Rapid and Holistic Technology Evaluation for Exploratory DTCO in Beyond 7nm Technologies

Myung-Hee Na
IBM Research
Albany, USA
myunghee@us.ibm.com

Albert Chu
IBM Research
Albany, USA
amchu@us.ibm.com

Yoo-Mi Lee
IBM Research
Albany, USA
yoomi@us.ibm.com

Albert Young
IBM Research
Albany, USA
yalbert@us.ibm.com

Vidhi Zalani
IBM Research
Albany, USA
Vidhi.Zalani@ibm.com

Hung H. Tran
IBM Research
Albany, USA
hhtran@us.ibm.com

Abstract—New device architectures such as horizontal Nanosheets have been seriously considered as a replacement for FinFET. A comprehensive, and realistic assessment of these architectures at the early stages of technology development is indispensable to understand their value propositions. In this study a new holistic technology-evaluation methodology for an early technology assessment is proposed. This methodology closely links performance-power metrics to realistic area scaling using block area assessment. This is especially critical for lower track cells since routing complexity can severely degrade performance. In addition, the optimization of an M1 power staple design combined with this evaluation can provide 12% additional area reduction with less than 1% of inverter performance penalty.

Keywords—CMOS, Horizontal Nanosheet, Parasitic, Technology evaluation, DTCO

I. INTRODUCTION

The introduction of FinFET technology to the 22nm node and beyond has enabled a superior short-channel effect using multi-gate structures [1]. However, as technology scaling goes to deep sub-7nm nodes, the gate control of a device continues to be a significant challenge for power and performance. Beyond 5nm nodes, the need of a new device architecture beyond FinFET has been widely acknowledged. Possible candidates include stacked horizontal Nanosheets, vertical nanowires, and 3-dimensional monolithic devices [2-5].

Regardless of the choice of device architectures in the ultimate technology adaptation, any new device architecture will very likely be disruptive in product designs once it becomes available as a technology offering. Therefore, to fully enable a new device architecture, an evaluation of the impact on product design at the beginning of the technology definition is crucial. Furthermore, an accurate value proposition for these kinds of new architectures has become increasingly difficult due to the complexity of process engineering, as well as the large number of permutations of unknown variables in a new definition. This is particularly true for technologies beyond 7nm. To mitigate this difficulty, device-focus TCAD has been widely used to understand potential benefits of new architectures in the early technology definition phase. While this is very important to study intrinsic device characteristics such as short-channel effects and intrinsic transport of new materials, it is not nearly sufficient to understand the technology value proposition in 10nm and beyond.

A holistic technology evaluation which includes parasitic effects as well as unique design concepts with innovative process solutions is crucial to demonstrate a true technology value proposition in the early stages of definition. In this paper, we review a variety of key considerations for advanced technology definitions. Moreover we propose a rapid and holistic technology evaluation methodology beyond 7nm technologies to predict realistic value propositions. In this study we focus on vertically stacked horizontal Nanosheets to demonstrate the proposed methodology.

II. EXPLORATORY DTCO CONCEPT &METHODOLOGY

In 14nm and beyond, design-technology co-optimization (DTCO) has played a critical role in the success of advanced technology development. As technology elements become more and more complex, it is essential to understand the design impact before they are introduced in products. Typically, DTCO has heavily focused on specific design styles, density and yield. Therefore, it has been natural for DTCO to influence design for manufacturing (DFM) methodology for product manufacturability.

In the early stages of technology development, the focus of DTCO should be broader than that of mature technology development. Consequently, DTCO should consider a diverse set of aspects of technology solutions not limited to performance, density, power, process complexity, but also extending to innovative product design concepts. One of the fundamental goals for exploratory DTCO is to define a new technology which can influence future products proactively. Therefore, it is critical to deploy an accurate methodology for area scaling and performance assessment relevant to product design styles.

978-1-5386-6790-3/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: Zhejiang University. Downloaded on December 24,2021 at 09:44:23 UTC from IEEE Xplore. Restrictions apply.

It has been speculated that the cell area estimate driven by standard cell track height is no longer sufficient to anticipate technology area scaling. This is because area scaling has become limited by routability issues in beyond 7nm technologies. It is therefore critical to consider routing impact in the early technology definition phase.

Fig. 1 describes the methodology used in this paper for holistic technology evaluations. There are three key components in this methodology: Patterning fidelity, performance, and area scaling. Exploratory patterning fidelity is focused on the assessment of key design arcs in standard cells and SRAM. A few examples of key design arcs are described in Fig. 2. As a result of patterning and process feasibility assessments, critical design rules and variability impact have been implemented in a Pathfinding production-design kit (PDK).

A Pathfinding PDK not only incorporates intrinsic device characteristics generated by device TCAD, but also includes accurate and flexible parasitic extraction capability. The accuracy of parasitic extraction comes from the direct link to the process assumption for a given technology, while flexibility of parasitic extraction is possible since the process change is strongly linked to parasitic RC extraction.

Area scaling, or density has been a strong driver for CMOS scaling. As discussed earlier, cell scaling alone is no longer sufficient to predict area scaling. In this study, we propose timing-aware block area scaling to consider the routing impact upfront for a realistic area scaling estimation.

III. Performance-Power Assessment

A. Horizontal Nanosheet : Width Optimization

This study focuses on vertically stacked horizontal Nanosheets, as shown in Fig. 3 (a)-(b) [3]. Table 1 shows the key parameters in a horizontal Nanosheet device used in this study. The performance trade-off between short-channel control and effective device width (W_{eff}) has been a consistent challenge for gate-all-around (GAA) structures. This is also valid for horizontal Nanosheets. A larger W_{eff} increases the drive strength of a Nanosheet device. However, horizontal Nanosheets that are too wide can degrade short-channel performance. Therefore, optimizing W_{eff} for a given track cell is important to maximize the device performance.

Parasitic capacitance can play an important role in the optimization of W_{eff}. At the same total W_{eff}, the performance of a single wide Nanosheet device is superior to that of multiple discrete narrow nanowires, as shown in Fig. 4. This results from the reduction of parasitic capacitance (Ceff), as the increase of edge capacitance from the sidewall of the narrow Nanosheet outweighs the benefit of current gain from the improved short channel effect.

B. Performance Impact: Parasitic R&C

MOL parasitic resistance and capacitance (R/C) play an important role in performance assessment due to tight gate-pitches and smaller contact areas. The impact of MOL R/C variations on performance can vary with design styles.

Figures 5 show the designs of 1X, 2X, and 4X inverters. Here, MOL parasitic resistance degrades the current as a function of the number of inverter fingers. The shared diffusion contacts between the fingers degrade the performance observed in 2X and 4X inverters. The 1X-finger design does not show this effect due to the lack of shared diffusion contacts. Figure 6 shows a 9% Ieff degradation and a ~70% increase of MOL parasitic resistance in 4X-finger devices relative to a 1X device. As the number of inverter fingers increases, the ratio of MOL resistance with respect to total resistance shows significant increase, as seen in Table 2.

MOL contacts have a significant impact on resistance, and consequently on device performance. A set of MOL resistance experiments was carried out with varying CA contact lengths as shown in Fig. 7(a). The result shows that the MOL resistance exponentially increases with decreasing CA as seen in Fig. 7(b). This increase will become more severe if the CA liner thickness increases to improve MOL reliability. As an example, we observed a ~30% degradation of MOL R with 1nm liner thickness increase for a square CA contact with 18nm width.

Therefore, the MOL parasitic R&C must be considered carefully, particularly in performance-optimized multi-finger designs. Figure 8 shows the unloaded inverter performance broken down into its Reff and Ceff components as a function of inverter fingers. Reff normalized by inverter W_{eff} increases as a function of fingers in inverter, while Ceff normalized by W_{eff} decreases. The data indicate that performance is at an optimum for a 2X inverter due to the shown Reff/Ceff trade-off in Fig. 8.

IV. Technology Definition for Perforamnce-Power-Area optimization

A. Realistic area scaling – Incorporating early assessment of block level area scaling

To achieve a high density chip design, tremendous focus has been put on smaller track height standard cells such as 6T and below. As the cell height is reduced in advanced technologies with complex metal design rules, metal routing becomes more challenging. This is particularly true for 6T and below track cells.

Figure 9 shows the comparison of cell area scaling and block level area scaling as a function of track heights. The results indicate that even when 6T can achieve ~67% of standard cell area scaling compared to 9T cells, the early assessment of block level scaling does not recognize the area scaling entitlement of 6T. Quite the opposite, the 6T block area becomes larger relative to 8T and 9T, even with a substantial increase of metal routing levels, shown in Fig. 10.

The inability to scale block areas in small track cells has compound effects on performance-power assessements. Reference 6 shows that performance at constant power can be comparable from 8.4T to 6T cells, while the performance at constant leakage can be lower for smaller track cells. This is driven by the fact that the parasitic capacitance in smaller track height cells can be reduced as the track height scales. However this is not accurate when the block area of small

Authorized licensed use limited to: Zhejiang University. Downloaded on December 24,2021 at 09:44:23 UTC from IEEE Xplore. Restrictions apply.

track height cannot be achieved. When the wire length cannot be scaled with the track height, 6T performance at constant power with realistic block area shows a severe degradation compared with the scaled area case of 6T, as shown in Fig. 11. In this study, the performance degradation is up to 10% at constant power, driven by the increase of wire parasitics with 6T larger block area.

B. *M1 power staples for small track height standard cells: Performance and block area scaling*

A small track height design below 7T suffers from the challenges of block level area scaling. In 6T case, the trend is even reversed. It has been widely recognized that new design concepts through design-process innovation during the early technology development is critical to address this issue.

M1 power staples can improve pin access and offer flexibility in the trade-off between power distribution and pin access [7]. However, M1 power staples have an impact on performance due to an increase in IR drop. Figure 12 from Ref. 6 shows that fewer M1 staples with a larger number of CPP per staple degrade the performance. The magnitude of the observed degradation depends on the device strength of the inverters.

As a consequence, the placement of M1 power staples should be carefully evaluated. We found that a 7T design with a 4:3 gear ratio between M1 metal pitch and critical gate pitch (CPP) can be effective when combined with one power staple per three M1 pins, as shown in Fig. 13. Such a design can enable block level area scaling by 12 %, as seen in Fig. 14. A 4X inverter using M1 staples with the gear ratio of 4:3 shows less than 1% inverter frequency degradation, while achieving ~12% area reduction. This can be a good trade-off between performance and area scaling which can be achieved by careful placement of M1 staples.

V. CONCLUSION

An enhanced holistic technology evaluation methodology beyond 7nm technologies is proposed, using vertically stacked horizontal Nanosheet devices. In this paper, we show that evaluating accurate parasitic R&C effects combined with realistic density scaling using product-like area scaling can influence performance-aware designs effectively. Specifically, MOL parasitic R&C effects play a significant role in performance-power optimized designs. We studied the performance degradation caused by MOL R due to shared diffusion contacts and MOL contact sizes in various inverter configurations.

Realistic area scaling using block level assessment becomes critical for small track standard cells below 7T due to routing challenges. The lack of block area scaling of 6T cells compared with cell level scaling can introduce an additional 10% of performance degradation due to increased wire parasitic effects. To enable smaller track scaling, innovative design concepts such as M1 power staples have been comprehensively investigated in various aspects including realistic area scaling and accurate performance impact.

This paper emphasizes the importance of holistic evaluation of performance, power and area for 7nm beyond nodes with realistic product-like metrics. This methodology should be considered in the early stage of technology definition to understand accurate value proposition of technology solutions.

Acknowledgment

The authors greatly appreciate IBM executive support to perform this work. We would also like to thank Synopsys for generous support in ICC2 licenses and valuable technical discussion. We are also grateful to Jens Haetty for providing insightful feedback for the paper.

REFERENCES

[1] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost et al, "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors", *Symp on VLSI Technology,* 2012, pp. 131-132.

[2] S.D. Kim, M. Guillorn, I. Lauer, P. Oldiges, T. Hook, M.H. Na, "Performance trade-offs in FinFET and gate-all-around device architectures for 7nm-node and beyond", *S3S Conference,* 2015, pp. 1-3.

[3] N. Loubet, T. Hook, P. Montanini, C. -W. Yeung, S. Kanakasabapathy, M. Guillom et al, "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET", *Symp on VLSI Technology,* 2017, pp. 230-231.

[4] M. Vinet, P.Batude, C.Fenouillet-Beranger, F.Clermidy, L.Brunet, O.Rozeau et al, "Monolithic 3D Integration: A powerful alternative to classical 2D scaling", *S3S Conference*, 2014, pp.1-3.

[5] M. Garcia Bardon, Y. Sherazi, P. Schuddinck, D. Jang, D. Yakimets, P. Debacker et al, "Extreme scaling enabled by 5 tracks cells: Holistic design-device co-optimization for FinFETs and lateral nanowires", *IEDM,* 2016, pp. 28.2.1-28.2.4.

[6] Y.M. Lee, M.H. Na, A. Chu, A. Young, T. Hook, L. Liebmann et al., "Accurate performance evaluation for the horizontal nanosheet standard-cell design space beyond 7nm technology", *IEDM,* 2017, pp. 29.3.1 - 29.3.4.

[7] L. Liebmann, V. Gerousisb, P. Gutwinb, X. Zhuc, and J. Petykiewiczc, "Exploiting Regularity: Breakthroughs in Sub-7nm Place-and-Route" *Proceeding SPIE, Design-Process-Technology Co-optimization for Manufacturability XI,* vol. 10148, May 2017.

Authorized licensed use limited to: Zhejiang University. Downloaded on December 24,2021 at 09:44:23 UTC from IEEE Xplore. Restrictions apply.

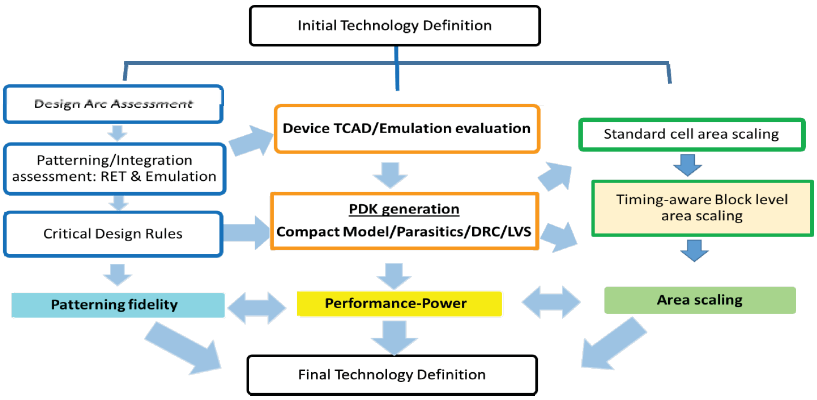

Fig. 1. Schematic diagram of exploratory DTCO flow used in this study. The critical components include patterning fidelity, performance-power and area scaling assessment.

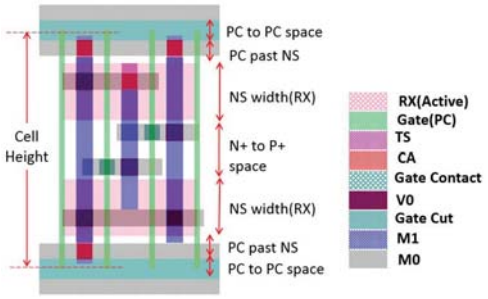

Fig. 2. Example of key design arcs in a vertically-stacked horizontal Nanosheet 7T standard cell design.

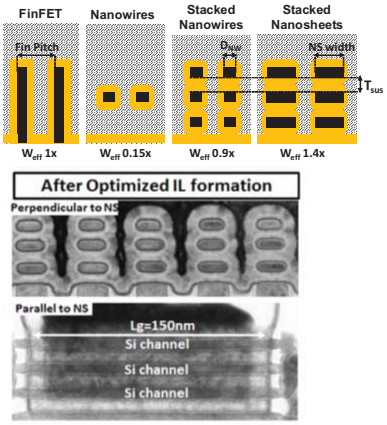

Fig. 3. Comparison of cross sections of FinFET and Nanosheet [3, 6]: (a) a schematic structure with key geometrical parameters of FinFET and vertically stacked horizontal Nanosheet and (b) cross sections of horizontal Nanosheet stack.

Table 1. Key technology parameters for sub-7nm horizontal Nanosheet [6] for this study. MOL contacts include TS/CA for diffusion and CB for gate contact. M0 is a local interconnect for both diffusion and gate contacts.

Device Focus	
Device Structures	Horizontal Nanosheet
# of Vertical sheet	3 Stacks
Total Weff (nm)	420 (8.4T)
	300 (7T)
	240 (6T)
Channel	unstrained Si N/P
MOL Contact	4 TS/CA/CB/M0
Epi-shape	Wrap around contact

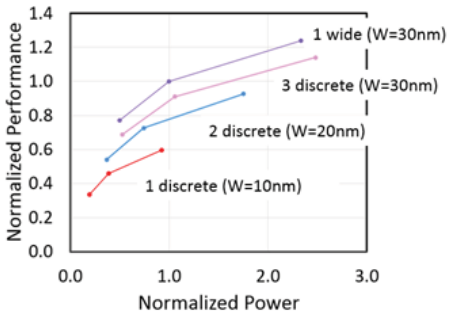

Fig. 4. Performance assessment of multiple narrow vs. one wide Nanosheet [6]. The power-performance of a single wide Nanosheet (W=30nm) is better than multiple discrete narrow Nanosheets (3 discrete Nanosheets with W=30nm).

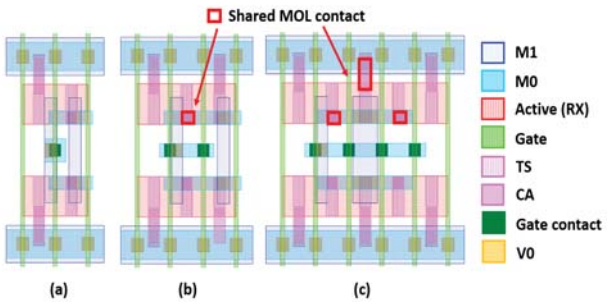

Fig. 5. Inverter designs for (a) 1X, (b), 2X and (c) 4X fingers. Shared CA contacts are shown in red. In (b), one square CA contact shared between two fingers. In (c), multiple shared diffusion contacts between the four fingers.

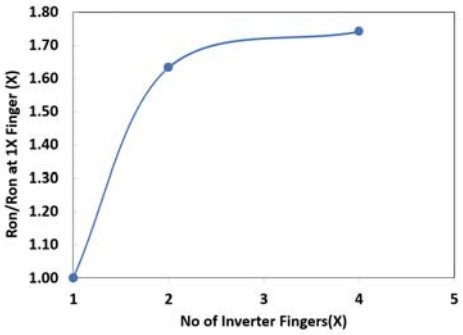

Fig. 6. Ron comparison relative to that of 1X finger for inverters. Ron is normalized by total Weff for each inverter. The normalized Ron is degraded as the number of inverter fingers increases due to parasitic resistances.

Authorized licensed use limited to: Zhejiang University. Downloaded on December 24,2021 at 09:44:23 UTC from IEEE Xplore. Restrictions apply.

Table 2. Breakdown of Ron with a variety of inverter fingers. Total Ron and each Ron components are compared with the reference (total Ron of 1X finger inverters). The parasitic resistance contributes more than 50% of total Ron for the 1X finger and goes up to 85% for 4X fingers.

Ron Ratio	Inverter Fingers		
	1X	2X	4X
Total	100%	134%	140%
Intrinsic Device	46%	46%	46%
MOL R	50%	81%	85%
MOL+V0/M1	54%	88%	93%

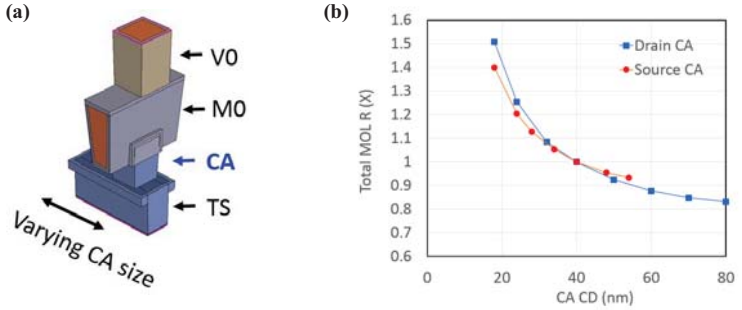

Fig. 7. (a) Schematic diagram of MOL contact of a horizontal Nanosheet. CA size is varied from 18nm to 80nm, while other MOL contact remains same. (b) Total MOL resistance as a function of CA CD. The MOL resistance is normalized by that of 40nm CA CD.

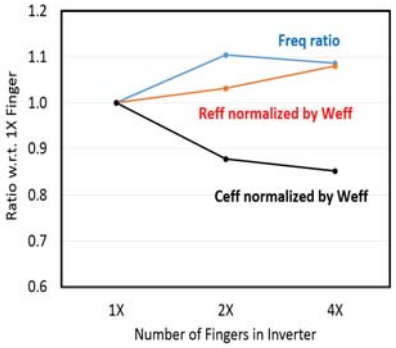

Fig. 8. Relative ratios of frequency, Reff and Ceff as a function of the number of fingers for unloaded inverters. Reff and Ceff are normalized by W_{eff} of each inverter.

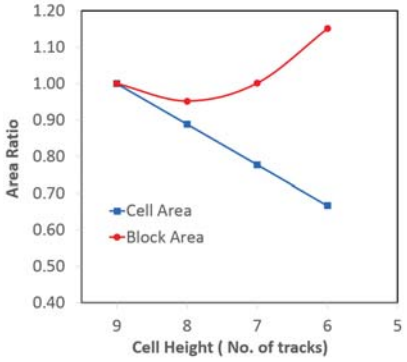

Fig. 9. Area scaling as a function of standard cell height, 9T, 7T and 6T. The comparison between cell area scaling and block area scaling is plotted.

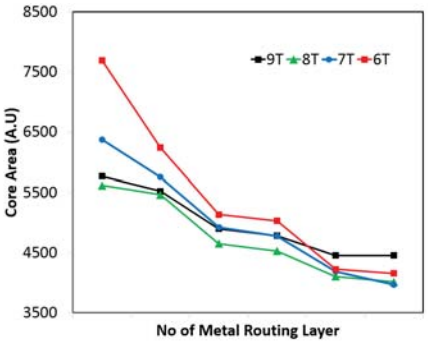

Fig. 10. Block core area reduction as the number of metal routing layers increases for various track height cells. Compared with 8T, the core areas for 7T and 6T do not improve even when the number of metal routing layer increases.

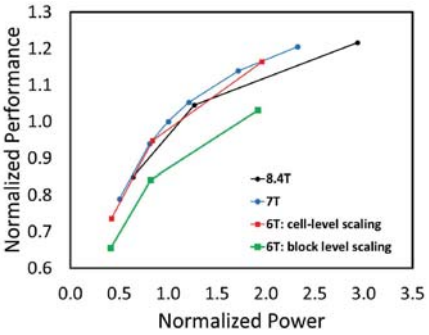

Fig. 11. Performance-power assessment for various track height cells. For 6T, the impact of wire parasitics on the performance is evaluated for both cases of cell area and block area scaling.

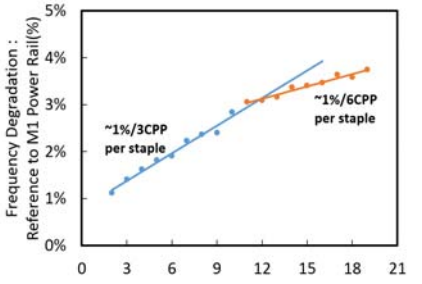

Fig. 12. 4X finger inverter performance degradation as a function of CPP spacing per M1 power staple [6]. The reference is the inverter frequency for the case of M1 power rail.

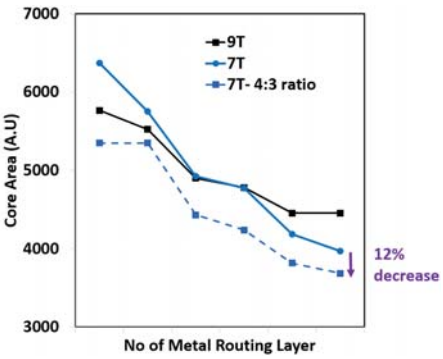

Fig. 14. The block core area reduction for 7T is plotted with 4:3 gear ratio of M1 power staples. The core area is reduced by ~12% with the optimized M1 power staple placement.

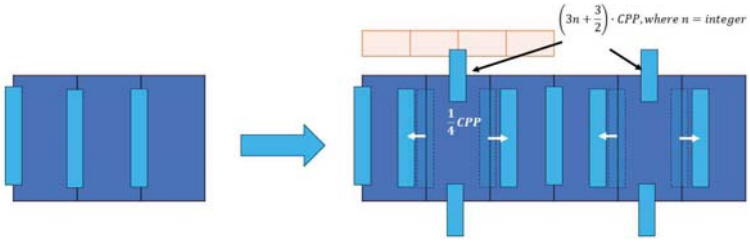

Fig. 13. Example of M1 power staple placement with gear ratio of 4:3. M1 staples are placed with 4CPP spacing. M1 pin access can also be optimized to align M1 grid.

Authorized licensed use limited to: Zhejiang University. Downloaded on December 24,2021 at 09:44:23 UTC from IEEE Xplore. Restrictions apply.