The impact of Sequential-3D integration on semiconductor scaling roadmap

A. Mallik[1], A. Vandooren[1], L. Witters[1], A. Walke[1], J. Franco[1], Y. Sherazi[1], P. Weckx[1], D. Yakimets[1], M. Bardon[1], B. Parvais[1,3], P. Debacker[1], B. W. Ku[2], S. K. Lim[2], A. Mocuta[1], D. Mocuta[1], J. Ryckaert[1], N. Collaert[1], P. Raghavan[1]
[1]imec, Leuven, Belgium, [2]School of ECE, Georgia Institute of Technology, Atlanta, USA, [3]Vrije Universiteit Brussel, Belgium
email: Arindam.Mallik@imec.be

***Abstract*—**The continued physical feature size scaling of CMOS transistors is experiencing asperities due to several factors (physical, technological, and economical), and it is expected to reach its boundary in the coming years. Sequential-3D (S3D) integration has been perceived as a promising alternative to continue the benefits offered by semiconductor scaling. This paper addresses the different variants of S3D integration and potential challenges to achieve a realizable solution. We analyze and quantify the benefits observed due to sequential scaling at a die level.

Keywords— Sequential 3D, Integration, Design Technology Cointegration, Cost of ownership

I. INTRODUCTION

The traditional path for improving the energy efficiency of digital systems through silicon CMOS scaling is becoming increasingly difficult (Fig. 1). For 7-nm node and beyond, we need to address both the familiar challenges of historical scaling and the new challenges associated with enabling such a technology with benefits in terms of Power, Performance, Area and Cost (PPAC) (Fig. 2). Sequential-3D (**S3D**) integration has been perceived as a technology enabler to alleviate problems in classical 2D CMOS. In this paper, we have analysed the feasibility of introducing different variations of S3D to reap the benefits of semiconductor scaling.

II. VARIATION OF S3D INTEGRATION

The granularity of S3D integration determines the complexity of the integration modules (Fig. 3). In **Transistor level S3D integration (T-S3D)**, the CMOS gate is split into tiers of pMOS and nMOS. This reduces the footprint of the unit cell to open a parallel path for dimensional scaling. As an alternative, we can place conventional 2D standard cells in different tiers. This is called **CMOS level Sequential-3D (C-S3D)** integration. At the other end of the spectrum, S3D can also be used for heterogeneous technology integration at an IP block level, termed as **Heterogeneous S3D (H-S3D)**. The primary objective of this paper is to analyze the impact of introduction of S3D integration in the semiconductor scaling roadmap in all 3 different variations. To the best of our knowledge, this is the first approach to evaluate the benefits of S3D integration at sub-7nm technology node.

III. INTEGRATION CHALLENGES FOR SEQUENTIAL 3D

A. Top layer implementation

A main challenge of S3D integration consists of creating a semiconductor layer on top of the processed bottom tier device and interconnects layers. Many approaches have been proposed in literature based either on polycrystalline material deposition and recrystallization or selective epitaxial overgrowth from a seed window. The approach based on polycrystalline material requires the use of ns laser anneal for recrystallization to meet thermal budget. On the other hand, selective epitaxial overgrowth suffers from limited density, high thermal budget, poor control of thickness and defective material close to the seed windows. Layer transfer based on wafer to wafer bonding is an attractive approach. It can be enabled at low temperature (400ºC) and leads to defect-free material with very tight control of thickness uniformity.

B. Thermal issues

The thermal budget for the top-tier processing needs to be low to avoid degradation of the bottom tier devices, metal interconnects, and the bonding interface. The low temperature processing may induce a mismatch of performance between two layers. Additionally, critical process steps such as selective epitaxial growth, and polysilicon deposition used in the RMG module needs to be reviewed for enabling low-temperature S3D integration.

1) Interconnect metals

Interconnect metals used for routing at the bottom layer need to fulfill two main constraints. They need to have a good thermal stability to sustain the top tier device-processing, and a low line resistance to maintain low RC delay. The standard back-end of line metal, Copper has low thermal stability (Fig.4). On the other end, Tungsten currently used for local interconnects features a higher resistivity than copper, but higher thermal stability up to 650ºC (Fig.5). An interesting candidate is cobalt, which provides a 2 to 3 times lower resistance than tungsten and a thermal stability as good as tungsten up to 650ºC (Fig.5). Cobalt also shows good filling properties in scaled inter-tier vias with 42nm CD and 1:5 aspect ratio (Fig.6).

2) Contacts

Contacts to devices using Ni silicide showed a limited thermal stability. F and W implantation to the PtNi silicide enables stable silicide up to 650°C [1]. On the other hand, direct contact using a Ti/TiN barrier shows thermal stability of the contact resistivity up to 550ºC 2h anneal for contacts to phosphorus-doped Si (nMOS) and up to 600ºC 1h for contacts to boron-doped SiGe (pMOS) (Fig.7).

3) Gate stack

In traditional Replacement Metal Gate flows, rapid high temperature anneals (‘reliability anneals’, t<2s, T>800ºC) are used after gate stack deposition to cure defects in the high-k dielectric and improve reliability. We have proposed a novel nMOS RMG gate stack suitable for S3D[3]. As shown in Fig. 8 (a), a thin $LaSiO_x$ layer is inserted between SiO_2 and HfO_2 to induce an interfacial dipole which modulates the effective work function (enabling the use of a thermally stable TiN metal gate for nMOS), and beneficially decoupling the high-k defect levels from

the channel Fermi level yielding sufficient PBTI reliability. Fig. 8 (b) summarizes the thermal stability of the gate stack electrical parameters. The parameters of the novel gate stack are within a target window (green shaded areas) both as deposited and after 60' long anneals at 450ºC and 520ºC.

C. S3D integration for finFET devices

S3D integration at imec features 300mm logic on logic finFET device stacking with a bottom tier device layer enabled with a 14nm bulk finFET technology. (Fig. 9).

IV. S3D INTEGRATION IMPACT ON AREA

The primary challenge in traditional scaling is the targeted 50% area shrink for both logic and SRAM bit cells. To enable that, Table 1 summarizes the lithography assumptions to transit from an iN7 node to iN5 node. Resorting to the T-S3D approach alleviates the problem of footprint shrink by stacking device/cells or IP blocks vertically. Fig. 10 demonstrates a representative GDS in accordance to iN7 design rules in both traditional2D and T-S3D technology. Top and Bottom tiers are connected to each other through three types of inter-tier vias (Fig. 11), TB_Via (Via through Oxide); TB_ViaDS (drain/source) and TB_ViaG (Gate-Gate). The area gains obtained due to T-S3D is reported in Table 2. The limited scaling of both logic and memory can be attributed to the track resource required for the TB_ViaG.

From an area scaling perspective, C-S3D is perceived as a better candidate. Previously published results have demonstrated 50% effective area reduction [2]. On the other hand, in the case of H-S3D, effective area reduction depends on the area of the largest component. In the current analysis, we evaluated a use-case of H-S3D where the logic and memory part is scaled to iN5 technology, whereas the remaining part (analog, and IO) is assumed to be manufactured in the N28 technology in the top tier.

V. PERFORMANCE LOSS/BENEFIT OF S3D SCALING

Fig. 12 shows the design flow for dealing with the top tier performance loss and using tungsten/higher resistance metallization in between the two tiers. The detailed design flow is explained in [2]. It can perform a partitioning such that that the circuits with more headroom for performance loss are placed on the top tier and the ones with critical performance critical are placed on the bottom tier.

VI. IMPACT ON POWER DUE TO S3D INTEGRATION

Fig. 13 shows the power breakdown of a 100K-gate digital block in iN7 ground rules after place and route of the full design. In such designs, ~50% of all capacitance is in the wiring between standard cells and the rest of the capacitance is inside the standard cells. Given that S3D allows to stack the 2D design into two layers of digital designs, this reduces the backend or wiring capacitance. Additionally, it indirectly reduces the front-end capacitance as the number buffers needed to drive the wires reduce.

Table 3 shows the power benefits of C-S3D with respect to a 2D implementation. There are two C-S3D implementations that are shown. The ideal C-S3D has no process constraints and the top tier device is as good as the bottom tier device and the metallization in between the two front ends is in Copper. The last column refers to a worst-case scenario where the metallization between the tiers is Tungsten and the top tier has about 20% lower drive compared to the bottom tier. Based on the capacitance breakdown, the C-S3D gets about 15% reduction in power.

VII. COST OF IMPLEMENTATION

The acceptance of a technology innovation by the semiconductor industry is heavily dependent on the manufacturing cost. We looked on the S3D impact on cost based on imec cost modeling framework [4]. Fig. 14 records the wafer-level cost of the 3 different variants of the S3D integration against the traditional iN7 and iN5 node. The BEOL cost is the major cost contributor in advanced technology node. As seen the figure, the increased FEOL cost of the T-S3D gets balanced against the increased iN5 BEOL cost. For the C-S3D as the FEOL along with 3Mx layers gets repeated twice, the overall wafer cost in considerably higher than baseline iN7 wafer cost. The wafer cost of the H-S3D is highest as due to the full iN5 stack as the bottom tier with a N28 top tier with 6 193 Single-patterned metal layer.

Fig. 15 shows the increased mask count for S3D integration which have a direct impact on the cycle time and yield. However, the corresponding Process complexity factor (K-value) does not increase significantly due to simpler 193-ArF or I-Line masks required for the S3D integration options. It rather eliminates the EUVL double-patterning required in the iN5 case.

But the final decision on acceptability of manufacturing cost is determined at a die level. For a representative 125 mm^2 SoC die (Fig 16), we can see even with 80% increase in wafer cost, it is possible to have 33% reduction in die cost for H-S3D (Fig. 17). The primary reasons for not obtaining any die cost reduction in the T-S3D and C-S3D are two-fold, (a) 30% non-scalable part being present and (b) the limited scaling of logic and memory.

Fig 18. reports the sensitivity of the die size on the cost. With 30% non-scalable part, H-S3D performs the best by removing the non-scalable part to a different tier. On the other hand, Fig. 19 reports the impact of reducing the non-scalable part to 10%. We observe C-S3D manages to drop the die cost up to 10% of iN7. On the other hand, the smaller non-scalable part hampers the overall benefit of H-S3D.

VIII. CONCLUSION

We have analyzed impact of S3D integration on the semiconductor scaling roadmap. From an integration perspective, major challenges and significant progress made in interconnect materials, gate-stack have been demonstrated. Successful stacking of finFET devices show a promising path for S3D scaling. An early PPAC analysis for different mode of S3D integration have been performed. We have demonstrated that the high processing cost of S3D integration can be recovered by an efficient scaling (up to 50%) at a die level. Based on current assumptions, H-S3D performs the best by integrating a low-cost non-scalable tier on top of an expensive iN5 tier for scalable logic and memory. Results show that the relative benefit of the approach is heavily dependent on the component distribution and size of the application hardware.

REFERENCES

[1] P. Batude et. al, "Advances, challenges and opportunities in 3D CMOS sequential integration", IEDM 2011

[2] BW Ku, et. al, "Physical Design Solutions to Tackle FEOL/BEOL Degradation in Gate-level Monolithic 3D ICs", ISLPED, 2016

[3] J. Franco, et. al, "Gate Stack Thermal Stability and PBTI Reliability Challenges for 3D Sequential Integration: Demonstration of a suitable gate stack for top and bottom tier nMOS", IRPS 2016

[4] A. Mallik, et. al, "Maintaining Moore's law: enabling cost-friendly dimensional scaling, SPIE AL, EUVL, February 2015

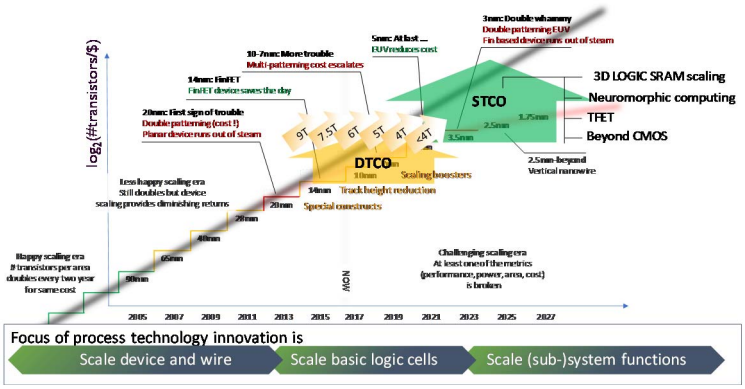


Fig. 1. Scaling Roadmap

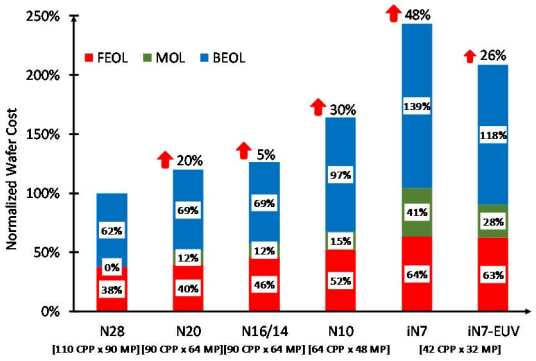


Fig. 2. Wafer cost evolution across technology nodes

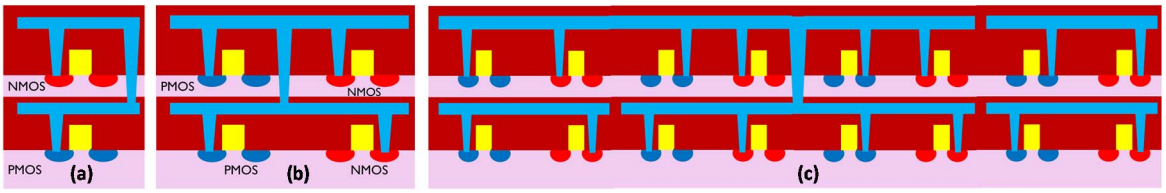


Fig. 3. (a) T-S3D (b) C-S3D; (c) H-S3D architecture

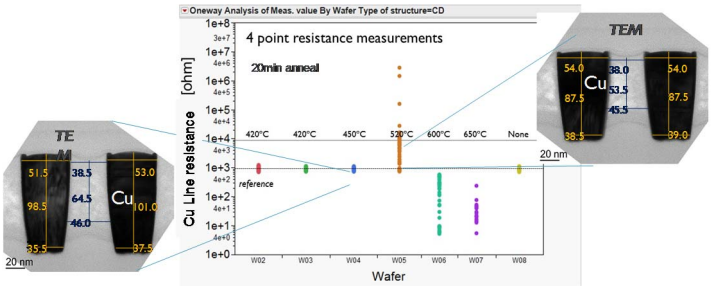


Fig. 4. Copper line resistance variation with thermal budget. First degradation is seen for an anneal at 520ºC for 20 minutes

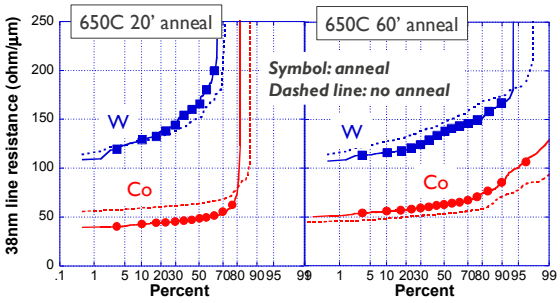


Fig. 5. W and Co line resistance before and after 2 different anneals at 650ºC 20' and 60'

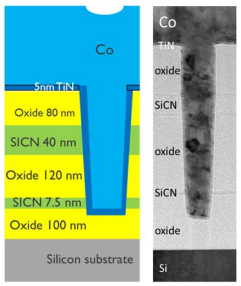


Fig. 6. Co filling of inter-tier vias with CD of ~42nm and an aspect ratio of 1: 5

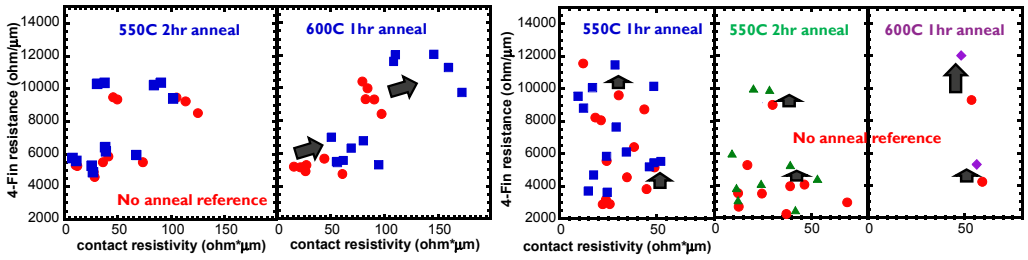


Fig. 7. Fin resistance and contact resistivity for (a) nMOS (Si:P) and (b) pMOS (SiGe:B) devices before and after different anneals

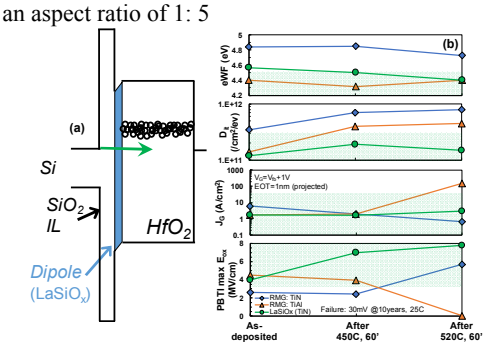


Fig. 8. (a) novel nMOS RMG gate stack suitable for S3D; (b) Thermal stability of the gate stack electrical parameters (effective work function, interface state density, gate leakage density, and PBTI max operating oxide electric field), compared to RMG gate stacks with TiN or TiAl metal gates

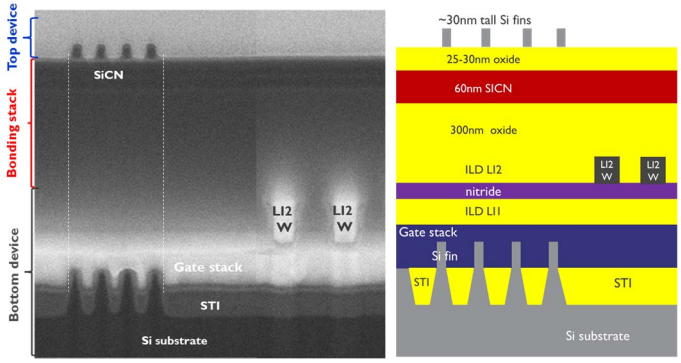


Fig. 9. finFET on finFET device stacking enabled on 300mm wafers featuring wafer to wafer bonding and 14nm finFET technology

Layer Name		iN7-193i		iN7-EUVL		iN5 (36x21)		
	Pitch (nm)	193i	# of masks	193i	EUVL	Pitch (nm)	193i	EUVL
Fin	24	SAQP	1	SAQP		24	SAQP	
Fin Keep+CUT	48	LE2 + LE2	4			48		CUT + Keep
Gate	42	SADP	1	SADP		36	SAQP	
Gate CUT	90	LE3	3		EUV SE	72		EUV SE
M0A	42	LE4	4			36	SRAM Cut	EUV LE2
VINT-A	52	LE4	4		EUV SE	36		EUV LE2
VINT-G	90	LE2	2	LE2		36		EUV SE
MINT	32	SAQP	1	SAQP		21	SAQP	SAQP
MINTCUT	52	LE3	3		EUV SE	42		EUV LE2 (SAB)
V_0	50	LE4	4		EUV SE	42		EUV LE2
M1	42	SADP	1	SADP		21	SAQP	
M1CUT	52	LE3	3		EUV SE	42		EUV LE2(SAB)
V_1	50	LE3	3		EUV SE	30		EUV LE2
Vx	45	LE4	4		EUV SE	30		EUV LE2
Mx	32	SAQP	1	SAQP		21	SAQP	
Mx CUT	32	LE4	4		EUV SE	42		EUV LE2(SAB)

Table 1. Process assumption table for iN7, iN5

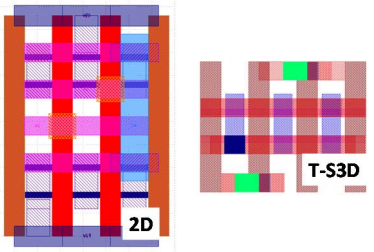


Fig. 10. ND2 logic cell layout in 2D and T-S3D option

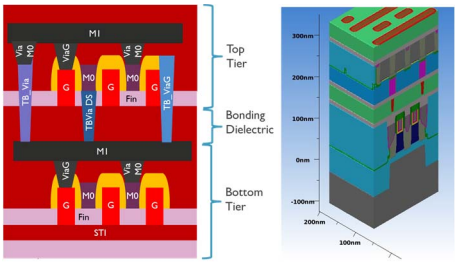


Fig. 11. Cross sectional COVENTOR images for T-S3D SRAM bit cell

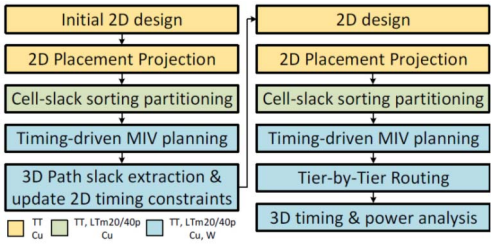


Fig. 12. Framework for C-S3D benefit evaluation

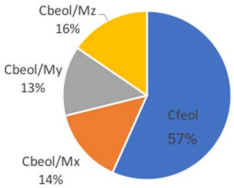


Fig. 13. Capacitance breakdown in iN7 of a digital IP block (~100K gates)

	iN7-2D	iN5-2D		T-S3D	
			Area gain		Area gain
Logic cells (ND2, ND3, OAI21, XNR, DFF)	6 TR	6 TR	42%	5 TR	17%
SRAM bit cell	6.375 TR		36%	5.5 TR	13.7%

Table 2. Area scaling due to T-S3D

	2D	Seq3D ideal	Seq3D worst case
Metal stack	5 Cu layers	5 Cu layers + 5 Cu layers	5 Cu layers + 3 W layers
Top tier performance	-	Same as bottom tier	-20% Ion wrt bottom tier
Wire cap (pF)	113.64	69.08	83.22
FEOL cap (pF)	61.96	56.84	56.71
Total power (mW)	39.28	29.88	32.31

Table 3. Impact of C-S3D on system power

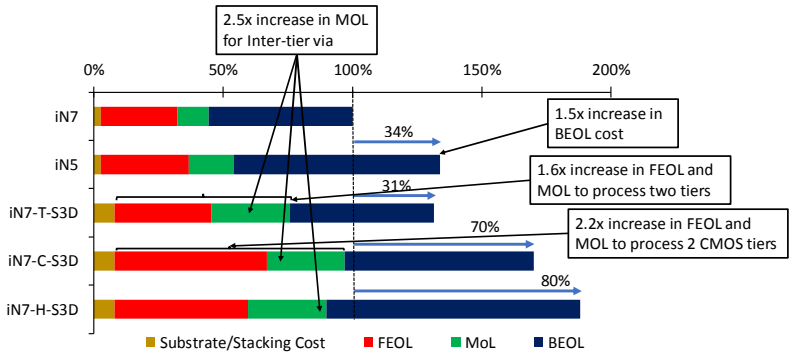


Fig. 14. Wafer cost evaluation for S3D integration

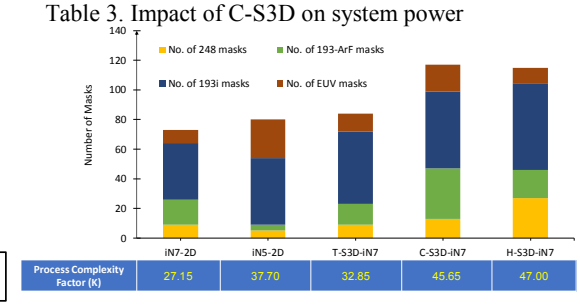


Fig. 15. Mask count and process complexity factor (K) for S3D alternatives. K is a determining factor for die-yield based on Bose-Einstein formulation

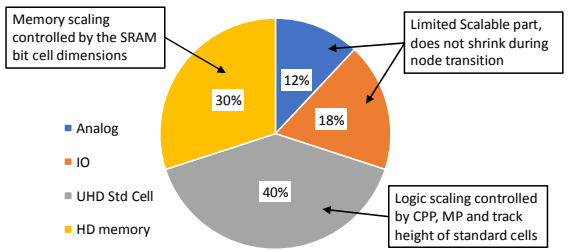
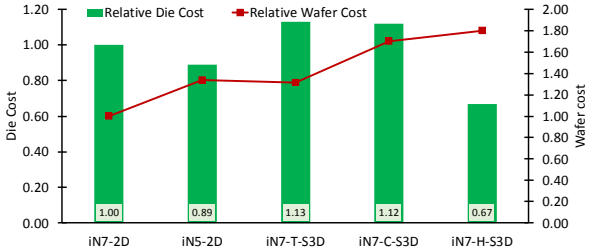


Fig. 16. Component distribution in a SoC die



Die Size(mm²)	125	85.75	107.5	81.875	52
Die Yield	77%	78%	76%	75%	83%

Fig. 17. Die cost for S3D variations

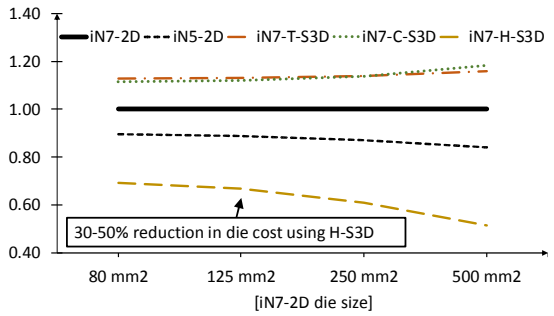


Fig. 18. Die cost with 30% non-scalable component

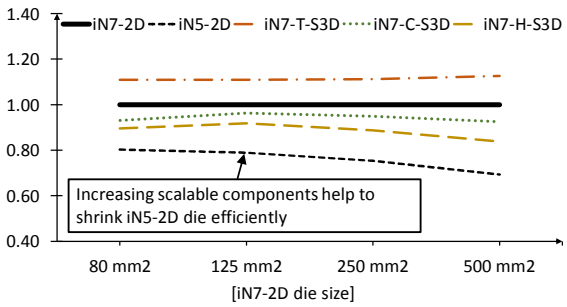


Fig. 19. Die cost with 10% non-scalable component