Monolithic 3D Integrated Circuits

Deepak C. Sekar, Brian Cronquist,

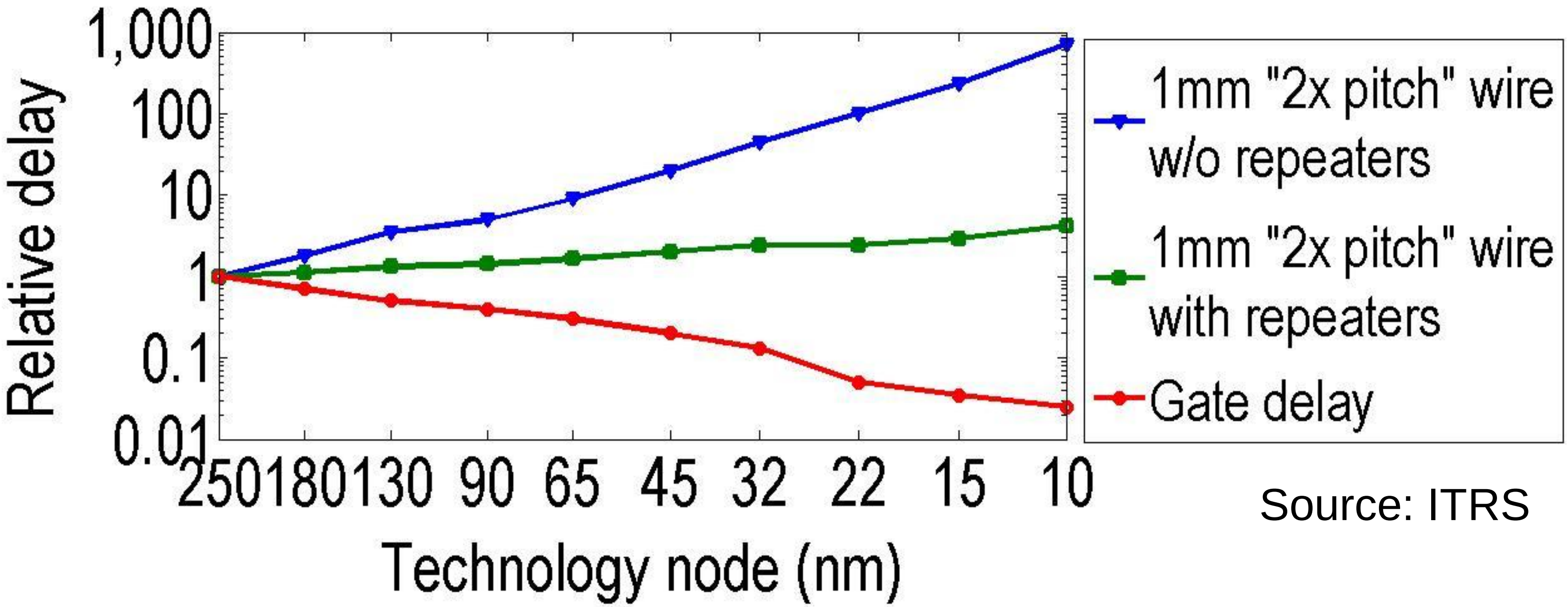
Israel Beinglass, Paul Lim, and Zvi Or-Bach

MonolithIC 3D Inc.

Outline

- Introduction
- Paths to Monolithic 3D
- IntSim v2.0: A 2D/3D-IC Simulator
- Conclusions

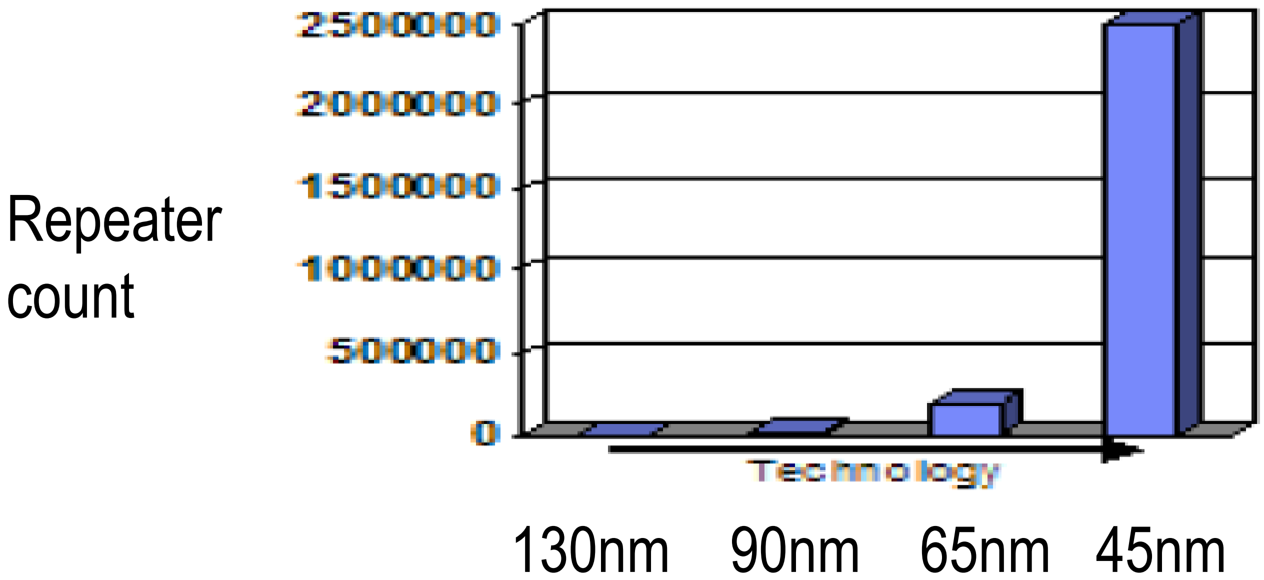
The Interconnect Problem



Source: ITRS

- Transistors improve with scaling, interconnects do not
- Even with repeaters, 1mm wire delay ~50x gate delay at 22nm node

The repeater solution consumes power and area…



Source: IBM POWER processors
R. Puri, et al., SRC Interconnect Forum, 2006

- Repeater count increases exponentially with scaling
- At 45nm, repeaters >50% of total leakage power of chip [IBM]
- Future chip power, area could be dominated by interconnect repeaters [IBM] [P. Saxena, et al. (Intel), IEEE J. for CAD of Circuits and Systems, 2004]

We have a serious interconnect problem

What's the solution?

FRIDAY, FEBRUARY 12, 1960 Irvine Auditorium—9:00 A.M.-12:00 Noon

SESSION VII: Microelectronic Considerations

7.2: Speed, Power and Component Density in Multielement High-Speed Logic Systems

J. M. Early

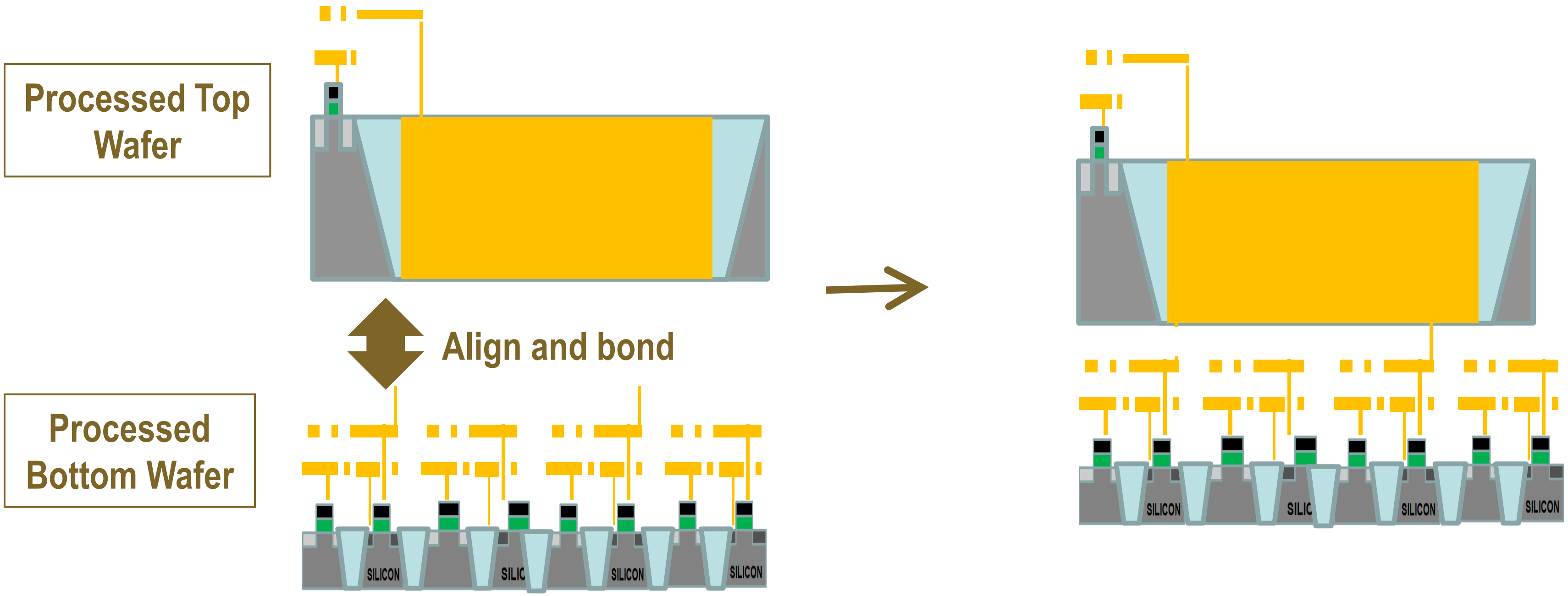
Bell Telephone Laboratories, Inc.

Murray Hill, N. J.

Arrange components in the form of a 3D cube → short wires

James Early, ISSCC 1960

3D with TSV Technology



- TSV size typically >1um: Limited by alignment accuracy and silicon thickness

Industry Roadmap for 3D with TSV Technology

Intermediate Level, W2W 3D-stacking	*2009-2012*	*2013-2015*
Minimum TSV diameter	1-2 µm	0.8-1.5µm
Minimum TSV pitch	2-4 µm	1.6-3.0 µm
Minimum TSVdepth	6-10 µm	6-10 µm
Maximum TSV aspect ratio	5:1 – 10:1	10:1 – 20:1
Bonding overlay accuracy	1.0-1.5 µm	0.5-1.0 µm
Minimum contact pitch	2-3 µm	2-3 µm
Number of tiers	2-3	8-16 (DRAM)

ITRS 2010

- **TSV size ~ 1um, on-chip wire size ~ 20nm → 50x diameter ratio, 2500x area ratio!!!**
  Cannot move many wires to the 3rd dimension
- ***TSV: Good for stacking DRAM atop processors, but doesn't help on-chip wires much***

Can we get Monolithic 3D?

Requires sub-50nm vertical and horizontal connections

Focus of this talk…

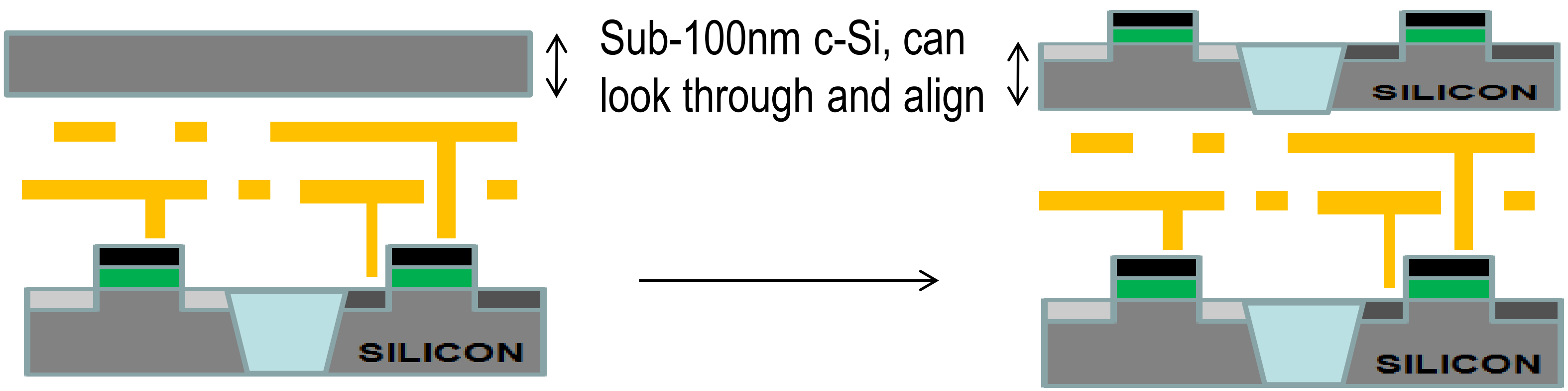
The Monolithic 3D Challenge

- A process on top of copper interconnect should not exceed 400^{o}C
  - How to bring mono-crystallized silicon on top at less than 400^{o}C
  - How to fabricate advanced **transistors** below 400^{o}C
- Misalignment of pre-processed wafer to wafer bonding step is ~1μ
  - How to achieve 100nm or better connection pitch
  - How to fabricate a thin enough layer for inter-layer vias of ~50nm

Outline

- **Paths to Monolithic 3D**

Getting sub-50nm vertical connections

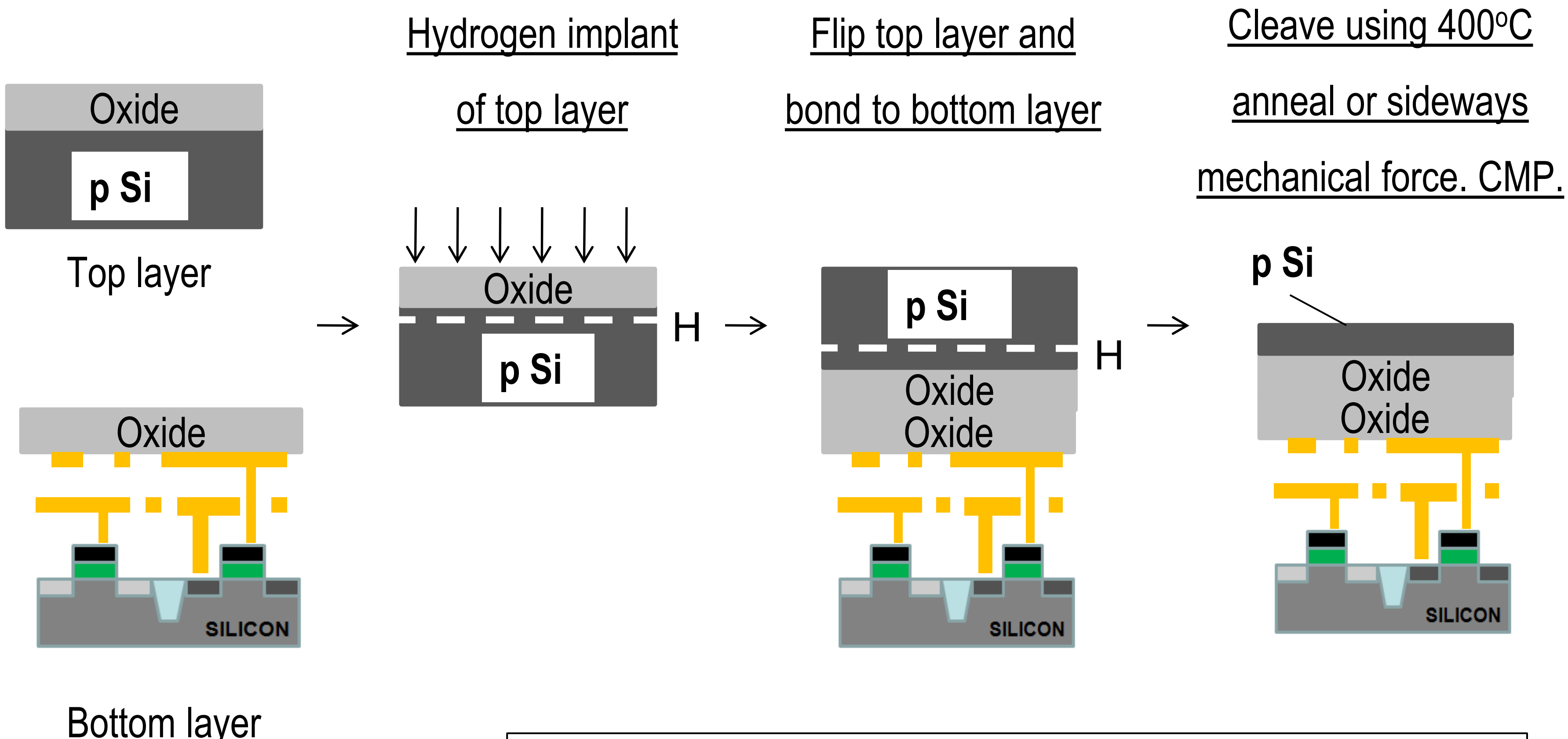


- Build transistors with c-Si films above copper/low k

  → Avoids alignment issues of bonding pre-fabricated wafers

- Need <400-450°C for transistor fabrication → no damage to copper/low k

Layer Transfer Technology (or "Smart-Cut")
→ Defect-free c-Si films formed @ <400°C



Similar process (bulk-to-bulk) used for manufacturing all SOI wafers today

Sub-400^{o}C Transistors

Transistor part	Process	Temperature
Crystalline Si for 3D layer	Bonding, layer-transfer	Sub-400^{o}C
Gate oxide	ALD high k	Sub-400^{o}C
Metal gate	ALD	Sub-400^{o}C
Junctions	Implant, RTA for activation	>400^{o}C

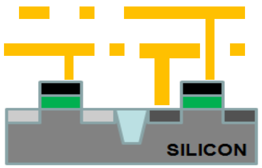
Junction Activation: Key barrier to getting sub-400^{o}C transistors

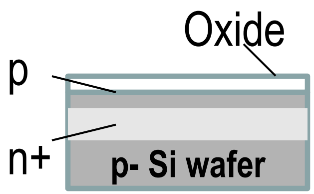
In next few slides, will show 2 solutions to this problem… both under development.
For other techniques to get 3D-compatible transistors, check out www.monolithic3d.com

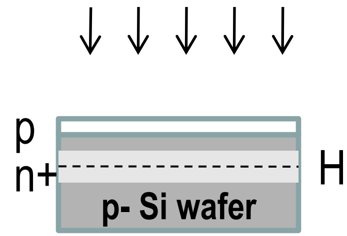
One path to solving the dopant activation problem: Recessed Channel Transistors with Activation before Layer Transfer

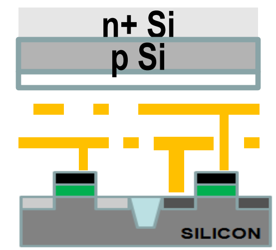
Idea 1: Do high temp. steps (eg. Activate) before layer transfer

Layer transfer



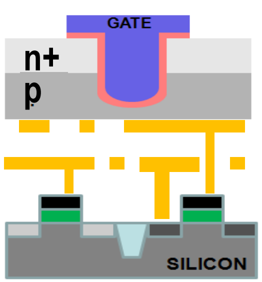


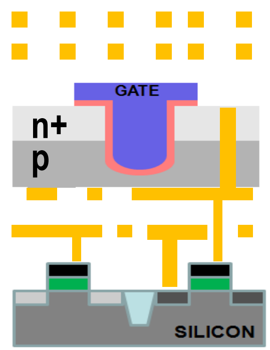




Idea 2: Use low-temp. processes like etch and deposition to define (novel) recessed channel transistors

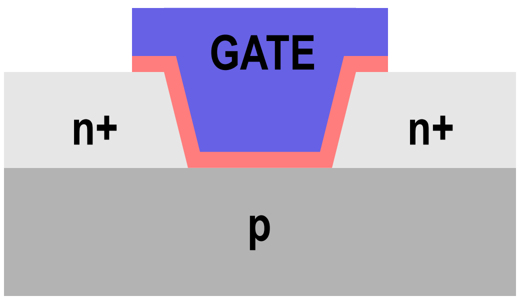
Idea 3: Silicon layer very thin (<100nm), so transparent, can align perfectly to features on bottom wafer



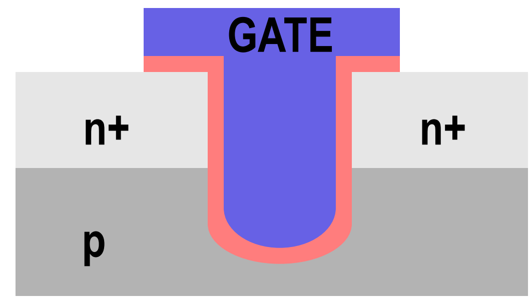


Note:
All steps after Next Layer attached to Previous Layer are @ < 400°C

Recessed channel transistors used in manufacturing today → easier adoption



V-groove recessed channel transistor:
Used in the **TFT industry** today



RCAT recessed channel transistor:
- Used in **DRAM production** @ 90nm, 60nm, 50nm nodes
- Longer channel length → low leakage, at same footprint

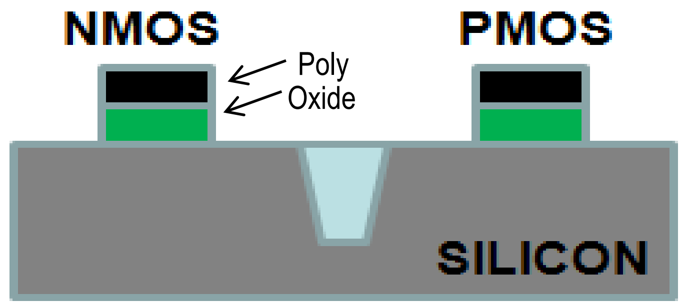
J. Kim, et al. Samsung, VLSI 2003
ITRS

Monolithic 3D with State of the Art Transistors

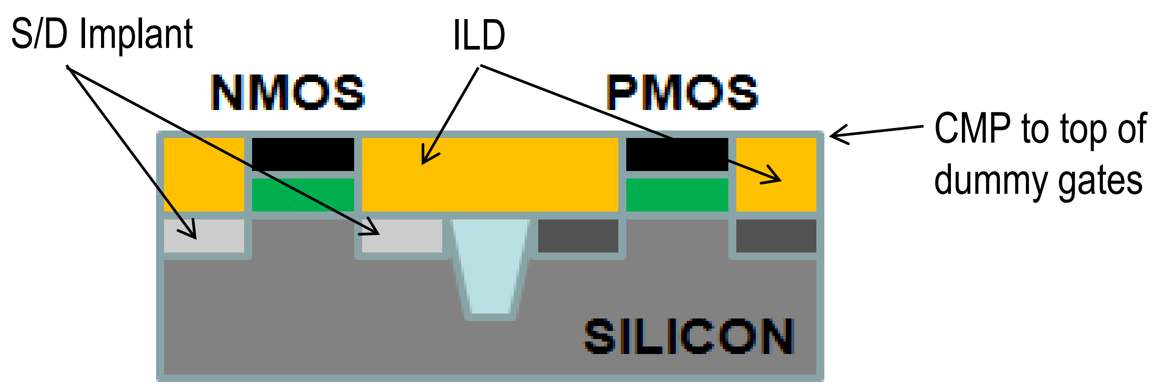
- Uses a novel combination of four ideas
  - Gate-Last Process and proper sequence of “Ion-Cut”
  - Low Temperature Face-up Layer Transfer
  - Repeating Layouts
  - Innovative Alignment

A Gate-Last Process for Cleave and Layer Transfer

Step 1 (**std**): On donor wafer, fabricate standard dummy gates with oxide, poly-Si

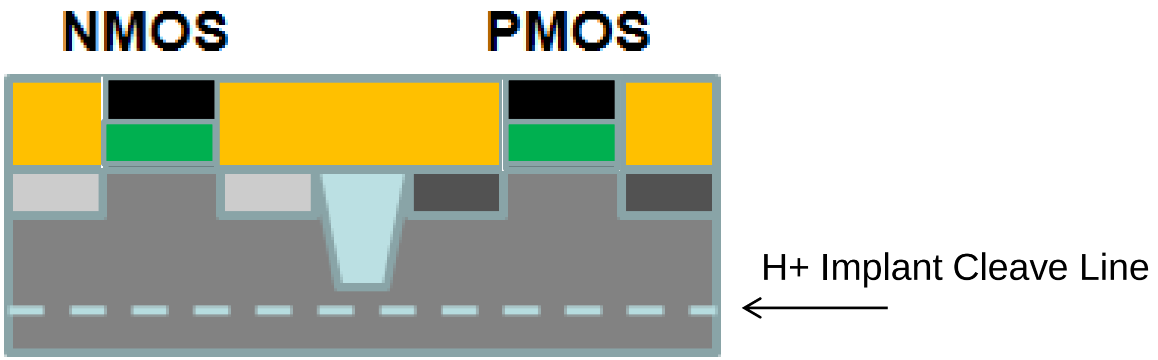


Step 2 (**std**): Std Gate-Last
- Self-aligned S/D implants
- Self-aligned SiGe S/D
- High-temp anneal
- Salicide/contact etch stop or faceted S/D
- Deposit and polish ILD

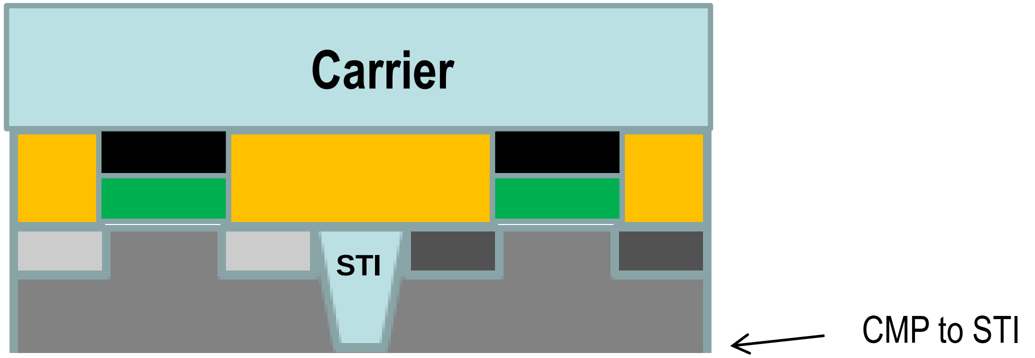


A Gate-Last Process for Cleave and Layer Transfer

Step 3.
Implant H for cleaving

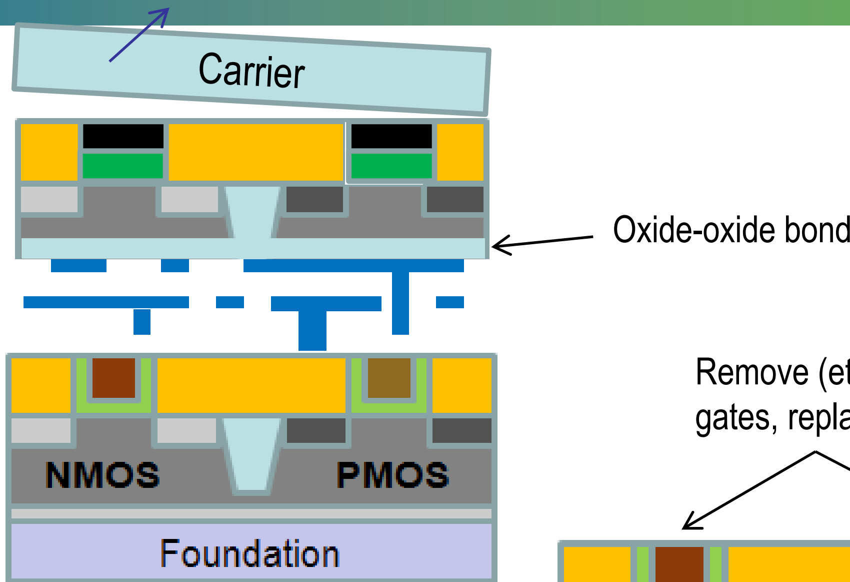


Step 4.
- Bond to temporary carrier wafer (adhesive or oxide-to-oxide)
- Cleave along cut line
- CMP to STI

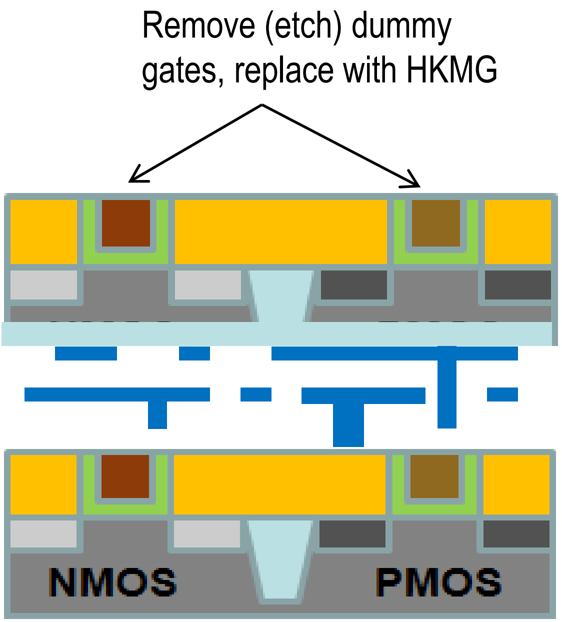


A Gate-Last Process for Cleave and Layer Transfer

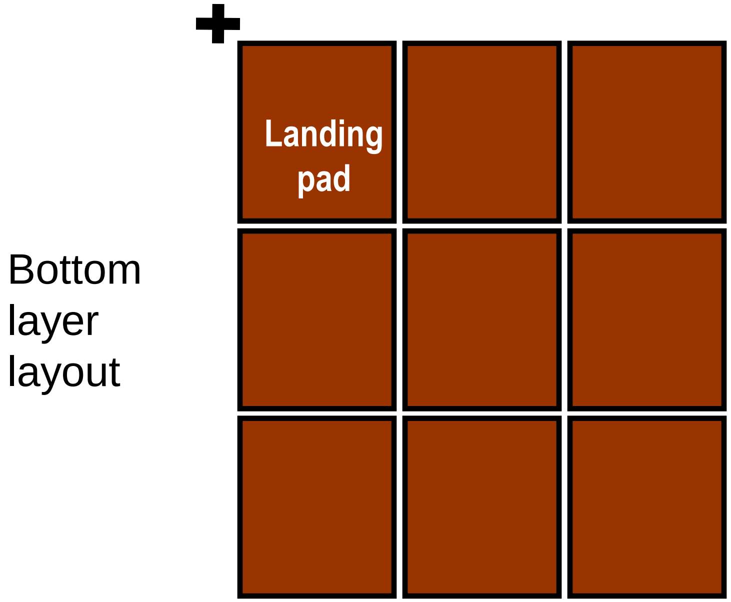
Step 5.
- Low-temp oxide deposition
- Bond to bottom layer
- Remove carrier

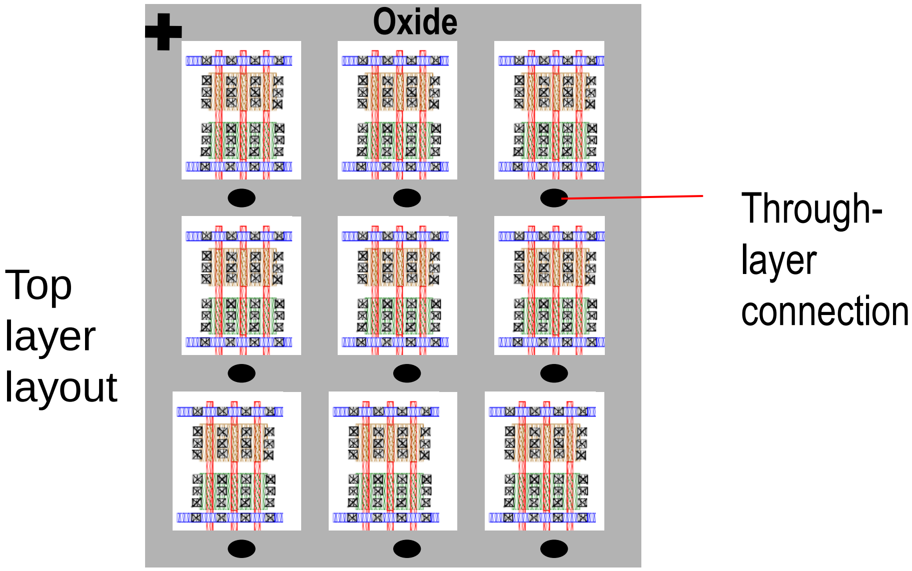


Step 6 (**std**): On transferred layer:
- Etch dummy gates
- Deposit gate dielectric and electrode
- CMP
- Etch tier-to-tier vias thru STI
- Fabricate BEOL interconnect



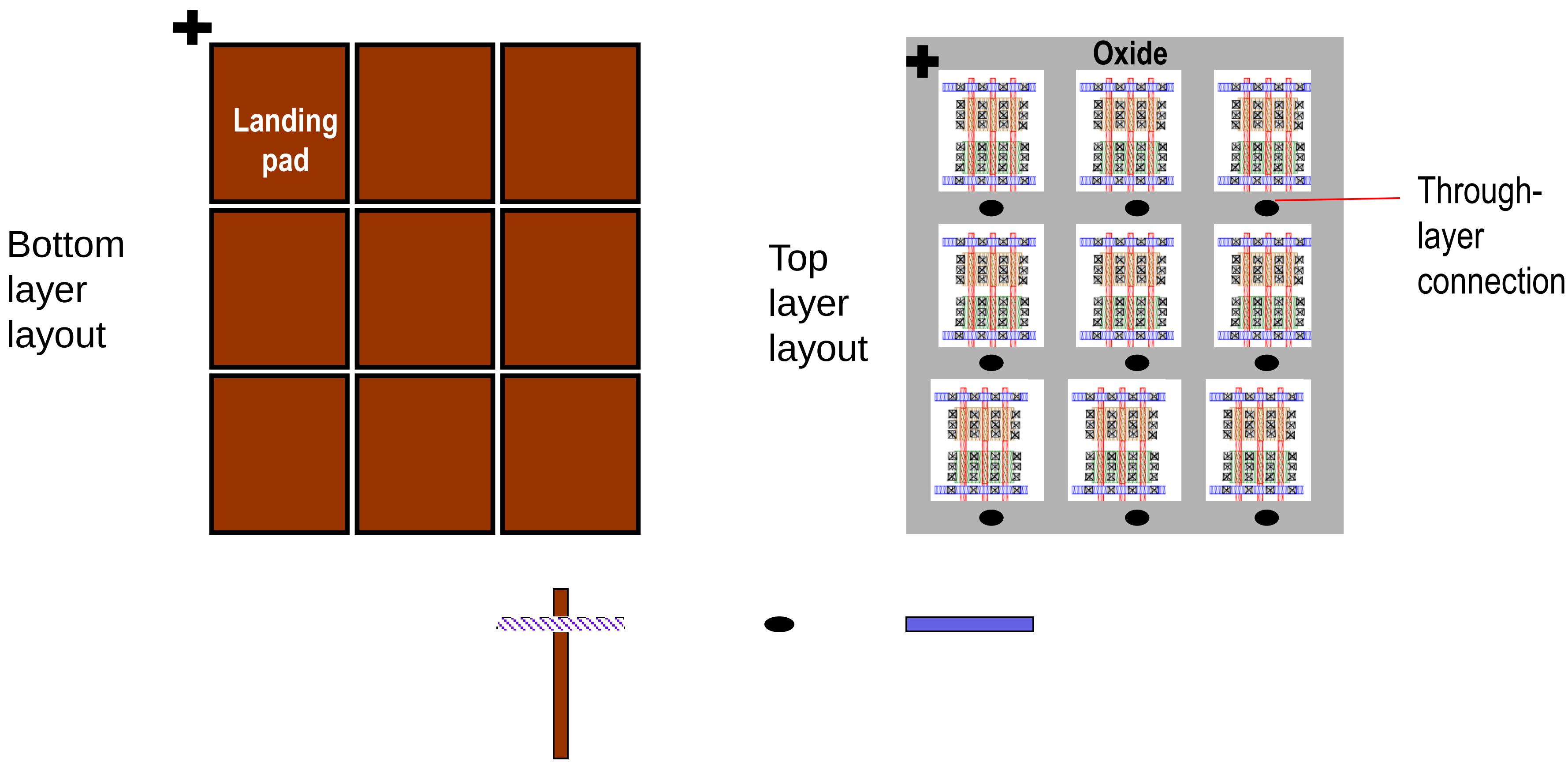
Novel Alignment Scheme using Repeating Layouts





- Even if misalignment occurs during bonding → repeating layouts allow correct connections.
- Above representation simplistic (high area penalty).

A More Sophisticated Alignment Scheme



Outline

- **<u>IntSim v2.0</u>: A 2D/3D-IC Simulator**

IntSim: A CAD Tool Simulator for 2D or 3D-ICs
[D. C. Sekar, J. D. Meindl, et al., ICCAD 2007]

Inputs

- Gate count
- Die area
- Frequency
- Rent's parameters
- Number of strata

  (1 if 2D, >=2 for 3D)

IntSim v2.0

Contains models for

- Stochastic signal interconnect prediction for 2D and 3D-ICs
- Via blockage
- Logic gates
- Chip power
- Power distribution
- Clocks
- Heat removal
- Energy-optimized repeater insertion

Iterative top-level algorithm used to handle dependencies between models

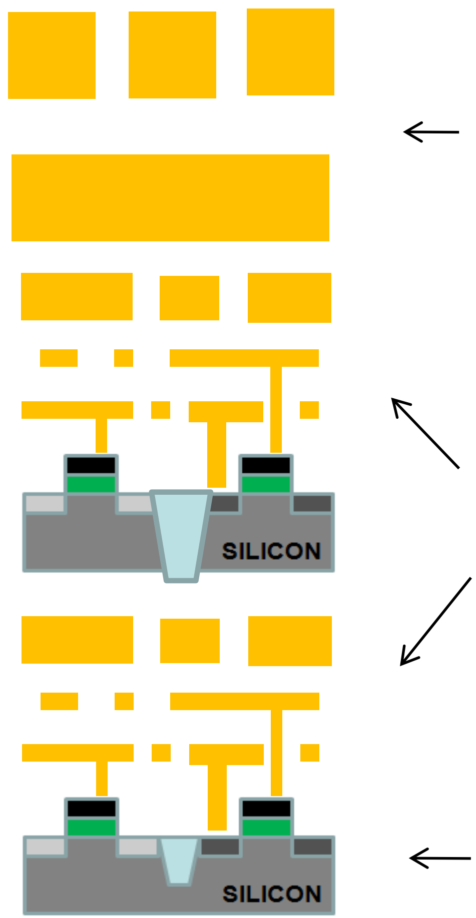
Outputs

- Chip power
- Metal level count
- Wire pitches of different metal levels

Open-source tool, available for use at www.monolithic3d.com

IntSim v1.0: Built at Georgia Tech in Prof. James Meindl's group (by Deepak Sekar, now @ MonolithIC 3D)
IntSim v2.0: Extended IntSim v1.0 to monolithic 3D using 3D wire length distribution models in the literature

IntSim v2.0: Uses a novel algorithm to combine many models



Global interconnect levels

Shared among all strata

Model → [D. C. Sekar, J. D. Meindl, et al., IITC 2006]

Local and semi-global interconnect levels

Each stratum has its own

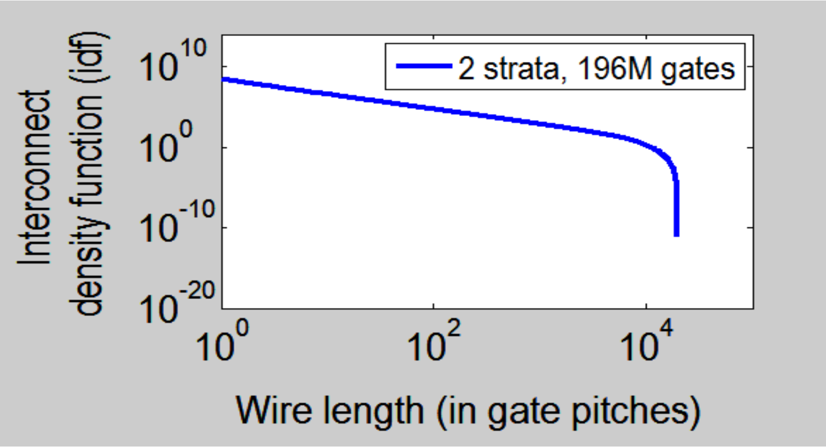
Models → [PhD dissertations of A. Rahman (MIT), R. Venkatesan, D. Sekar, J. Davis, R. Sarvari (all Georgia Tech students in Prof. Jim Meindl's group)]

Logic gates

Critical path model developed by K. Bowman (Georgia Tech)

Stochastic Signal Wire Length Distribution Model

Number of wires of length l = Function(Number of gates, die size, strata, feature size, Rent's constants)



Number of wires of length between l and l+dl = idf(l) dl

- **Models from J. Davis, A. Rahman, J. Meindl, R. Reif, et al.**
  [A. Rahman, PhD Thesis, MIT 2001] [J. Davis, PhD Thesis, Georgia Tech, 1999]
- **2D model → fits experimental data reasonably well [J. Davis, PhD Thesis, GT, 1999]**
  3D model → same methodology

Compare 2D and 3D-IC versions of the same logic core with IntSim

22nm node 600MHz logic core	2D-IC	3D-IC 2 Device Layers	Comments
Eff. Metal Levels	10	10	
Average Wire Length	6um	3.1um	
Av. Gate Size	6 W/L	3 W/L	Since less wire cap. to drive
Die Size (active silicon area)	$50mm^2$	$24mm^2$	3D-IC → Shorter wires → smaller gates → lower die area → wires even shorter **3D-IC footprint = $12mm^2$**
Power	Logic = 0.21W	Logic = 0.1W	Due to smaller Gate Size
	Reps. = 0.17W	Reps. = 0.04W	Due to shorter wires
	Wires = 0.87W	Wires = 0.44W	Due to shorter wires
	Clock = 0.33W	Clock = 0.19W	Due to less wire cap. to drive
	Total = 1.6W	**Total = 0.8W**	

3D with 2 device layers → 2x power reduction, ~2x active silicon area reduction vs. 2D

Scaling with 3D or conventional 0.7x scaling?

Analysis with IntSim v2.0 Same logic core scaled	2D-IC @22nm	2D-IC @ 15nm	3D-IC 2 Device Layers @ 22nm
Frequency	600MHz	600MHz	600MHz
Eff. Metal Levels	10	12	10
Footprint	50mm^2	25mm^2	12mm^2
Total Silicon Area (a.k.a "Die size")	**50mm^2**	**25mm^2**	**24mm^2**
Average Wire Length	6um	4.2um	3.1um
Av. Gate Size	6 W/L	4 W/L	3 W/L
Power	**1.6W**	**0.7W**	**0.8W**

- 3D can give you similar benefits vis-à-vis a generation of scaling for a logic core!
- Without the need for costly lithography upgrades!!!
- Let's understand this better…

Theory: 2D Scaling vs. 3D Scaling

	2D Scaling (0.7x Dennard scaling)		Monolithic 3D Scaling (2 device layers)
	Ideal	Today, V_{dd} scales slower	
Chip Footprint	2x reduction		*2x-4x reduction*
Long wire length $\alpha\sqrt{\text{Footprint}}$	0.7x reduction		0.7x-2x reduction
Long wire capacitance	0.7x reduction		0.7x-2x reduction
Long wire resistance	>0.7x *increase*		0.7x-2x *reduction*
Gate Capacitance	*0.7x reduction*		*Same*
Driver (Gate) Resistance (Vdd/Idsat)	Same	*Increases*	*Same*

Overall benefits seen with IntSim have basis in theory

- **2D scaling scores: Gate capacitance**
- **3D scaling scores: Wire resistance, driver resistance, wire capacitance**

Outline

- Conclusions

Conclusions

- Monolithic 3D Technology possible and practical:

  - Recessed Channel Transistor

  - SOA gate-last HKMG transistor

- IntSim v2.0, a CAD tool to simulate 2D and 3D-ICs

  - Useful for architecture exploration, technology predictions and teaching

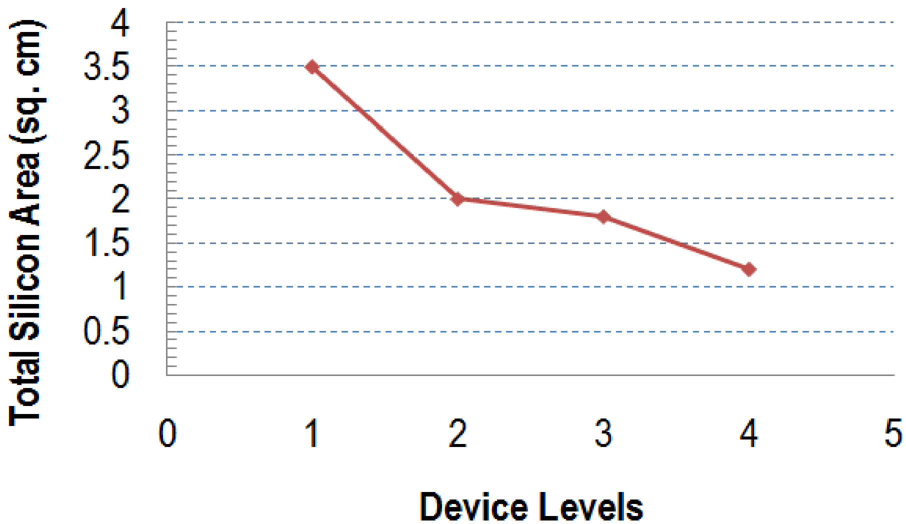
  - Open source tool, anyone can contribute!

- 3D scaling

  → Benefits similar to a generation of feature size scaling (2D), but without costly litho upgrades or expensive R&D

Backup slides

Technical Literature:
[J. Davis, J. Meindl, K. Saraswat, R. Reif, et al., Proc. IEEE, 2001]



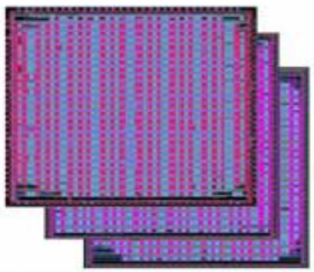
Simulation study:
Frequency = 450MHz, 180nm node
ASIC-like chip

Tremendous benefits when vertical connectivity ~ horizontal connectivity.

3x reduction in total silicon area + 12x reduction in footprint

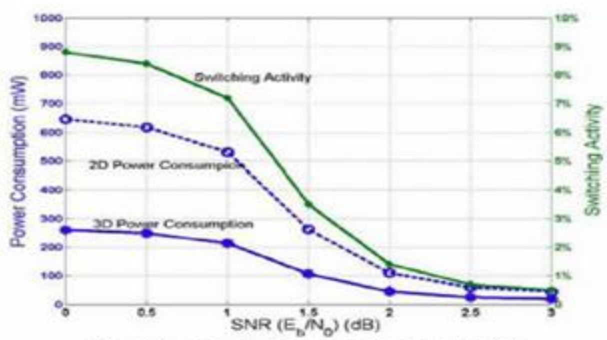
vs. a 2D implementation, even @ 180nm node

Technical Literature:
[L. Zhou, R. Shi, et al, Proc. ICCD 2007]



Final layout view of 3D LDPC structure.

"Implementing a 2-Gbs 1024-bit ½-rate Low-Density Parity-Check Code Decoder in Three-Dimensional Integrated Circuits"
Lili Zhou, Cherry Wakayama, Robin Panda, Nuttorn Jangkrajarng, Bo Hu, and C.-J. Richard Shi
University of Washington
International Conference on Computer Design, ICCD, Oct. 2007



Post-layout power of the LDPC decoder (2D vs 3D).

Comparison between 3D and 2D designs

	2D design	3D design
Area (mm*mm)	18.238*15.92 =**290.35**	(6.4*6.227)*3 = **119.56**
Total wire length (m)	**182.42**	22.39+22.57+22.46 =**67.42**
Max WL before buffer insertion (mm)	**13.82**	**8.68**
Max WL after buffer insertion (mm)	**4**	**4**
Buffer used	**32900**	**24636**
Clock skew (ns)	**2.33**	**1**
Power dissipation (mw)	**646.2**	**260.2**

Performance Factor (Area * Timing * Power) = 14

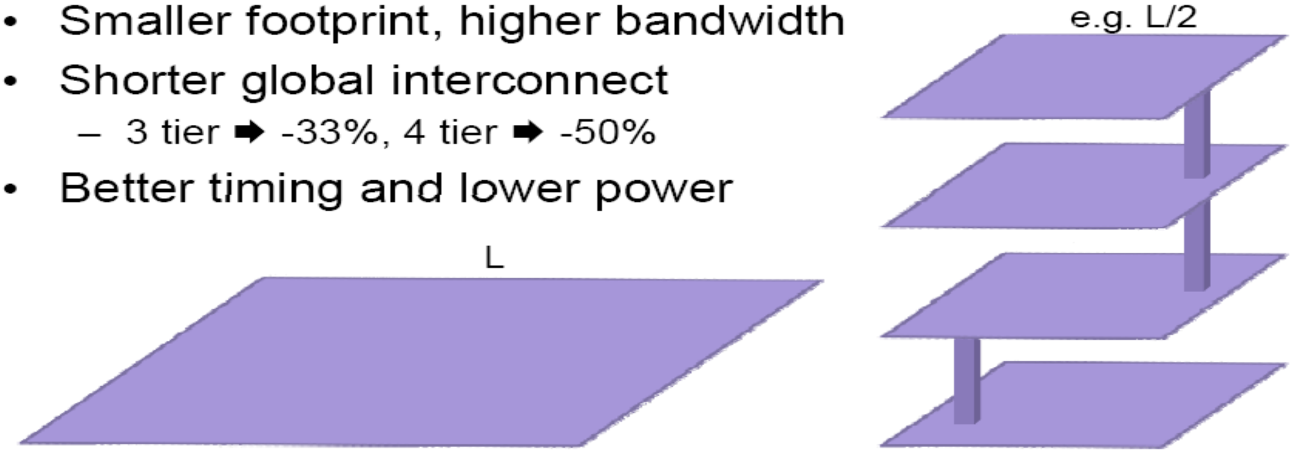
Did layout of 2D and 3D-ICs, and showed more than 10x benefit

Technical Literature:
Synopsys @ RTI 3D Workshop, Dec. 2010

"3D" IC Integration Looks Great…

Technology Node n^{th} 2D $\cong$ Technology Node $(n-2)^{th}$ 3D

- Much easier D and A&M/S integration
- Smaller footprint, higher bandwidth
- Shorter global interconnect
  - 3 tier ➡ -33%, 4 tier ➡ -50%
- Better timing and lower power



Silicon area = L^2, Footprint = L^2
Corner to corner distance = 2L

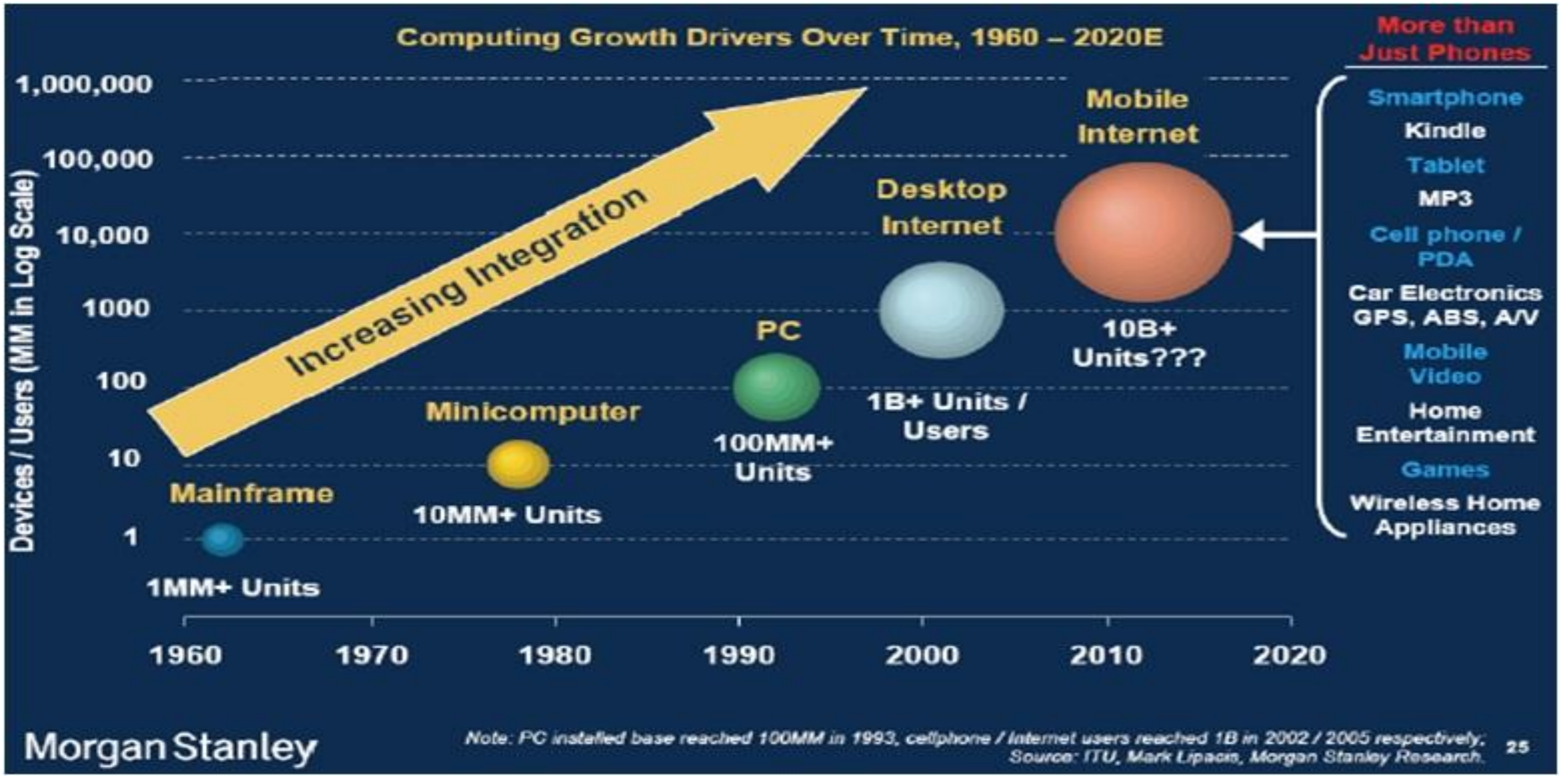
Silicon area = L^2, Footprint = L/4
Corner to corner distance = L + ε

CONFIDENTIAL 15

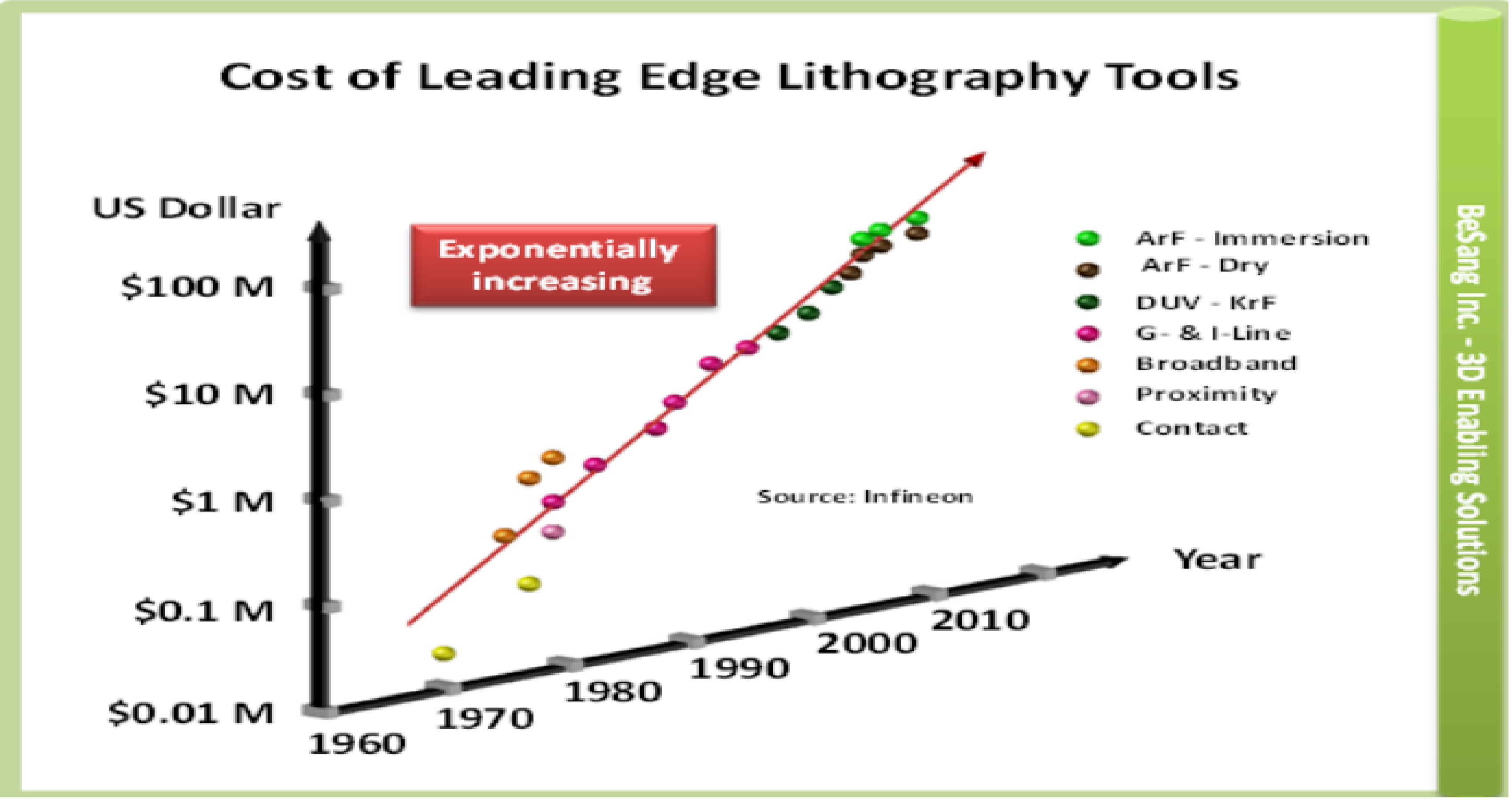
SYNOPSYS® Predictable Success

3D-ICs: The Heat Removal Question

- Sub-1W smartphones, cellphones and tablets the wave of the future
- Heat removal not a key issue there → can 3D stack. Also, shorter wires → net power reduced.

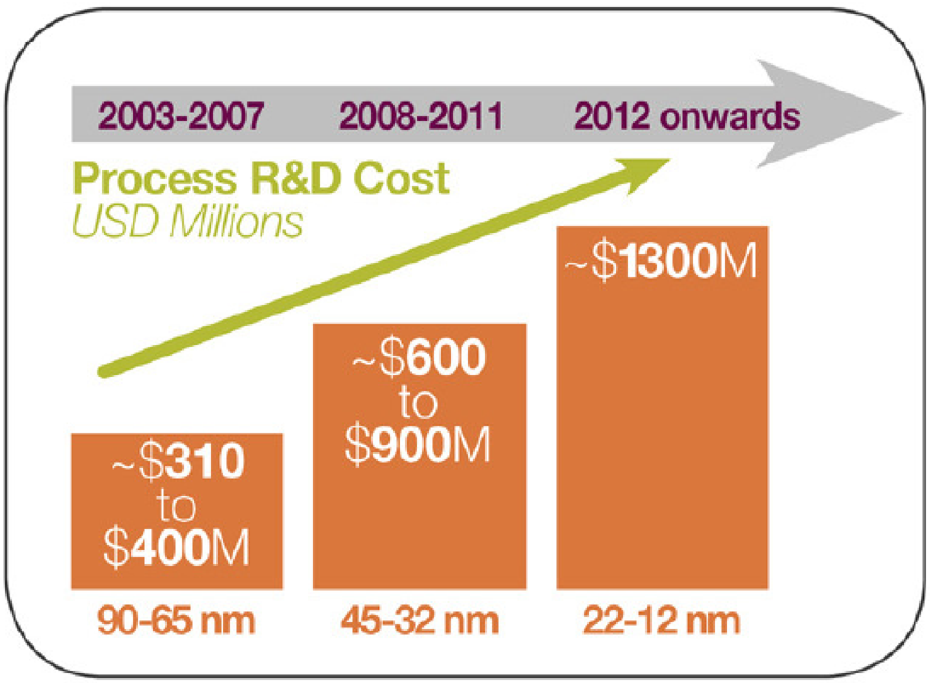


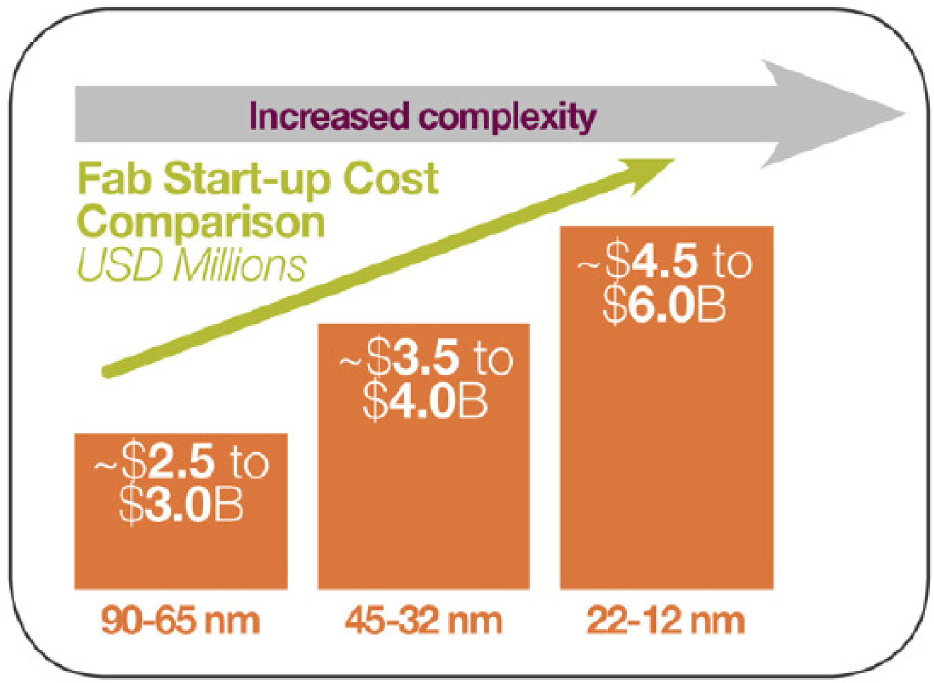
Escalating Cost of Litho to Dominate Fab and Device Cost



The Dilemma of the Semiconductor Industry

- Chip-makers need to keep pace with technology and focus on design
- …while chip manufacturing and technology R&D continue to grow in cost and complexity

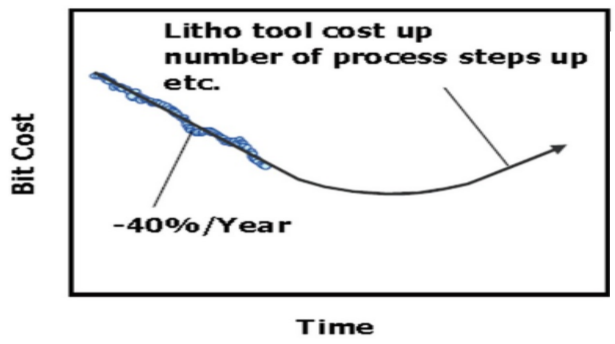




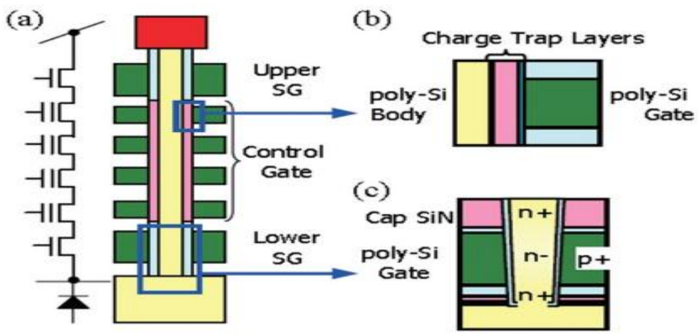
Courtesy: GlobalFoundries

Other parts of the industry (eg. flash memory) → actively exploring SCALE-UP as alternative to SCALE-DOWN

Bit-cost of flash memory if current trends continue [Source: Toshiba, VLSI 2007]



Toshiba's monolithic 3D solution, BiCS



(a) Cross section of BiCS flash memory string, (b) Cross section of vertical SONOS cell, (c) Cross sections of vertical FET.

- Flash memory moving to quad patterning at the 1x nm node → costly.
  Future litho roadmap (eg. EUV) risky.
- Smaller feature size flash memory cells → degrade severely.

Toshiba, Samsung, SanDisk, Micron, Hynix's flash memory roadmaps → monolithic 3D top option beyond 1x nm node

RCATs vs. Planar Transistors: Experimental data from Samsung 88nm devices

From [J. Y. Kim, et al. (Samsung), VLSI Symposium, 2003]

RCATs → Less junction leakage

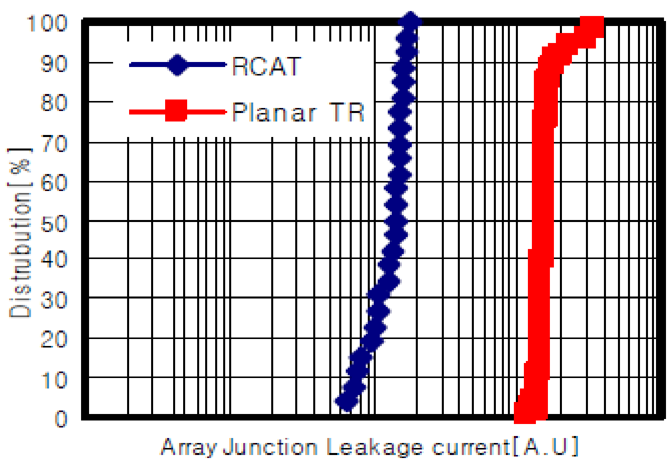


Fig.18. Junction leakage current reduce by 1 order using the RCAT.

RCATs → Less DIBL i.e. short-channel effects

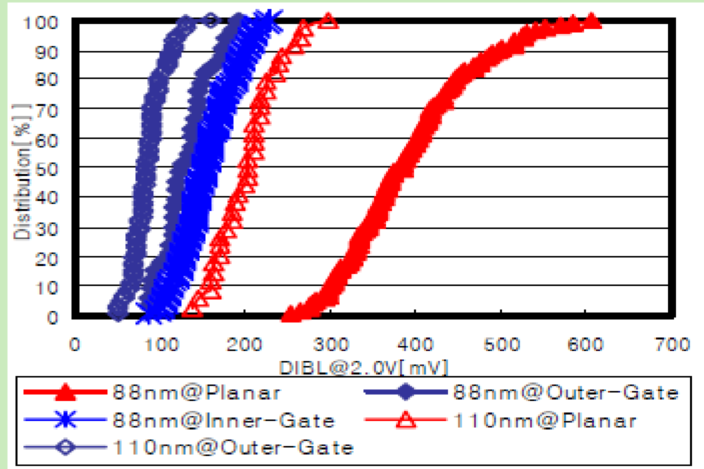


Fig.16. Distribution of DIBL compared with RCATs and Planar TRs

RCATs vs. Planar Transistors (contd.): Experimental data from Samsung 88nm devices

From [J. Y. Kim, et al. (Samsung), VLSI Symposium, 2003]

RCATs → Similar drive current to standard MOSFETs → Mobility improvement (lower doping) compensates for longer L_{eff}

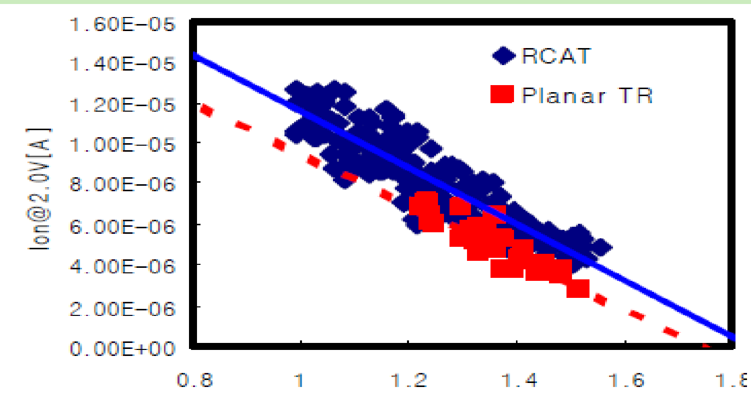


Fig.14. Correlation of Ion and Vth compared with RCAT and planar TR

RCATs → Higher I/P capacitance

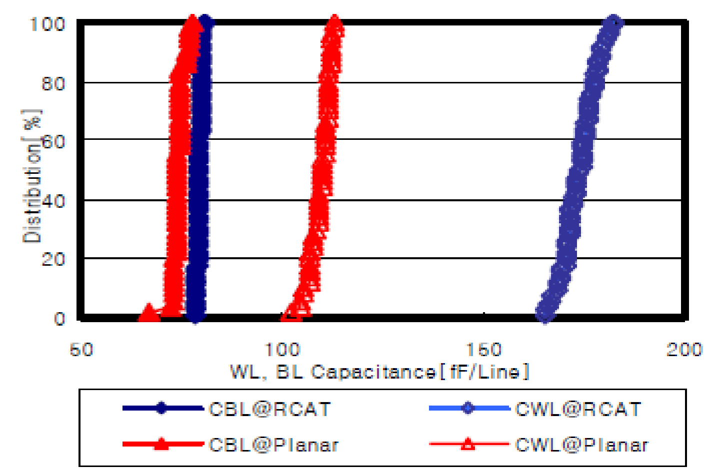
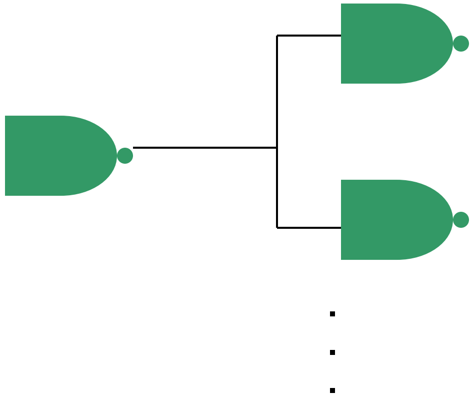


Fig.20. The comparison of WL and BL capacitance with RCAT and Planar TR

Logic gate model

Logic gates:

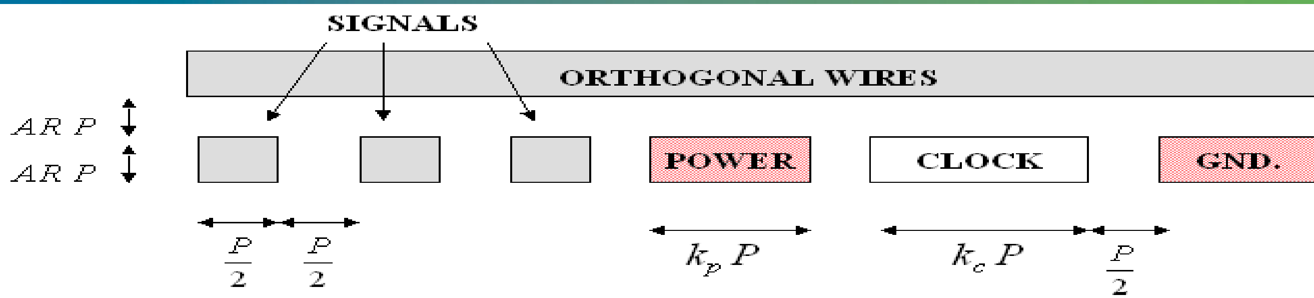
Two input NAND gates with average wire length, fan-out user defined



$$t_d = L_d 0.7 \frac{R_{NAND}}{W}\left(f.o.C_{NAND}W + f.o.\chi c L_{avg} \right)$$

Find W for a certain performance target

Global interconnect model



$$P = Max.\left[2.\left(k_p + 0.5\right).N_{power\_pads} \cdot \rho . \frac{I_T . d_{pad\_to\_pad}{}^2}{\pi . e_{router} . A . AR . k_p . V_{IR}} . \ln\left(\frac{0.65 . d_{pad\_to\_pad}}{l_{pad}} \right), \quad \frac{D}{2}\sqrt{\frac{c_{clock}\rho}{AR \cdot k_c R_o C_o}} . \frac{1}{\frac{\beta_0}{f R_o C_o} - 11}\left(\sqrt{72.6 + \frac{4.4\beta_0}{f R_o C_o}} + 11 \right) \right]$$

Global wire pitch obtained based on two conditions:

(1) Signal bandwidth maximized with power grid IR drop requirement being reached

(2) Wire pitch big enough to drive a clock H tree of a certain length

Results match well with commercial processors [D. C. Sekar, et al., IITC 2006]

Local and semi-global interconnect model

Condition 1:

Wiring area available = Wiring needed for routing the stochastic wiring distribution

$$e_w 2A = \chi P \sqrt{\frac{A}{N_{sockets}}} \int_{l_{\min}}^{l_{\max}} l i(l) dl$$

Condition 2:

RC delay of longest signal wire in each wiring pair = fraction of clock period

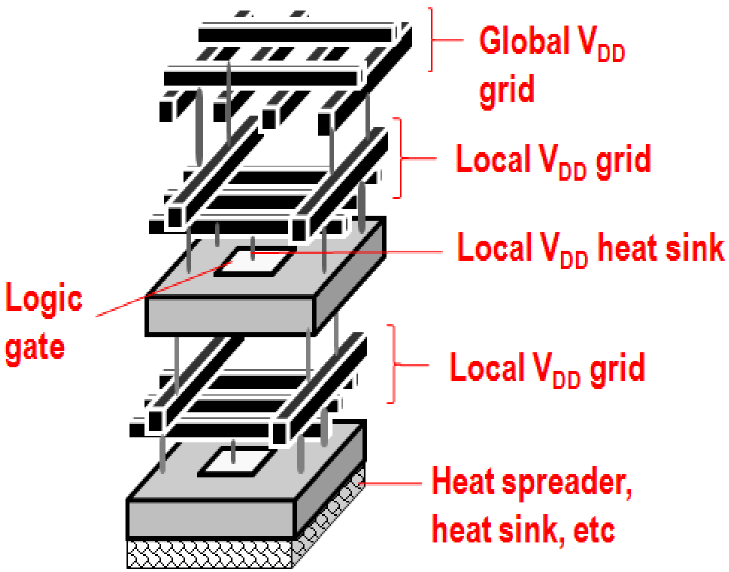
For wires with repeaters, new Energy-Delay Product repeater insertion model used

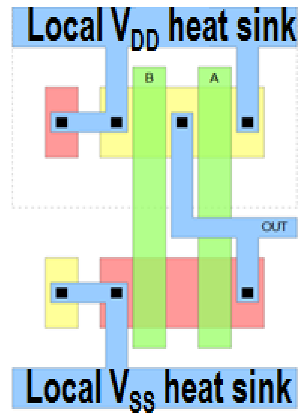
Condition 3:

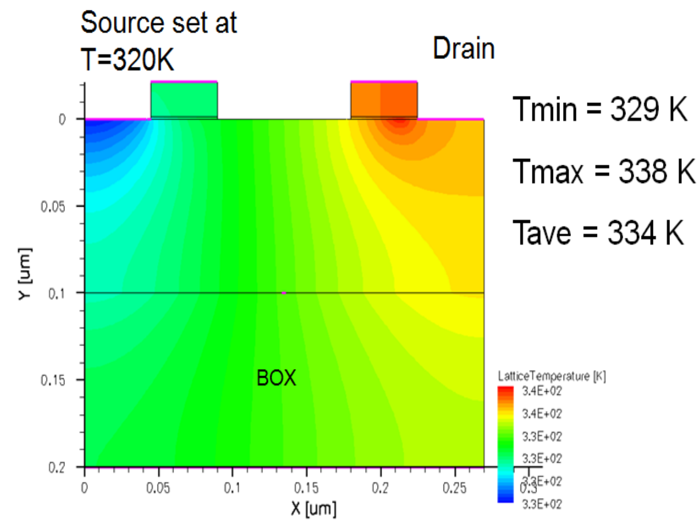
Wire efficiency (e_w) = 1 – fraction of wiring area lost to power wiring, via blockage

[Sarvari, et al. - IITC'07] [Q. Chen, et al. – IITC'00]

Thermal model







- Idea: Use V_{DD}/V_{SS} contacts of each stacked gate to remove heat from it. Design standard cell library to have low temp. drop within each stacked gate.
- Low (thermal) resistance V_{DD} and V_{SS} distribution networks ensure low temp. drop between heat sink and logic gate
- IntSim v2.0: Computes temp. rise of 3D stacked layers using models.

Algorithm used to combine together all these models

1. User inputs parameters
2. Logic gate sizing
3. Select rough initial power estimate
4. Design multilevel interconnect network (including power distribution) for 3D chip with this power estimate
5. Find power predicted by IntSim v2.0
6. Is predicted power = initial power? If yes, this is the final interconnect network. If no, choose new initial power estimate = average of previous initial power estimate and IntSim v2.0 estimate. Go to step 4.
7. Output data

Iterative process used for designing chip

Demo

IntSim v2.0
App

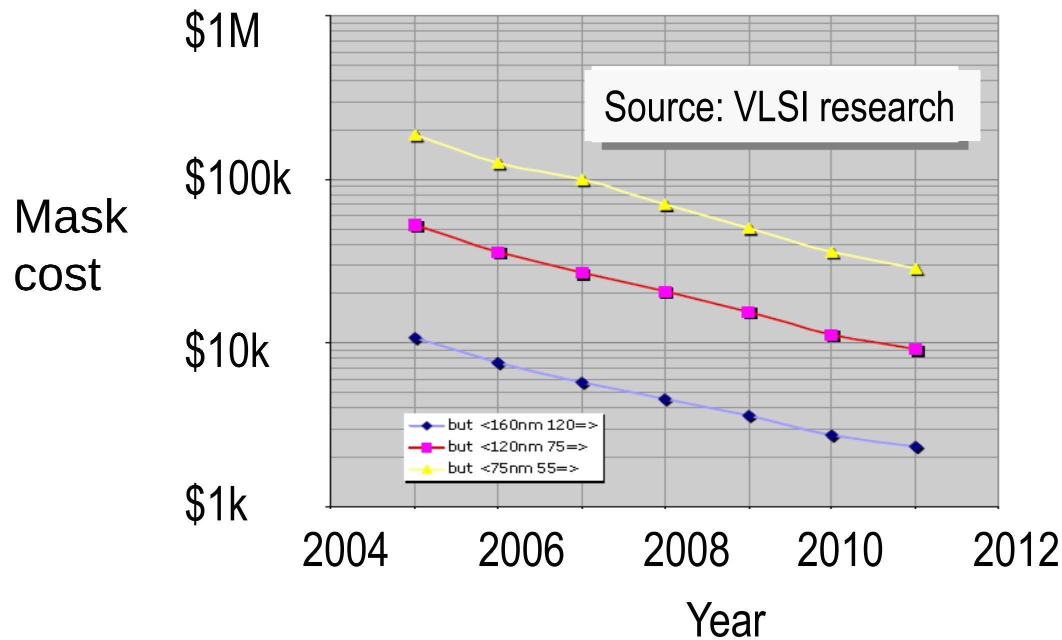
Utility of IntSim v2.0:

- **Pre-silicon optimization and estimation of frequency, power, die size, supply voltage, threshold voltage and multilevel interconnect pitches**
- **Study scaling trends and estimate benefits of different technology and design modifications**
- **Undergraduate and graduate courses in universities for intuitive understanding of how a VLSI chip works**

Monolithic 3D → Can use cheap depreciated equipment and still get the benefits of feature size scaling

Equipment value depreciates 50% every 2 years

Mask cost for a certain feature size goes down 50% every 2 years



For the calculations in this presentation,

- **22nm 2D = Year ‘x’, 15nm 2D = Year ‘x+2’**
- **22nm 2 layer 3D = Year ‘x+2’, depreciated equipment previously used for 22nm 2D**

Cost per Die using Sematech Cost-Of-Ownership Methodology

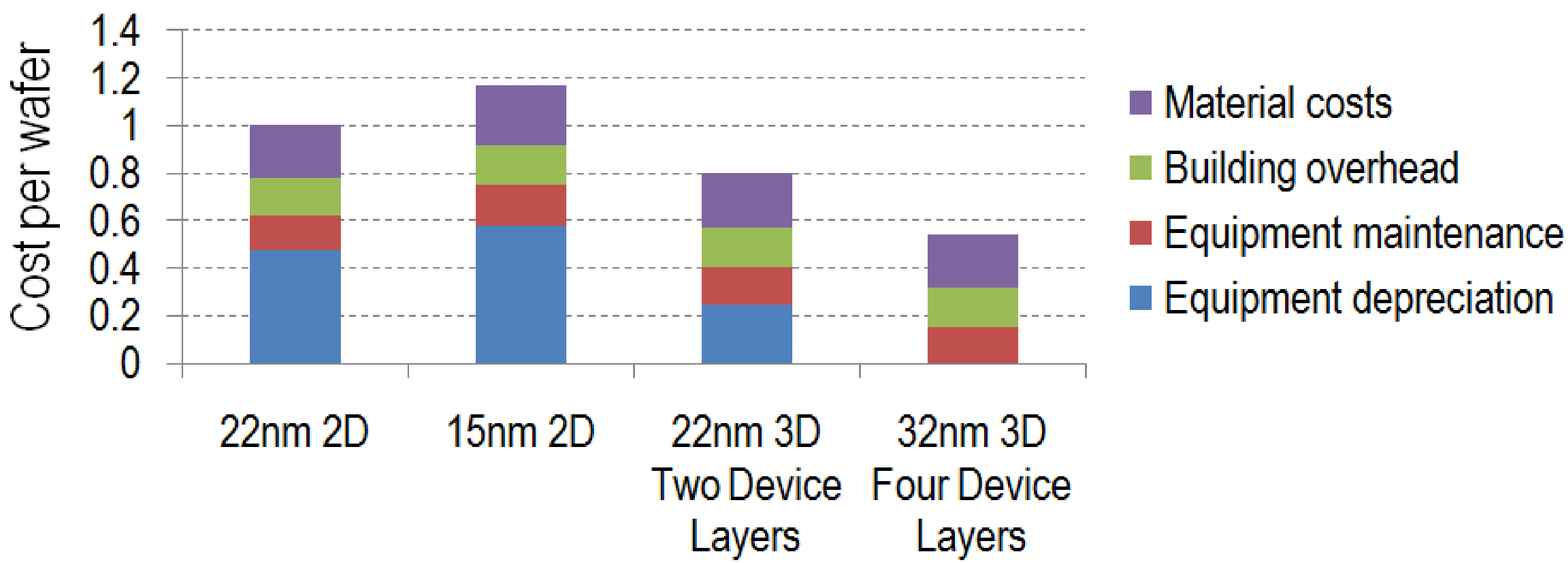
Assumptions:
Die has 50% logic, 50% SRAM. SRAM area →no reduction with monolithic 3D (pessimistic)

	Relative Wafer Cost	Relative Die Size	Relative Cost per die	Cap-ex for upgrade
22nm 2D	1	1	1	
15nm 2D	1.16	0.5	0.6	$4B if all tools changed $800M-$1.1B if only tools related to critical litho steps are changed
22nm 3D 2 layers	0.8	0.75	0.6	$150M
32nm 3D 4 layers	0.54	1.25	0.67	

SCALE-UP → Gives similar cost per die benefits as SCALE-DOWN. But with far less capital expenditure. Largely due to use of depreciated equipment.

Cost-of-Ownership using Sematech Methodology

Equipment depreciation = Tool costs, Maintenance = 7.5% of capex, Building overhead = Cost of facility and labor, Material costs = Masks and chemicals, equivalent of 20k wspm



Monolithic 3D → use depreciated equipment → lower equipment cost → lower wafer cost

Cost Summary

600MHz Die with 50% logic , 50% SRAM	2D-IC @22nm	2D-IC @ 15nm	3D-IC 2 Device Layers @ 22nm
Power	1.6W	0.7W	0.8W
Cost per die	1	0.6	0.6
Capital-expenditure for upgrade		$4B if all tools changed, $800M-$1.1B if only tools related to critical litho steps changed	$150M

Monolithic 3D scaling gives

- Performance, power and cost benefits of feature-size scaling
- But without the large cap-ex, litho risk and production ramp times
- Flash industry → already taken this route, numbers indicate viability for logic too