and cell strength (number of fins).

To arrive at the best compromise between alignment yield and circuit density, we find the optimum OLM by maximizing the alignment yield per unit average cell area for a given HGI process (i.e., value of σ); this is analogous to optimizing the design rules to obtain the maximum number of good dice per wafer. Since different cell types have different optimum OLM, we consider a reduced size MIPS processor [49] as our benchmark and compute an average cell area weighted by the number of cell instances of each type. The calculated results are given in Fig. 8(b) which provide a mapping between the alignment yield from Fig. 8(a) and average cell area for different possible overlay margins ranging from 0 to 132.5 nm.

TABLE I
MODIFIED 15NM DESIGN RULES FOR DIFFERENT PROCESS SCENARIOS

Process	P-N spacing (intra-cell) [nm]	P-P spacing (inter-cell) [nm]	Minimum cell dimensions [nm]
Non-HGI	H: n/a V: 35	H: 72 V: 72	H: 72 V: 506
HGI (no trim)	H: n/a V: 35+OLM=92.5	H: 72+2OLM=187 V: 72+2OLM=187	H: 72+4OLM=302 V: 506
HGI (trim)	H: n/a V: 35+OLM=92.5	H: 72 V: 72+OLM=129.5	H: 72 V: 506

Note: For HGI processes, σ = 25 nm and OLM = 57.5 nm are used. "H/V" specifies design rule value in horizontal/vertical direction.

Moving from bottom to top along each curve in Fig. 8(b) indicates a progressive increase in OLM; along the vertical segments, OLM is increasing but the average cell area is not, meaning that transistor folding is not occurring in any of the benchmark cells yet. Dividing the alignment yield values by the average cell area values and plotting the results versus OLM in Fig. 8(c), we can search for the OLM values which maximize the yield-to-area ratio. Based on the data, for σ = 3 nm, 25 nm, and 50 nm, the optimum values of OLM are 7.5 nm (m = 2.5), 57.5 nm (m = 2.3), and 107.5 nm (m = 2.15) respectively, corresponding to alignment yield of 99, 98, and 97%, respectively. For the σ = 3 nm case, there is negligible difference in yield-to-area between OLM = 7.5 nm and 57.5 nm, so we just pick the smallest value. As a rule of thumb, it appears OLM $\cong 2\sigma$ is a good choice for the allotted overlay margin due to misalignment.

Table I summarizes the modified design rules for HGI circuits assuming a transfer accuracy of σ = 25 nm as well as the baseline design rules for non-HGI circuits. Here, we have DR_{old} = 35 nm (introduced in Section III.B) and OLM = 57.5 nm. The P-N spacing is the same as DR_{new} and accounts for misalignment in the $-y$ direction, while the P-P spacing accounts for misalignment in the $\pm x$ and $+y$ directions.

B. Inverter Delay vs. Area Evaluation

Using the framework described in Section III, we examine the tradeoffs between delay and area for FinFET inverters of varying strength (i.e., number of fins) implemented in either InGaAs/Ge (HGI) or Si/Si (non-HGI) processes: the notation "A/B" refers to a cell using material "A" for the NFET and "B" for the PFET. For each set of design rules per process scenario, we obtain a series of inverter delay—area curves such as those shown in Fig. 9. Starting from the top of each curve and moving downward, each successive marker represents an increment in the number of PFET and NFET fins in the inverter, beginning with 1 and ending at 20 fins, mapping out the inverter's delay and area as a function of cell strength from 1X to 20X. The cell height (CH) in each case is either 11 or 15 (Metal 3) tracks for InGaAs/Ge inverters, while for Si/Si inverters CH is fixed at 11 tracks.

1) Without Fin Trimming

When fin trimming is neglected, there is no point at which any of the $\sigma \geq$ 25 nm HGI configurations holds a clear advantage over the Si/Si baseline, as evidenced by comparing the curves in Fig. 9(a) at a given delay value: the baseline can always provide the same delay while consuming a smaller footprint. We also observe that taller cells pay a larger initial area overhead compared to shorter cells but become more attractive as the cell strength increases, since fewer transistor

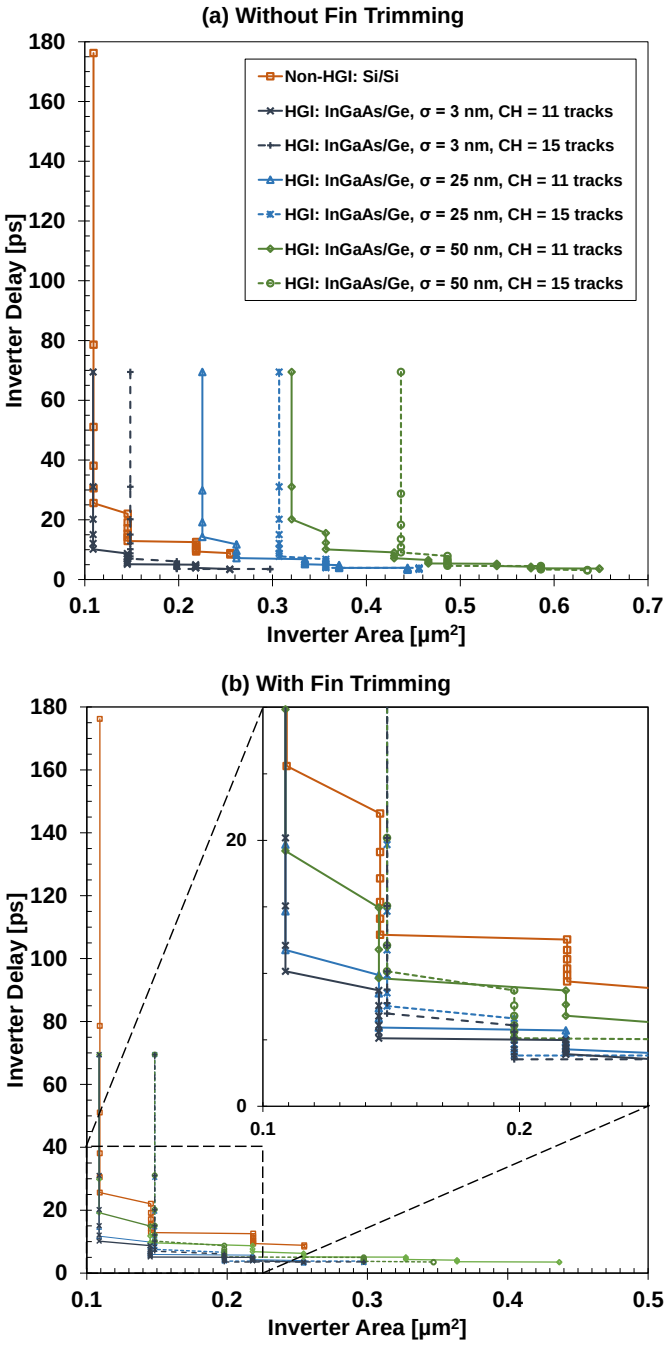


Fig. 9. Delay versus area for 15nm InGaAs/Ge (HGI) and Si/Si (non-HGI) inverters for different σ and CHs (a) without fin trimming and (b) with fin trimming. The inset is a magnified view of the dashed region in (b).

folds are needed in a taller cell. High performance circuit blocks using many fins per transistor can be designed with taller cells to minimize folding, while low power blocks using only a few fins per transistor can be designed with shorter cells to minimize the area overhead. As expected, large σ values result in larger areas due to more frequent folding. When σ = 3 nm, however, most of the InGaAs/Ge curve lies well below the Si/Si baseline, indicating substantial benefits. This represents the upper limit of what HGI can offer assuming such accurate transfers are possible, bearing in mind the caveats of Section II.B.

TABLE II
AREA, DELAY, AND POWER OF HGI STANDARD CELLS COMPARED TO NON-HGI CELLS.

A) SUMMARY OF AREA COMPARISON

Cells in library with area overhead from HGI	Average weighted area overhead (MIPS[1])
24 of 114	6.6%

B) SUMMARY OF DELAY COMPARISON

Cells in library with delay reduction from HGI	Average weighted delay reduction (MIPS)
114 of 114	62%

C) SUMMARY OF POWER COMPARISON

Cells in library with power reduction from HGI	Average weighted power reduction (MIPS)
114 of 114	18%

D) STANDARD CELLS WITH AREA OVERHEAD

Cell	Area overhead	Cell	Area overhead
AOI222_X4	94.6%	OAI222_X2	21.9%
NOR4_X2	85.5%	BUF_X4	19.7%
INV_X8	74.7%	AND2_X4	16.4%
NAND2_X4	74.7%	OR2_X4	16.4%
AOI222_X2	49.3%	OAI222_X4	15.3%
NOR4_X4	46.4%	AND3_X4	12.3%
INV_X16	40.3%	OAI21_X4	12.3%
NAND3_X4	36.4%	OR3_X4	12.3%
BUF_X8	33.1%	AND4_X4	9.9%
BUF_X16	33.0%	OR4_X4	9.9%
INV_X4	32.9%	INV_X32	8.2%
NAND4_X4	31.4%	BUF_X32	6.9%

2) With Fin Trimming

When fin trimming is included, the layout area penalty is reduced such that all the HGI delay–area curves in Fig. 9(b) have at least some advantageous regions that lie beneath the baseline. In fact, for σ = 25 nm and CH = 11 tracks, nearly the entire curve lies below the Si/Si baseline with the InGaAs/Ge inverter able to provide >50% reduction in delay for the same area. For σ = 50 nm and CH = 11 tracks, InGaAs/Ge still offers benefits, but the constant-area delay reduction is only roughly 25%. For taller cells (CH = 15 tracks), the initial overhead represents an extra 35% area cost for the weakest cells but starts to pay off once the cell strength exceeds 10X when σ = 25 nm and 6X when σ = 50 nm. The benefits of migrating to a taller CH (11→15 tracks) are more apparent when σ is larger: folding frequency is reduced from every six to every eight fins (25% less folding) when σ = 25 nm, but from every three to every six fins (50% less folding) when σ = 50 nm. Depending on the balance of weak and strong cells in the circuit design, it may be advantageous to design with taller cell heights everywhere when adopting HGI, especially if the transfer accuracy is poor.

C. Block Level Evaluation

For circuit block analysis, we again investigate designs involving either non-HGI (Si/Si) or HGI (InGaAs/Ge) configurations. We modify the digital circuit backend flow to properly account for the area adjustments induced by NTP. For misalignment, following the arguments above, we have seen that the use of fin trimming essentially eliminates OLM in the *x* direction and any cell area penalties arise only from *y* direction OLM and added transistor folding. We use UCLA DRE [47] to generate all HGI standard cells based on the Nangate Open Cell Library templates [48] with calibrated design rules including misalignment penalties as discussed in Sections III.C and IV.A. In all results to follow, the use of post-transfer fin trimming is assumed. The area overheads, delay reduction, and power reduction of HGI cells compared to non-HGI are given in Table II, assuming OLM = 50 nm (roughly corresponding to σ = 25 nm) and CH = 11 tracks. We see 24 out of the 114 HGI cells incur an area penalty (of up to 94.6%) and an overall 6.6% area increase is seen after weighted averaging based on usage in MIPS. The stronger drive currents and lower capacitance from InGaAs/Ge HGI result in lower delay and power for all 114 standard cells in the library. Circuit benchmarks are then synthesized in a commercial synthesis tool using these standard cells.

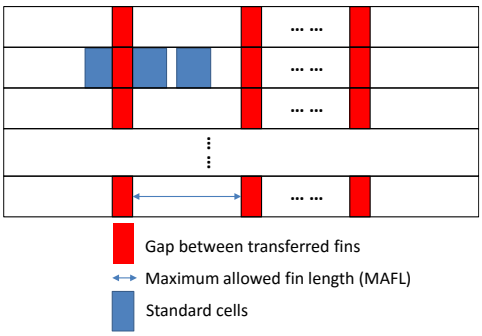


Fig. 10. Protocol for block-level HGI design. A grid of dummy filling cells (red cells) are inserted pre-placement to represent the effect of finite fin length, and standard cells (blue cells) are then placed in between the filling cells.

In addition to the OLM requirements, there may be limits to the transferable fin length imposed by the fin aspect ratio as mentioned in Section III.B. In practice, this means that sets of long but finite fins will be transferred, with additional gaps between adjacent fins. However, these sets are transferred simultaneously without incurring relative OLM, so that the only added area overhead comes from gaps between the sets. To include these gaps, prior to cell placement we insert a grid of dummy filling cells on the placement rows separated by a distance equal to the maximum allowed fin length[2] (MAFL) as shown in Fig. 10. These filling cells are temporarily fixed in the layout and the design cells are then placed using a commercial placement tool. The filling cells guarantee that the fin length, which is the width of the connecting cells, does not exceed the MAFL. The width of the fill cells is set to the minimum gap required in the transfer process (2OLM). After placement, the dummy cells are removed and routing is performed.

In the block-level simulations, multiple delay constraints are set during circuit synthesis using different technology libraries. Synthesized circuits are then placed and routed (P&R) within a fixed-size die with a grid of filling cells. This die size accommodates the Si/Si baseline design with 80% utilization. We first compare the *pre*-P&R delay versus area tradeoffs of HGI and non-HGI implementations in MIPS and AES [49]

[2] The MAFL represents the hypothetically longest fin length which can be transferred with 100% yield (which is assumed throughout this work), considering the process challenges mentioned in Sections II.D and IV.B. The higher the MAFL, the better it is.

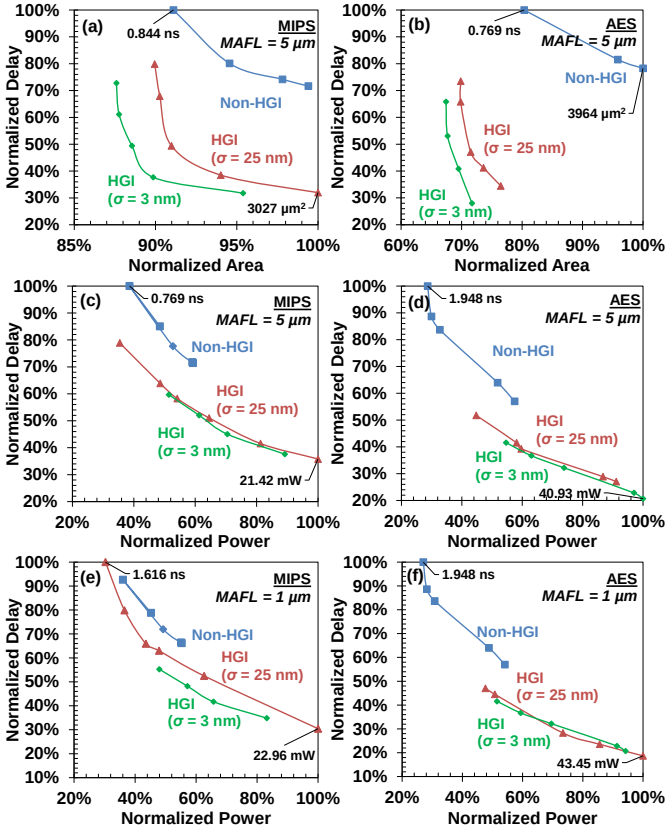


Fig. 11. Post-synthesis (pre-P&R) normalized delay and area of (a) MIPS and (b) AES designs. Post-P&R normalized delay and power of MIPS and AES designs with MAFL of (c,d) 5 μm and (e,f) 1 μm, respectively. In each panel the reported data is normalized to the largest observed delay, power, or area values as indicated by the data labels. The design rules (i.e., OLM values) are chosen to ensure 95% yield in all cases.

benchmark designs to estimate the impact of misalignment-induced penalties. Then the *post*-P&R delay and power are compared for the same benchmarks, which also includes the penalty from reserved areas (gaps) between adjacent but disconnected sets of transferred fins. The results to follow assume that MAFL = 5 μm and the gap between adjacent sets of fins is given by 2OLM. We consider two situations for HGI: 1) an ideal scenario of σ = 3 nm (OLM = 6 nm) which essentially means no penalty[3] from misalignment, and 2) a more realistic scenario of σ = 25 nm (OLM = 50 nm). In both cases the alignment yield is 95% according to the analysis in Section IV.A. By comparing these two cases with the non-HGI scenario, we can separate the gains in chip performance/density due to the use of InGaAs and Ge as channel materials from the degradation due to the transfer technology.

In Fig. 11(a)–(b) we present the *pre*-P&R normalized delay–area curves for MIPS and AES benchmarks under the HGI and non-HGI scenarios introduced earlier. Clearly, the InGaAs/Ge design outperforms the non-HGI design in both delay and area efficiency. From Section III.A, InGaAs and Ge both offer stronger drive current than Si, while InGaAs also possesses lower intrinsic capacitance than Si due to its lower

[3] The MAFL is assumed to be infinite for σ = 3 nm since a 6 nm OLM overhead is already easily satisfied by the default (non-HGI) standard cell design rules, and thus no extra "filling cells" are ever needed in Fig. 10 nor do any of the HGI standard cells require enlargement from their default non-HGI sizes.

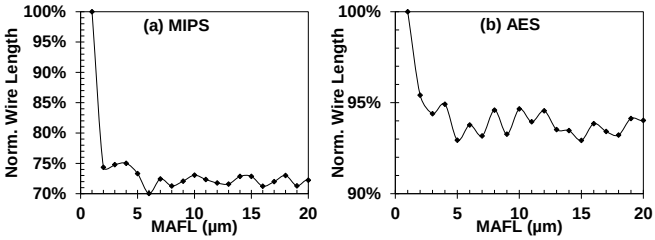


Fig. 12. Total interconnect length as a function of maximum allowed fin length for HGI-based (a) MIPS and (b) AES designs.

density of states. These advantages outweigh the higher area overheads (i.e., from OLMs) due to transfer misalignment since weaker (smaller) and/or fewer cells can be used in the design while still meeting the performance target. For instance, to achieve the same target clock period of 600 ns in AES design synthesis, InGaAs/Ge (σ = 25 nm) requires only 1358 buffers and inverters, while the non-HGI design needs 2845 buffers and inverters. This is exemplified in **Error! Reference source not found.**(b), where the HGI designs can actually show chip area savings compared to the non-HGI case despite the higher penalties from transfer misalignment. The benefits of InGaAs and Ge are even more apparent for the ideal σ = 3 nm scenario, which represents the full potential of HGI technology.

In Fig. 11**Error! Reference source not found.**(c)–(f), the *post*-P&R delay and power tradeoffs are compared for the same benchmark designs with MAFL = 5 μm and 1 μm. The penalty arising from the gaps between adjacent sets of transferred fins generally leads to higher interconnect delay and power. Again, the intrinsic performance advantage from using InGaAs/Ge-based HGI overwhelms the overhead area penalties, leading to much better performance and power efficiency compared to the non-HGI design even for short MAFL. We note that in (c), (d), and (f), the σ = 3 and 25 nm HGI designs give very similar performance within a fixed-size die which suggests that the extra penalties to routing from the dummy filling cells in Fig. 10 are insignificant.

Finally, we explore the HGI design impact resulting from constraints on the maximum allowed fin length due to NTP challenges. We place designs synthesized with the same delay constraint in a fixed-size die with MAFL ranging from 1 to 20 μm, representing fin ARs of 120:1 to 2300:1 for 15nm FinFETs. For comparison, the experimentally demonstrated AR in Fig. 4 is 533:1. In Fig. 12, the total wire length decreases when longer fins can be successfully transferred. Additionally, the wire length drops quickly with incremental improvement in fin length for short fins, but then saturates for longer fins. The reduction in total wire length with longer MAFL is more apparent in MIPS compared to AES; this is because fewer cells in MIPS are connected by long metal lines, unlike AES. This also explains why the σ = 3 nm HGI designs showed more improvement compared to the σ = 25 nm designs in Fig. 11(e) but not in (c), (d), or (f): the short MAFL of 1 μm leads to a routing bottleneck in MIPS when there is significant transfer misalignment (i.e., σ = 25 nm) and hence leads to tempered performance gains. This illustrates how the transfer capabilities can have a stronger impact on designs which normally suffer from higher routing congestion. A maximum al-

lowed fin length of 5 μm or more (AR > 600:1) should not pose a bottleneck for HGI except for the densest designs.

For the materials and models considered, full InGaAs/Ge HGI shows the best characteristics, though naturally other material and design scenarios remain. While full HGI offers the most benefits in terms of performance, power, and area over non-HGI, the higher cost of implementing a two-step transfer process may pose a legitimate manufacturing concern. Our constant-leakage, constant-voltage results also do not consider the possibility of using HGI to scale supply voltage, which opens up more possibilities for performance optimization. While the quantitative results will change somewhat depending on chip architecture and utilization ratio, these results clearly illustrate the attractiveness of NTP-based HGI for near-future digital designs and provide motivation for the development of more sophisticated HGI design methods and models.

V. Conclusions

Our evaluation framework reveals that substantial improvements in circuit delay and power can be obtained using heterogeneous designs while trading off layout area. HGI cell area grows in response to more stringent design rules stemming from nanotransfer overlay misalignment, resulting in more frequent transistor folding and larger minimum cell widths. Fin trimming significantly reduces the lateral misalignment penalty and will likely be mandatory for HGI adoption. Designing strong cells with taller cell heights to reduce the folding frequency can also be beneficial when σ is large, despite the higher initial area overhead. Using InGaAs and Ge as heterogeneous materials to replace Si, sizeable reductions in processor delay (up to 40%-50%) and power (up to 15%-20%) are observed in HGI-based designs. Despite additional area overheads stemming from transfer misalignment, HGI designs actually consume less overall area compared to their non-HGI counterparts because some cells now require fewer fins than before to provide the same cell strength and designs will require fewer buffers to minimize critical path delays. Our findings[4] provide strong motivation for the process and design communities to pursue feature-level heterogeneous integration as a viable option for nanoscale semiconductor fabrication.

VI. References

[1] A. W. Topol, D.C. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM J. Res. & Dev*., vol. 50, no. 4, pp. 491-506, Jul. 2006.

[2] S. Seen, D. La Tulipe, A. W. Topol, D. J. Frank, K. Belote, and D. Posillico, "Overlay as the key to drive wafer scale 3D integration," *Microelec. Eng*., vol. 84, no. 5-8, pp. 1412-1415, May 2007.

[3] A. Gutierrez-Aitken, P. Chang-Chien, D. Scott, K. Hennig, E. Kaneshiro, P. Nam, N. Cohen, D. Ching, K. Thai, B. Oyama, J. Zhou, C. Geiger, B. Poust, M. Parlee, R. Sandhu, W. Phan, A. Oki, and R. Kagiwada, "Advanced heterogeneous integration of InP HBT and CMOS Si technologies," in *Proc. CSICS*, 2010, pp. 1-4.

[4] L. Czornomaz, N. Daix, K. Cheng, D. Caimi, C. Rossel, K. Lister, M. Sousa, and J. Fompeyrine, "Co-integration of InGaAs n- and SiGe p-MOSFETs into digital CMOS circuits using hybrid dual-channel ETXOI substrates," in *IEDM Tech. Dig*. 2013, pp. 52-55.

[4] The simulation input files and results of our work will be made freely available online for download.

[5] M. Yokoyama, S. Kim, R. Zhang, N. Taoka, Y. Urabe, T. Maeda, H. Takagi, T. Yasuda, H. Yamada, O. Ichikawa, N. Fukuhara, M. Hata, M. Sugiyama, Y. Nakano, M. Takenaka and S. Takagi, "III–V/Ge high mobility channel integration of InGaAs n-channel and Ge p-channel metal–oxide–semiconductor field-effect transistors with self-aligned ni-based metal source/drain using direct wafer bonding," *Appl. Phys. Express*, vol. 5, p. 076501, 2012.

[6] C. Chung, C. Chen, J. Lin, C. Wu, C. Chien, and G. Luo, "First experimental Ge CMOS FinFETs directly on SOI substrate," in *IEDM Tech. Dig*., 2012, pp. 383-386.

[7] M. J. H. van Dal, G. Vellianitis, G. Doornbos, B. Duriez, T. M. Shen, C. C. Wu, R. Oxland, K. Bhuwalka, M. Holland, T. L. Lee, C. Wann, C. H. Hsieh, B. H. Lee, K. M. Yin, Z. Q. Wu, M. Passlack, and C. H. Diaz, "Demonstration of scaled Ge p-channel FinFETs integrated on Si," in *IEDM Tech. Dig*., 2012, pp. 521-524.

[8] Z. Cheng, J.-S. Park, J. Bai, Z. Li, J. Hydrick, J. Fiorenza, and A. Lochtefeld, "Aspect ratio trapping heteroepitaxy for integration of germanium and compound semiconductors on silicon," in *Proc. ICSICT*, 2008, pp. 1425-1428.

[9] R. Loo, G. Wang, L. Souriau, J. C. Lin, S. Takeuchi, G. Brammertz, and M. Caymax, "High quality Ge virtual substrates on Si wafers with standard STI patterning," *J. Electrochem. Soc*., vol. 157, no. 1, pp. H13-H21, 2010.

[10] T. E. Kazior, R. Chelakara, W. Hoke, J. Bettencourt, T. Palacios, and H. S. Lee, "High performance mixed signal and RF circuits enabled by the direct monolithic heterogeneous integration of GaN HEMTs and Si CMOS on a silicon substrate," in *Proc. CSICS*, 2011, pp. 1-4.

[11] C. O. Chui, K-S. Shin, J. Kina, K.-H. Shih, P. Narayanan, and C. A. Moritz, "Heterogeneous integration of epitaxial nanostructures – Strategies and application drivers," in *Proc. SPIE*, 2012, p. 84670R.

[12] K-H. Shih, "III-V multigate non-planar channel transistor simulations and technologies," Ph.D dissertation, Dept. Elect. Eng., UCLA, Los Angeles, CA 2012.

[13] J. Nah, H. Fang, C. Wang, K. Takei, M. H. Lee, E. Plis, S. Krishna, and A. Javey, "III-V complementary metal-oxide-semiconductor electronics on silicon substrates," *Nano Lett*., vol. 12, no. 7, pp. 3592-3595, Jun. 2012.

[14] A. Carlson, A. M. Bowen, Y. Huangk, R. G. Nuzzo, and J. A. Rodgers, "Transfer printing techniques for materials assembly and micro/nanodevice fabrication," *Adv. Mater.*, vol. 24, pp. 5284-5318, 2012.

[15] Y. Sun, V. Kumar, I. Adesida, and J. A. Rogers, "Buckled and wavy ribbons of GaAs for high-performance electronics on elastomeric substrates," *Adv. Mater.*, vol. 18, pp. 2857-2862, 2006.

[16] H. Chen, X. Feng, Y. Huang, Y. Huang, J. A. Rogers, "Experiments and viscoelastic analysis of peel test with patterned strips for applications to transfer printing," *J. Mech. Phys. Solids*, vol. 61, pp. 1737-1752, Apr. 2013.

[17] H.-J. Kim-Lee, A. Carlson, D. S. Grierson, J. A. Rogers, and K. T. Turner, "Interface mechanics of adhesiveless microtransfer printing process," *J. Appl. Phys.*, vol. 115, p. 143513, Apr. 2014.

[18] H. Chen, X. Feng, and Y. Chen, "Directionally controlled transfer printing using micropatterned stamps," *Appl. Phys. Lett.*, vol. 103, p. 151607, Oct. 2013.

[19] D. R. Hines, S. Mezhenny, M. Breban, E. D. Williams, V. W. Ballarotto, G. Esen, A. Southard, and M. S. Fuhrer, "Nanotransfer printing of organic and carbon nanotube thin-film transistors on plastic substrates," *Appl. Phys. Lett.*, vol. 86, p. 163101, Apr. 2005.

[20] J.-H. Ahn, H.-S. Kim, K. J. Lee, S. Jeon, S. J. Kang, Y. Sun, R. G. Nuzzo, and J. A. Rogers," Heterogeneous three-dimensional electronics by use of printed semiconductor nanomaterials," *Science*, vol. 314, no. 5806, pp. 1754-1757, Dec. 2006.

[21] S. Wang, A. Pan., C. O. Chui, and P. Gupta, "PROCEED: A pareto optimization-based circuit-level evaluator for emerging devices," *Proc. ASP-DAC*, 2014, pp.818-824.

[22] S. Wang, A. Pan., C. O. Chui, and P. Gupta, "PROCEED: A pareto optimization-based circuit-level evaluator for emerging devices," *IEEE Trans. VLSI Systems*, accepted for publication 2015.

[23] S. J. Pearton, J. M. Kuo, W. S. Hobson, E. Hailemarian, F. Ren, A. Katz, and A. P. Perley, "Ion implantation doping of InGaP, InGaAs, and InAlAs," *Proc. MRS Symp*., 1992, pp. 797-804.

[24] K. S. Jones, and E. E. Haller, "Ion implantation of boron in germanium," *J. Appl. Phys*., vol. 61, no. 7, pp. 2469-2477, 1987

[25] R. T. P. Lee, Y. Oshawa, C. Huffman, Y. Trickett, G. Nakamura, C. Hatem, K.V. Rao, F. Khaja, R. Lin, K. Matthews, K. Dunn, A. Jensen, T. Karpowicz, P. F. Nielsen, E. Stinzianni, A. Cordes, P. Y. Hung, D-H.

Kim, R.J.W. Hill, W-Y. Loh, and C. Hobbs, Ultra low contact resistivity (< 1×10^{-8} Ω-cm^2) to $In_{0.53}Ga_{0.47}As$ fin sidewall (110)/(100) surfaces: Realized with a VLSI processed III-V fin TLM structure fabricated with III-V on Si substrates," *IEDM Tech. Digest*, 2014, pp. 776-779.

[26] J.J. Gu, X.W. Wang, H. Wu, J. Shao, A.T. Neal, M.J. Manfra, R.G. Gordon, and P.D. Ye, "20–80nm channel length InGaAs gate-all-around nanowire MOSFETs with EOT=1.2nm and lowest SS=63mV/dec," *IEDM Tech. Digest*, 2012, pp. 633-636.

[27] I. Ok, D. Veksler, P.Y. Hung, J. Oh, R. L. Moore, C. McDonough, R. E. Geer, C. K. Gaspe, M.B. Santos, G. Wong, P. Kirsch, H. H. Tseng, G. Bersuker, C. Hobbs, and R. Jammy, "Reducing R_{ext} in laser annealed enhancement-mode $In_{0.53}Ga_{0.47}As$ surface channel n-MOSFET," *Proc. VLSI-TSA*, 2010, pp. 38-39.

[28] C. O. Chui, H. Kim, D. Chi, B. B. Triplett, P. C. McIntyre, and K. C. Saraswat, "A sub-400°C germanium MOSFET technology with high-k dielectric and metal gate," *IEDM Tech. Dig.*, pp. 437-440, San Francisco, CA, December 8-11, 2002.

[29] J-P. Colinge, C-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A-M. Kelleher, B. McCarthy, and R. Murphy, "Nanowire transistors without junctions," *Nature Nanotechnol.*, vol. 5, pp. 225-229, 2010.

[30] G. Leung and C. O. Chui, "Variability of inversion-mode and junctionless FinFETs due to line edge roughness," *IEEE Electron Device Lett.*, vol. 32, no. 11, pp. 1489-1491, Nov. 2011.

[31] G. Leung and C. O. Chui, "Variability impact of random dopant fluctuation on nanoscale junctionless FinFETs," IEEE Electron Device Lett., vol. 33, no. 6, pp. 767–769, Jun. 2012.

[32] S. Wang, G. Leung, A. Pan, C. O. Chui, and P. Gupta, "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies," *IEEE Trans. Electron Devices*, vol. 60, no. 7, pp. 2186-2193, Jul. 2013.

[33] ASML TWINSCAN NXT:1950i Datasheet, ASML.

[34] Imprio 450, Molecular Imprints, http://www.molecularimprints.com/products/imprio450.php.

[35] M. Malloy and L. C. Litt, "Technology review and assessment of nanoimprint lithography for semiconductor and patterned media manufacturing," *J. Micro/Nanolith. MEMS MOEMS*, vol. 10, no. 3, p. 032001, 2011.

[36] B. J. Choi, M. J. Meissl, M. Colburn, T. C. Bailey, P. Ruchhoeft, S. V. Sreenivasan, F. Prins, S. K. Banerjee, J. G. Ekerdt, and C. G. Wilson, "Layer-to-layer alignment for step and flash imprint lithography," in *Proc. SPIE*, vol. 4343, 2001, pp. 436-442.

[37] M. Malloy and L. C. Litt, "Step and flash imprint lithography for semiconductor high volume manufacturing?" in *Proc. SPIE*, vol. 3676, 1999, pp. 379-389.

[38] N. Tas, T. Sonnenberg, H. Jansen, R. Legtenberg, and M. Elwenspoek, "Stiction in surface micromachining," J. Micromech. Microeng., vol. 6, pp. 385-397, 1996.

[39] K. Takei, R. Kapadia, H. Fang, E. Plis, S. Krishna, and A. Javey, "High quality interfaces of InAs-on-insulator field-effect transistors with ZrO_2 gate dielectrics," *Appl. Phys. Lett.*, vol. 102, no. 15, pp. 153513, 2013.

[40] S. Datta, *Quantum transport: atom to transistor*. Cambridge, U.K.; Cambridge University Press, 2005.

[41] A. Pan, G. Leung, and C. O. Chui, "Junctionless silicon and $In_{0.53}Ga_{0.47}As$ transistors—Part I: Nominal device evaluations with quantum simulations," *IEEE Trans. Electron Devices*, accepted for publication.

[42] International Technology Roadmap for Semiconductors, 2012 Edition [Online]. http://www.itrs.net/

[43] J. A. del Alamo, "Nanometre-scale electronics with III–V compound semiconductors," *Nature*, vol. 479, no. 7373, pp. 317-323, Nov. 2011.

[44] A. Rahman, G. Klimeck, and M. Lundstrom, "Novel channel materials for ballistic nanoscale MOSFETs-bandstructure effects," in *Proc. IEDM*, 2005, pp. 604-607.

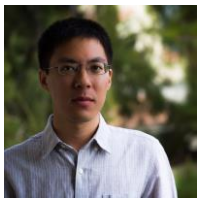
[45] NanoMOS, https://nanohub.org/resources/nanomos.

[46] M. De Michielis, D. Esseni, and F. Druissi, "Analytical models for the insight into the use of alternative channel materials in ballistic nano-MOSFETs," *IEEE Trans. Electron Dev.*, vol. 54, no. 1, pp. 115-123, Jan., 2007.

[47] R. S. Ghaida and P. Gupta, "A framework for early and systematic evaluation of design rules," in *Proc. ICCAD*, 2009, pp. 615-622.

[48] NanGate FreePDK45 Generic Open Cell Library [Online]. http://www.si2.org/openeda.si2.org/projects/nangatelib

[49] *MIPS/AES* [Online].http://opencores.org

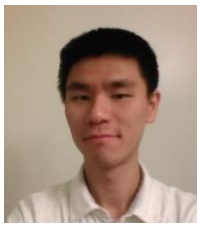


Greg Leung (S'10) received his B.S. degree in electrical engineering and computer science and materials science and engineering from the University of California at Berkeley, Berkeley CA, USA in 2008, and the M.S. degree in electrical engineering from the University of California at Los Angeles, Los Angeles, CA, USA in 2010 where he is currently pursuing the Ph.D. degree.

His current research interests include nanoscale CMOS devices, variability modeling, and heterogeneous integration technology.

Shaodi Wang received his B.S. degree from Peking University, China and the M.S degree in electrical engineering from UCLA. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, University of California, Los Angeles, CA, USA.

His current research interests include emerging memory and device technology and modeling for manufacturing.



Andrew Pan (S'12) received his B.S. degree in physics and M.S. degree in electrical engineering from the University of California, Los Angeles, where he is now pursuing a Ph.D. in electrical engineering.

His research interests include semiconductor device modeling, transport phenomena, and solid state physics.

Puneet Gupta is currently a faculty member of the Electrical Engineering Department at UCLA. He received the B.Tech degree in Electrical Engineering from Indian Institute of Technology, Delhi in 2000 and Ph.D. in 2007 from University of California, San Diego. He co-founded Blaze DFM Inc. (acquired by Tela Inc.) in 2004 and served as its product architect till 2007.

He has authored over 100 papers, 16 U.S. patents, a book and a book chapter. He is a recipient of NSF CAREER award, ACM/SIGDA Outstanding New Faculty Award, SRC Inventor Recognition Award and IBM Faculty Award. He currently leads the IMPACT+ center (http://impact.ee.ucla.edu) which focuses on future semiconductor technologies. Dr. Gupta's research has focused on building high-value bridges across application-architecture-implementation-fabrication interfaces for lowered cost and power, increased yield and improved predictability of integrated circuits and systems.

Chi On Chui (S'00–M'04–SM'08) received the B.Eng. degree in electronic engineering from the Hong Kong University of Science and Technology, Hong Kong, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2001 and 2004, respectively.

He joined Intel Corporation, Santa Clara, CA, USA, in 2004, as a Senior Device Engineer. During his tenure with Intel, he also served as a Researcher-in-Residence with the University of California at Berkeley, Berkeley, CA, USA, and Stanford University. From 2005 to 2006, he was also appointed as a Consulting Assistant Professor of Electrical Engineering with Stanford University. Since 2007, he has been with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, where he is currently an Associate Professor of Electrical Engineering and Bioengineering, and a member of the California NanoSystems Institute, UCLA. He has authored and co-authored over 130 peer-reviewed and invited archival journal and conference papers, and six book chapters. He holds nine issued patents. His current research interests include nanoelectronic device-circuit interaction, heterogeneous integration technology, and biomedical devices.

Dr. Chui has served on the Technical Program Committees and International Advisory Committees of several conferences on electronic devices and circuits. His work has received three Best Paper Awards. He received the Okawa Foundation Research Grant in 2007. He was the first recipient of the IEEE Electron Device Society Early Career Award in 2009, which is regarded as one of the society's highest honors. In 2011, he received the Chinese-American Faculty Association Robert T. Poe Faculty Development Award, the UCLA Faculty Career Development Award, and the University of California at San Diego's von Liebig Entrepreneurism Center Regional Health Care Innovation Challenge Award. He also received the UCLA Henry Samueli School of Engineering and Applied Science Northrop Grumman Excellence in Teaching Award in 2011.