MonolithIC 3D – General

By MonolithIC 3D Inc.

Issued March 2013

Contents

About the authors of the e-book

Zvi Or-Bach
President and CEO

Zvi Or-Bach is the founder of MonolithIC 3D™ Inc., a Finalist of the “Best of Semicon West 2011” for its monolithic 3D-IC breakthrough. Or-Bach was also a finalist of the EE Times 2011 and 2012 Innovator of the Year Award for his pioneering work on the monolithic 3D-IC.

Or-Bach has a history of innovative development in fast-turn ASICs for over 20 years. His vision led to the invention of the first Structured ASIC architecture, the first single via programmable array, and the first laser-based system for one-day Gate Array customization. In 2005, Or-Bach won the EETimes Innovator of the Year Award and was selected by EE Times to be part of the **"Disruptors" --"The People, Products and Technologies That Are Changing The Way We Live, Work and Play"**.

Prior to MonolithIC 3D, Or-Bach founded eASIC in 1999 and served as the company's CEO for six years. eASIC was funded by leading investors Vinod Khosla and KPCB in three successive rounds. Under Or-Bach's leadership, eASIC won the prestigious **EE Times' 2005 ACE Award for Ultimate Product of the year** in the Logic and Programmable Logic category.

Earlier, Or-Bach founded Chip Express in 1989 (recently acquired by Gigoptix) and served as the company's President and CEO for almost 10 years, bringing the company to $40M revenue, and to an industry recognition for three consecutive years as a high-tech Fast 50 Company that served over 1000 ASIC designs, including many one-day prototypes and one-week production delivery.

Even before his entrepreneurial ventures in ASIC technology, Or-Bach held engineering management positions at Elbit Computers, Ltd., Israel (subsidiary of Elron) and Honeywell (Lexington, Massachusetts). Zvi Or-Bach received his B.Sc. degree (1975) cum laude in Electrical Engineering from the Technion - Israel Institute of Technology, and M.Sc. (1979) with distinction in Computer Science, from the Weizmann Institute,

Israel. He holds over 100 issued or pending patents, primarily in the field of 3D integrated circuits and semi-custom chip architectures. He is the Chairman of the Board for Zeno Semiconductors, Bioaxial and VisuMenu. Or-Bach is passionate about the semiconductor industry, and has participated in **initiatives to improve immigration and education policies** to benefit the same.

Brian Cronquist
Vice President, Technology & IP

Brian Cronquist has over 31 years of semiconductor industry experience, most recently as Sr. Dir. Technology Development & Foundry at non-volatile FPGA provider Actel. He has global experience on “both sides of the silicon wafer table”: starting and building Chartered Semiconductor (Singapore) technology and customers as a captive then pure foundry, and non-volatile (antifuse and flash) FPGA technology at Actel as a fabless partner and customer to over 7 foundries and IDMs. He also led startup wafer fab engineering teams at Sierra Semiconductor, now PMC-Sierra, and developed new process technology at AMI and Synertek/Honeywell.

Mr. Cronquist has a diverse technical interest which includes developing ultra-thin thermal oxide and pre-cleaning technology (first to develop and implement HF-last), plasma etching of metals and oxides, database scaling techniques, process simulation & integration, novel ion implant techniques, first CMOS MOSFETs built with laser (CW) annealing, minimizing process induced damage PID) from plasma etching and ion implantation, time-to-market new product and process introduction (NPI), and customer engineering & program management.

While at Actel, he was also Principal Investigator of over $24M of government funded technology programs developing radiation hardened (RH) versions of both anti-fuse and flash based product families in commercial and RH foundries. He has published over 85 technical papers in the fields of semiconductor microelectronic radiation effects and hardening, as well as new logic, antifuse & flash processes, devices, and reliability. Mr. Cronquist graduated cum laude (Chemistry Medal) in Chemistry from Santa Clara University in 1979. Currently, he is a visiting researcher at the Rice University Chemistry Department and an Industry Affiliate Partner at the Stanford University Nanofabrication Facility.

Ze'ev Wurman

Chief Software Architect

Wurman has over 30 years of experience in developing algorithms, CAD software, and hardware and software architectures. Before MonolithIC 3D Inc., he led the software development groups in DynaChip, an FPGA startup later acquired by Xilinx, and eASIC, a programmable logic company. Prior to that Wurman was the architect for hardware simulation accelerator at Amdahl, the largest and fastest hardware accelerator at the time. He designed and managed CAD software for Silvar-Lisco, and spent three years with IBM Research in Haifa, Israel, working on algorithms for design verification, databases, and cryptography. Between 2007 and 2009 Wurman served as senior policy adviser in the office of Planning, Evaluation, and Policy Development, in the U.S. Department of Education.

Wurman holds B.Sc. and M.Sc. degrees in Electrical Engineering from Technion, Israel Institute of Technology, in Haifa, Israel. He has published technical papers in professional and trade journals and holds seven patents.

Israel Beinglass, Ph.D.
CTO Device Integration

Dr. Israel Beinglass has over 25 years of diversified experience in the semiconductor and semiconductor equipment industries. From 1989 through 2006 he was a senior technologist and business executive at Applied Materials (NASDAQ:AMAT), the global leader in semiconductor equipment, where he served in a variety of executive roles. He was General Manager of the High Temperature Films Group, Managing Director and Chief Technology Officer (CTO) for the Front End Equipment Group, and later CTO of Applied Global Services Group. He also served as the Chief Marketing Officer (CMO) of the CMP division and CTO for the Thin Films Group (TFG). Dr. Beinglass was involved in numerous successful acquisitions and was a member of Applied Materials Strategy and Marketing councils.

Dr. Beinglass is the industry pioneer of selective deposition; he is the co-inventor of the selective tungsten deposition process and an earlier developer of selective Epi deposition. He also was instrumental in developing the industry's first single-chamber polysilicon deposition system and an integrated, multi-chamber Policide system.

Before joining Applied Materials, Dr. Beinglass worked at Intel Corporation (NASDAQ:INTC) and IMP (NASDAQ:IMPX) in various positions, including process development manager and engineering manager of fab operations. While working at Intel he was the inventor of the selective tungsten deposition process. He was a co-recipient of the Beatrice Winner Award for Editorial Excellence at the Intl. Solid State Circuits Conference in 1982. He serves as a board member at Noise Free Wireless and Spectros and used to serve as a board member at Silicon Genesis between 1998 and 2001.

Dr. Beinglass holds a Ph.D. in Materials Science from the Hebrew University of Jerusalem; he completed post-doctoral research at UCSF and is the holder of 30 US patents (288 citations) as well as several pending patents.

Deepak Sekar, Ph.D.

Former Chief Scientist

Dr. Deepak Sekar received a B. Tech from the Indian Institute of Technology (Madras) in 2003 and a PhD from the Georgia Institute of Technology in 2008. He worked at SanDisk Corporation between 2006 and 2010, and conducted research on non-volatile memory. He joined MonolithIC 3D™ Inc. in early-2010 as a Principal Engineer, in 2011 as Chief Scientist of the company and left the company in March 2012.

For the past 8 years, Dr. Sekar's research has focused on 3D Integrated Circuits. His PhD research involved doing some of the first experimental work on microchannel cooled 3D stacked chips. He also developed a CAD tool called IntSim that simulates 2D and 3D stacked systems. At SanDisk, Dr. Sekar worked in the area of 3D crosspoint memory and developed rewritable memory devices, selector diodes and array architectures.

Dr. Sekar is the author of a book, an invited book chapter, 15 publications and 55 issued or pending patents, predominantly in the field of 3D integration. Awards he has received include a Best Student Paper Award at the Intl. Interconnect Technology Conference (2008), a Best Paper Award at the IETE Technical Review (2009), an Intel PhD Fellowship (2006-2008), a Motorola Electronic Packaging Fellowship Award at the Electronic Components and Technology Conference (2008), two Inventor Recognition Awards from the Semiconductor Research Corporation (2006, 2009) and the National Talent Scholarship from the Government of India (1997-2003). He serves as a Program Committee Co-Chair at the International Interconnect Technology Conference and as an Advisory Board Member for 3D InCites.

Part 1: Monolithic 3D - General

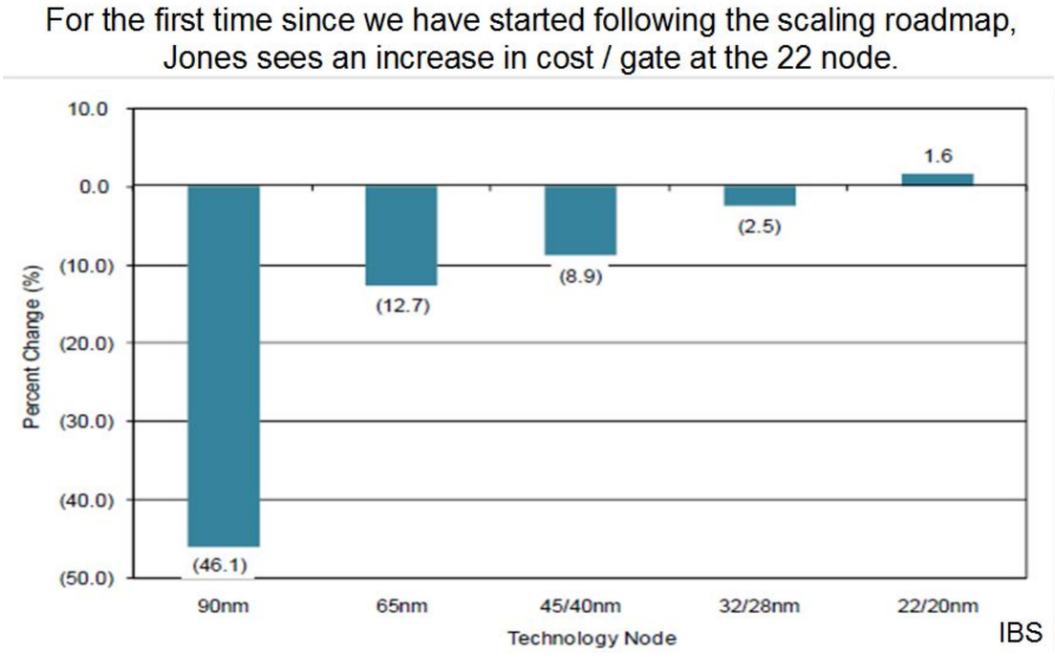
Chapter 1 - Is the Cost Reduction Associated with Scaling Over?

by Zvi Or-Bach, the President and CEO of MonolithIC 3D Inc.

"Yes, unless we Augment Dimensional Scaling with monolithic 3D-IC Scaling"

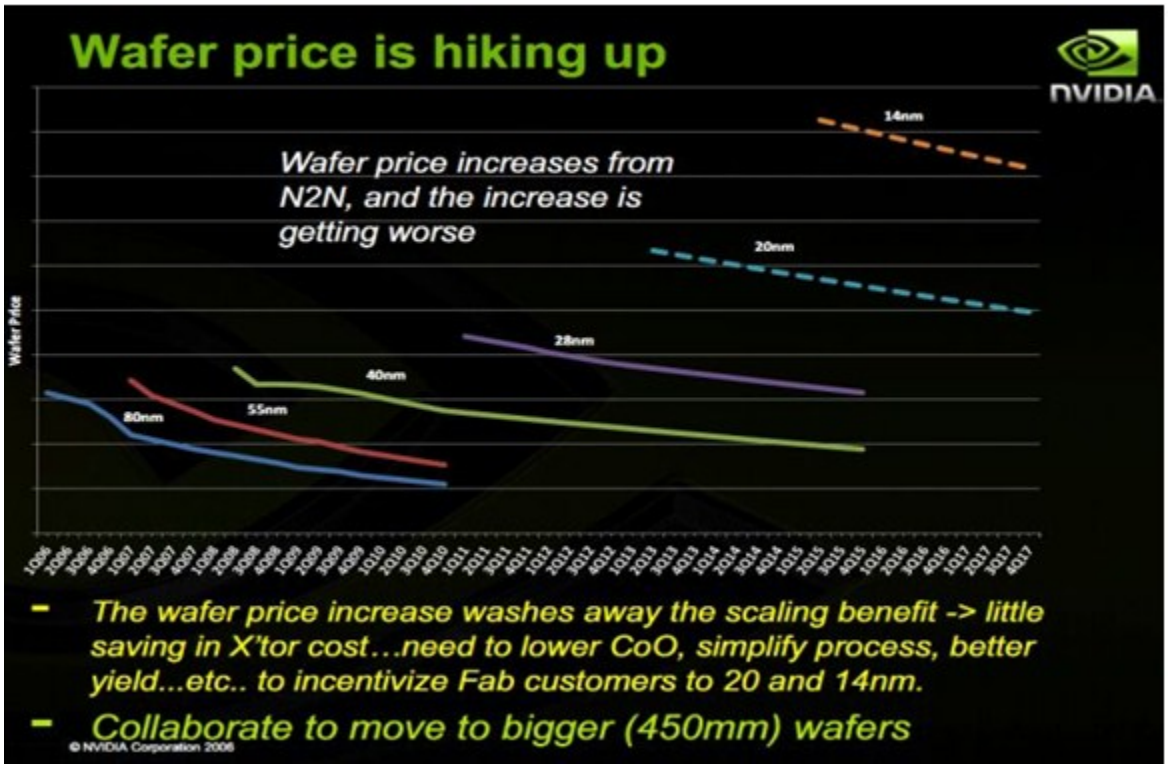
The last 50 years of the semiconductor industry have been all about the manifestation of Moore's Law in dimensional scaling of Integrated Circuits (ICs). As consumers of electronic devices we all love to see with every new product cycle better products at a lower cost. But now storm clouds are forming, as was recently publicly expressed "Nvidia deeply unhappy with TSMC, claims 20nm essentially worthless".

Clearly dimensional scaling is no longer associated with lower average cost per transistor. The chart below, published by IBS about a year ago, shows the diminishing benefit of cost reduction from dimensional scaling. In fact, the chart indicates that the 20nm node might be associated with higher cost than the previous node.

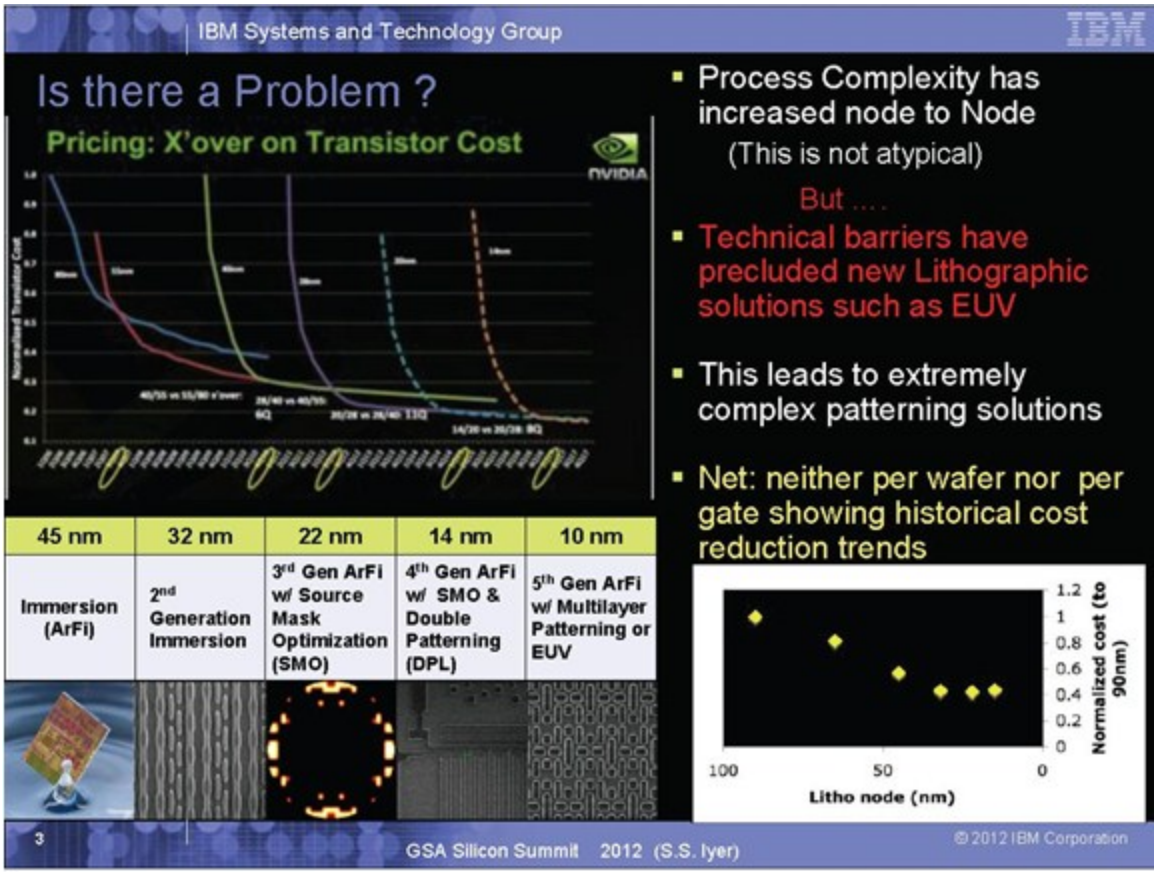


The following Nvidia chart provides the first order explanation. The cost reduction of dimensional scaling resulted from doubling the number of transistors per wafer. But if

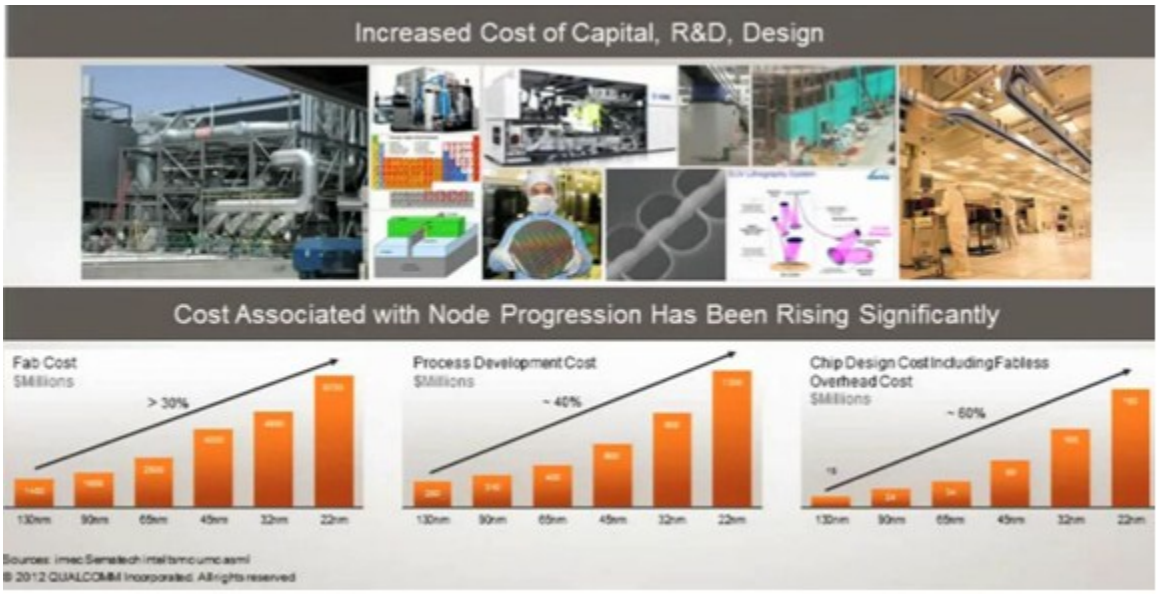
the wafer cost of the new technology node increases by too much then it neutralizes that cost reduction. The Nvidia chart shows the wafer cost of recent nodes over time. In the past (...80nm, 55nm, 40nm) the incremental wafer cost increases were small and rapid depreciation of those costs resulted in almost constant average wafer price. Recent nodes (28nm, 20nm, 14nm,...), however, signal a new reality.



The following busy slide of IBM summarizes it clearly: "Net: neither per wafer nor per gate showing historical cost reduction trends"

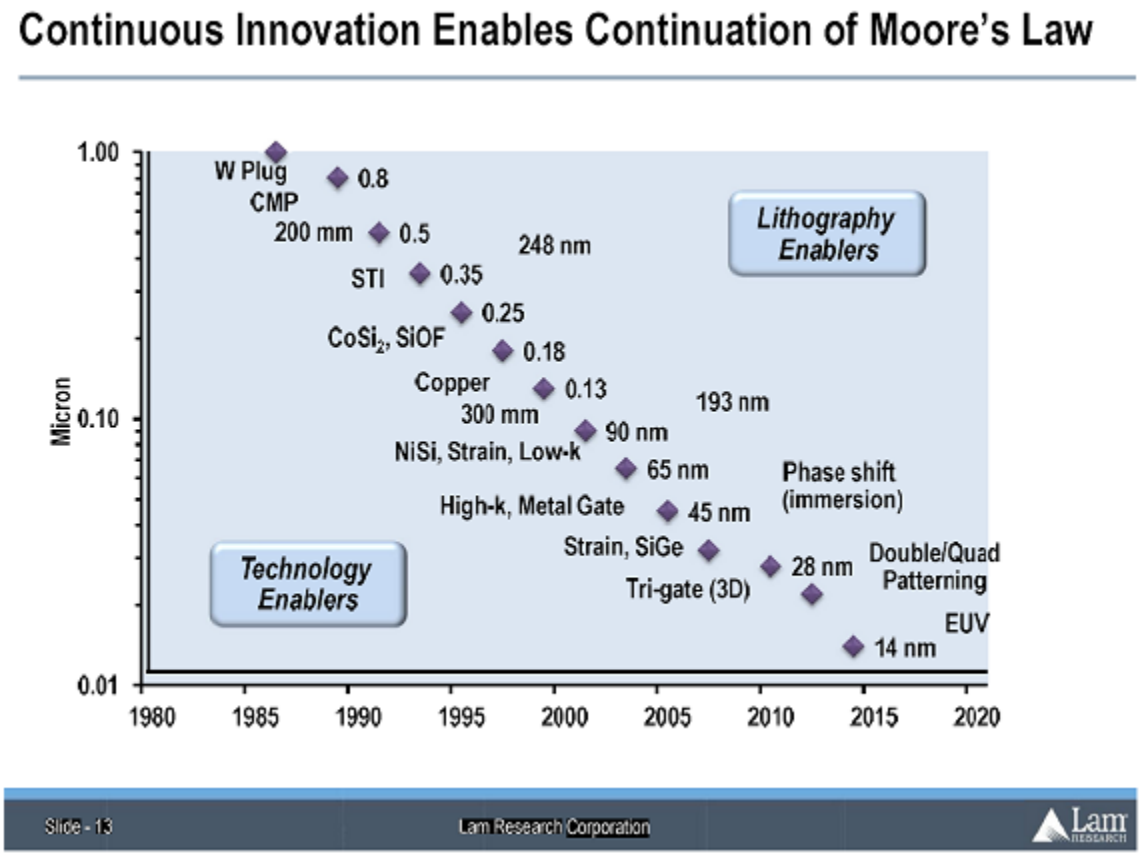


The number one driver to the increase of wafer cost is the increase in the equipment cost required for processing the next technology node. The following chart presents the increase in costs of capital, process R&D, and design.

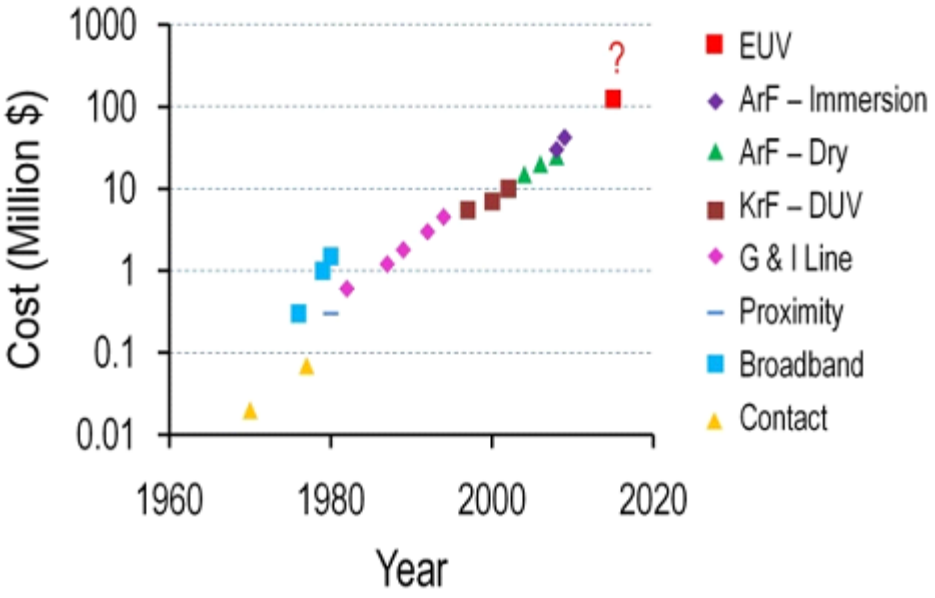


The sharp increase of costs associated with scaling is a new phenomenon. There were always costs to move from one node to the next, but they were about constant or incrementally small.

The following slide presents the innovations that enable dimensional scaling. Clearly, for many nodes we were able to use the same lithography tools. But once dimensional scaling reached the limit of light wavelength the lithography tool became critical and dominant. About for every node the lithography became a major challenge that required newer equipment and substantial process R&D. Moreover, in the recent lithography nodes the transistor itself required significant innovation at every node (high-k, Metal Gate, Strain, SiGe, Tri-gate,...) and it is clear that future scaled nodes will require even more of those innovations and their associated costs.



An important part of these costs is the escalating cost of the capital equipment for the next node fabrication lines. The following figure present the cost dynamic for the lithography equipment. Note the logarithmic scale of the cost axis.



Lithography tools grew from less than 10% of wafer fab equipment (WFE) spending to over 25% and accordingly lithography now represents about 50 % of the wafer cost.

An interesting implication of growing domination of lithography in semiconductor processing is the fact that the ASML, which is the lead vendor of lithography tool, recently passed Applied Material's (the leader of all other tools) market cap. Following is the chart of the stock price of ASML (in red) vs. Applied Material (AMAT).

The clear conclusion of all of this is that future dimensional scaling is not about to change these trends. Accordingly, as stated in the IBM slide above: "Net: neither per wafer nor per gate showing historical cost reduction trends." Unless ...

Unless we change the way we do scaling (remember Einstein's famous quote). Moore's Law is about doubling the number of transistors in a semiconductor device. At that time dimensional scaling was one of the three trends Moore described that would enable the observed and predicted exponential increase of device integration. It would seem that it is about time to look on another one of those - increasing the die size. If we do it by using the 3rd dimension – monolithic 3D-IC – we can achieve both higher integration and cost reduction!

It is not that we should stop scaling down, it just that if we augment it with scaling up we can introduce the required changes that can achieve the continuation of the cost reduction trend. Clearly almost all of the increases of wafer costs are related to the pace of dimensional scaling. If those costs could be spread over four years instead of two then the increase in wafer cost would be only about half of what it is now.

It might not be so clear, however, why monolithic 3D should reduce wafer cost. Shouldn't the cost of the double die size spread over two layers be at least double …?

Monolithic 3D IC would reduce wafer cost because of the following elements:

1. Reduced Die Size - It has been shown in many research studies that each folding into 3D has the potential to reduce the total required silicon area by 50% due to the reduced re-buffering and reduced sizing of the buffers.

2. Depreciation - Scaling up enables the use of the same fab and process R&D for few additional years with the associated improvement in deprecation costs and improved manufacturing efficiencies and yield.

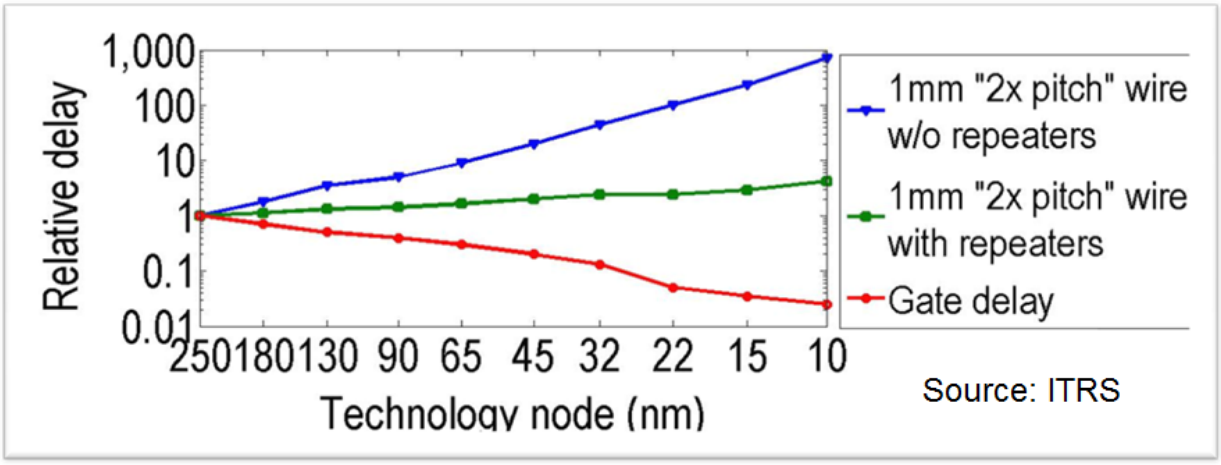
3. Heterogeneous Integration - Scaling up would enable heterogeneous integration. This will open up the third trend of Moore- improved circuit design. As each strata of 3D IC could be processed in a different flow, cost and power could be saved by using a different process flow for logic, memory and I/O.

4. Multiple Layers Processed Together - This would be most effective for a memory type circuits. Using the right architecture, multiple transistors layers could be process simultaneously with the result of a huge reduction of cost per layer.

Let's detail each of these.

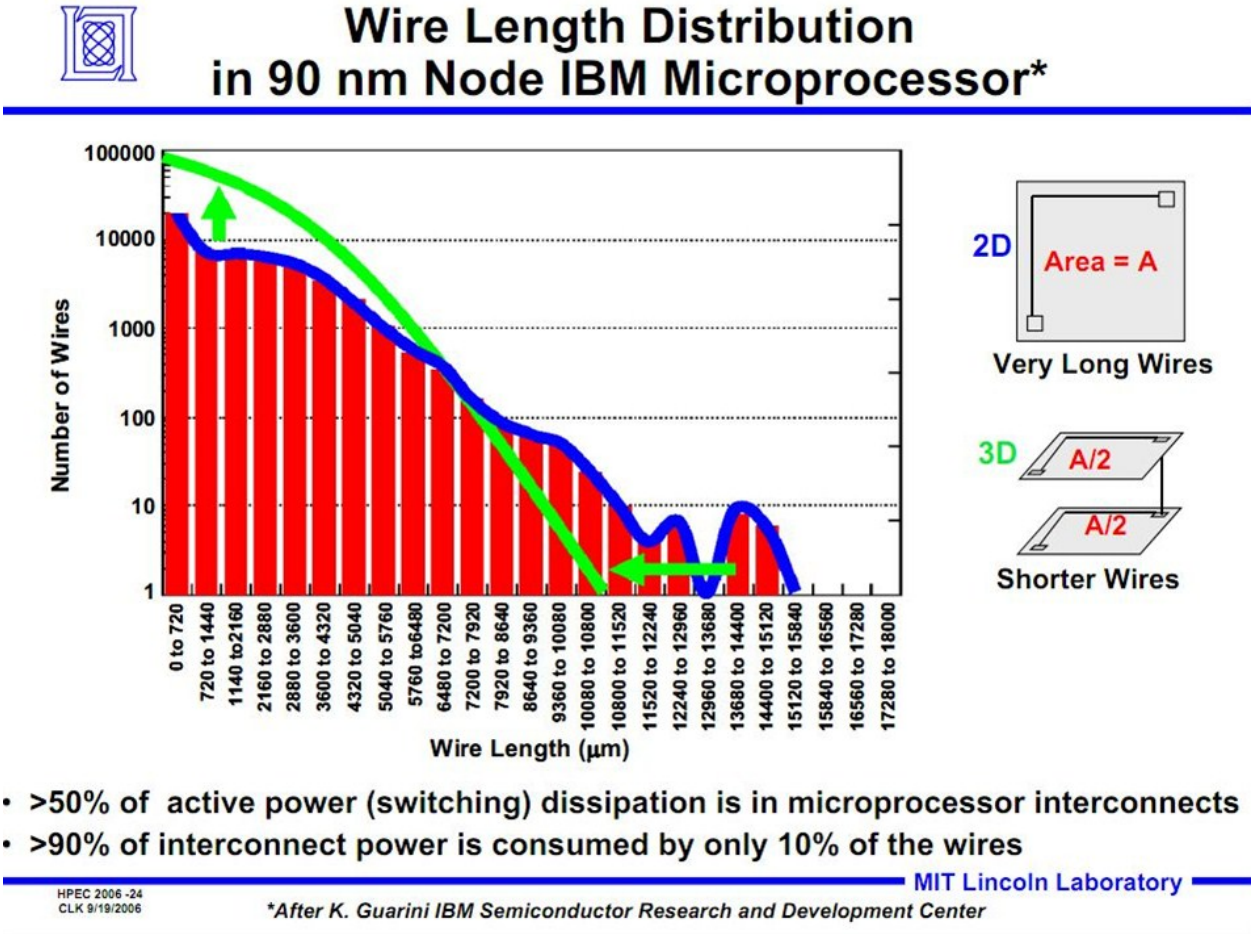
Reduced Die Size

Dimensional scaling has always been associated with an increase of wire resistivity and capacitance. The industry had spent a huge effort to overcome these by first replacing the conducting material with copper and then changing the isolation material to low-K dielectrics. But the interconnect problem is still growing as demonstrated in the following chart.



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

Every node of dimensional scaling is associated with larger cells, output drivers, and more buffers and repeaters. Monolithic 3D enables one to fold the circuit where the next strata is about 1µ above with a very rich vertical connectivity between the strata. The following IBM/MIT slide illustrates the effectiveness of such folding.



Further, the reduced silicon area generates an additional reduction of buffers and the average transistor size. MonolithIC 3D Inc. released an open-source top level simulator IntSim v2.0 to simulate a given design's expected size and power based on process parameters and the number of strata (more than 300 copies have been downloaded so far).

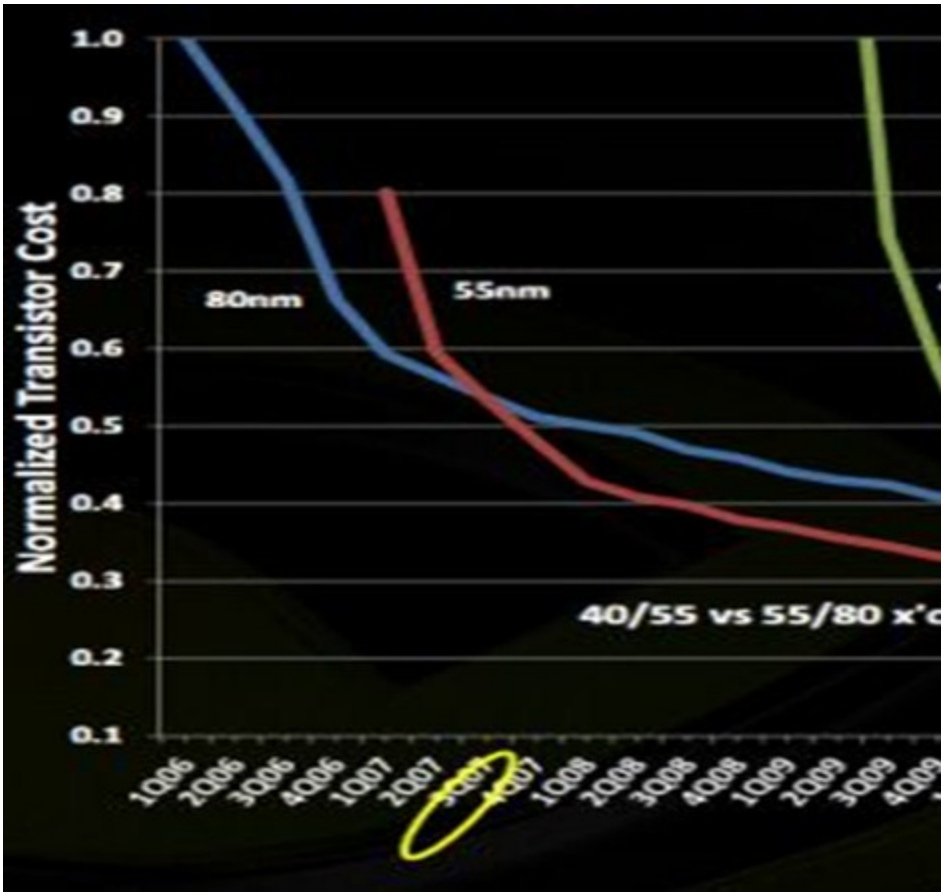
Using the simulator we can see in the following table that a design that uses 50 mm2 with average size gate size of 6 W/L, will need an average gate size of 3 W/L and accordingly only 24 mm2 if folded into two strata (the footprint will be therefore just 12 mm2).

22nm node 600MHz logic core	2D-IC	3D-IC 2 Device Layers	Comments
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	

These results are in-line with many other monolithic 3D research results.

Depreciation

The semiconductor industry is very capital intensive and a very significant part of the wafer cost is associated with the cost of capital. Since every two years we have been scaling to a new node, then the wafer cost needs to support this rapid loss of capital value. Achieving the next level of device functionality using the same generation of tools allows for a far better utilization of the investment capital. In addition the learning curve of yield and manufacturing efficiency contributes further to the end-product cost reduction. The following chart portion demonstrates this well-known trend.



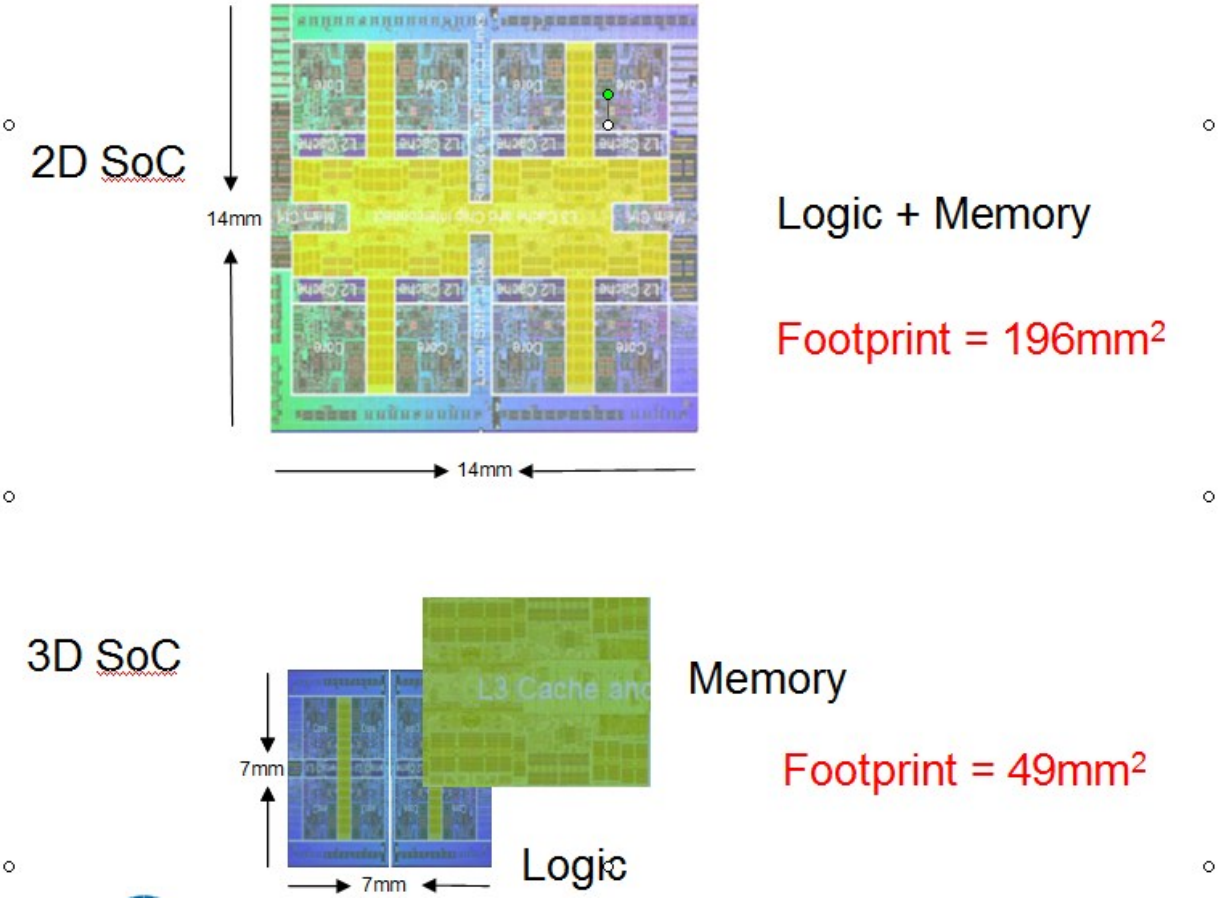
Heterogeneous Integration

Let's start with quoting Mark Bohr, in charge of Intel's process development:

"Bohr: One important perspective is that chip technology is becoming more heterogeneous. If you go back 10 or 20 years ago, it was homogenous. There was a CMOS transistor, it was the same materials for NMOS and PMOS, maybe different dopant atoms, and that basic CMOS transistor fit the needs of both memory and logic. Going forward we'll see chips and 3D packages that combine more heterogeneous elements, different materials, and maybe transistors with very different structures whether they're for logic or memory or analog. Combining these very different devices onto one chip or into a 3D stack—that's what we'll see. It will be heterogeneous integration"

The most important market for semiconductor products is smart mobility. For this market the SoC device needs to integrate many functions. In most cases the pure high-performance logic would be about 25% of the die area, 50% would be memories and the rest would be analog functions such as I/O. In 2D they all need to be processed together and bear the same manufacturing costs. In a monolithic 3D-IC stack using

heterogeneous integration each stratum is processed in an optimized flow, allowing for a significant cost reduction. The following illustration suggests the use of only two strata to build a device that in 2D would have a size of 196 mm2. By having one stratum for logic and one for memory, and by using DRAM instead of SRAM, the device could be reduced to 98 mm2 with footprint of 49 mm2. The device cost would be further reduced by the memory using only 3 or 4 metal layers.

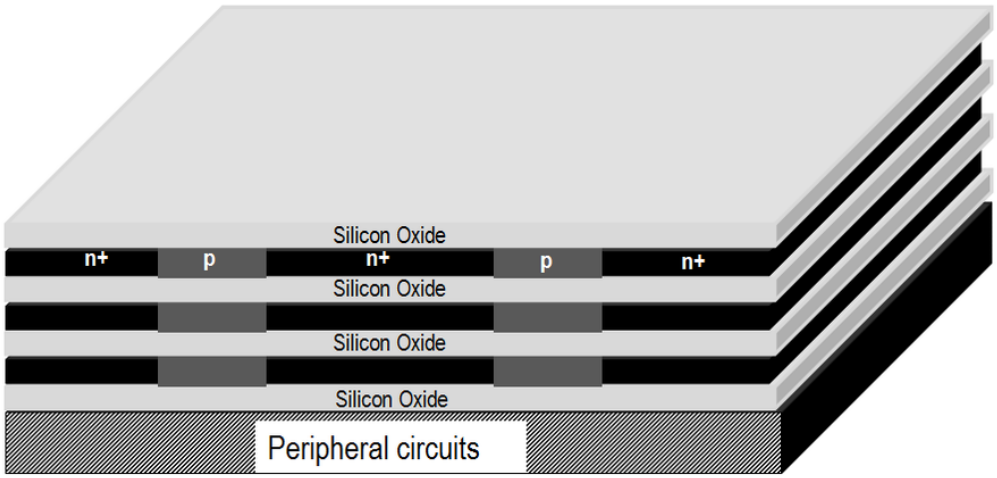


Multiple Layers Processed Together

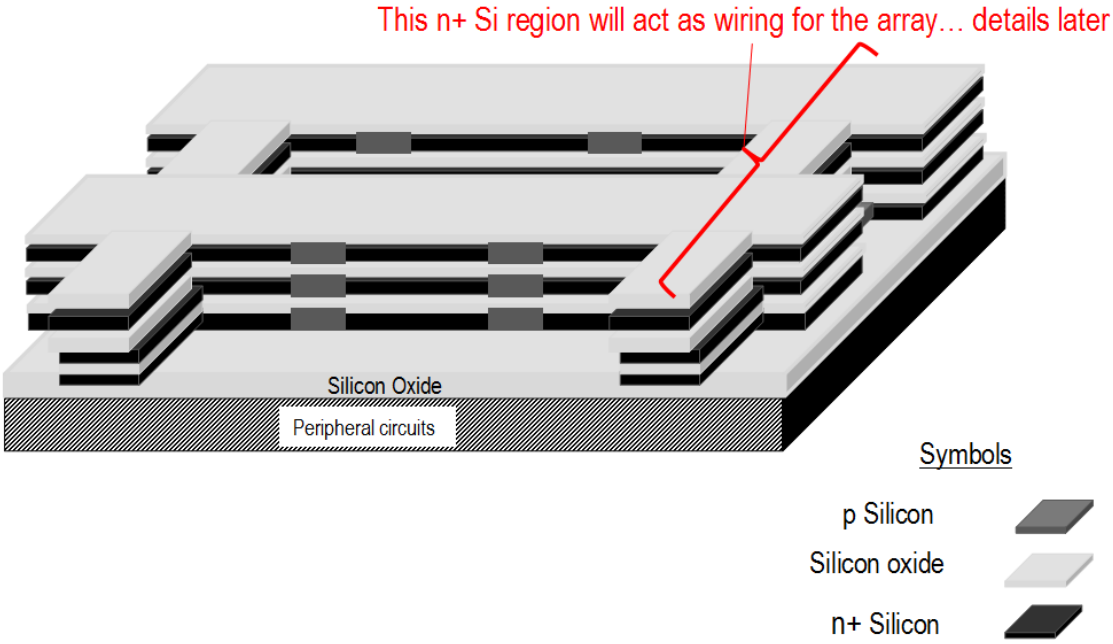
Using the right architecture, multiple transistor layers could be processed together with a huge reduction in cost per layer. This could be applied to many different types of regular devices.

The following illustrate the concept with respect to a floating-body DRAM:

Process Flow: Step 6
Using methods similar to Steps 2-5, form multiple Si/SiO_2 layers, RTA

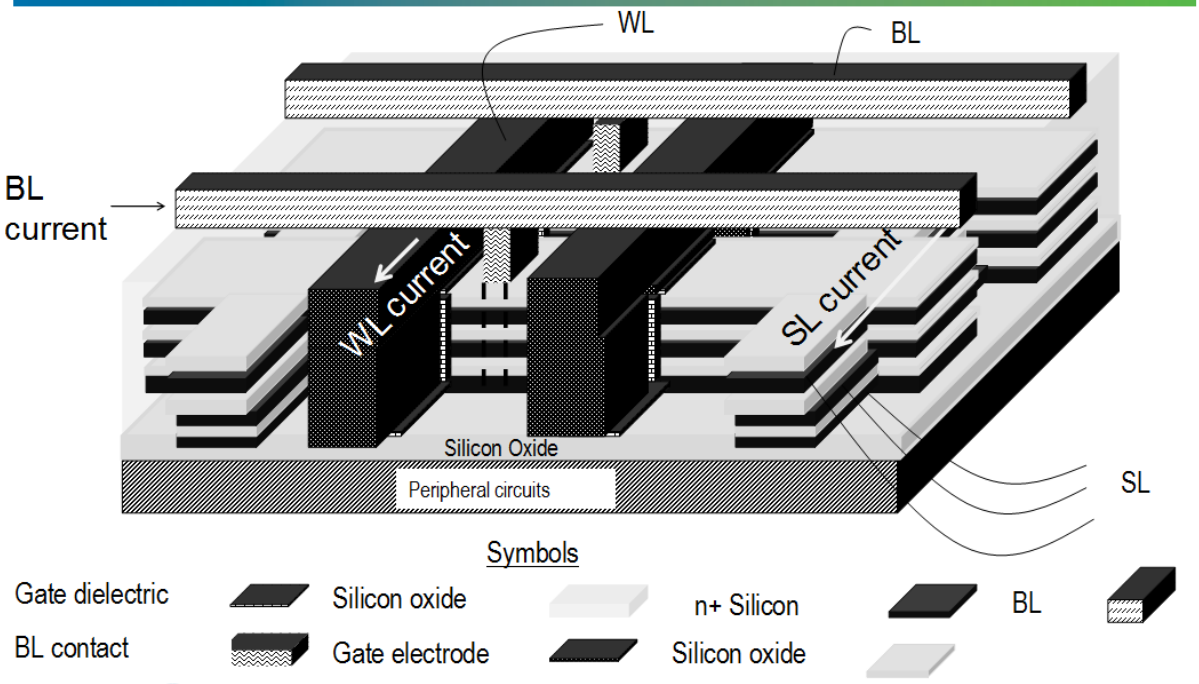


Process Flow: Step 7
Use lithography and etch to define Silicon regions



Process Flow: Step 11

Construct BLs, then contacts to BLs, WLs and SLs at edges of memory array using methods in [Tanaka, et al., VLSI 2007]



MonolithIC 3D Inc's website presents more details for the DRAM flow, and also related flows for RRAM and NAND Flash memories.

In short, we do have a path to continue the semiconductor industry drive for better products and with lower costs, but we should continuously apply innovation to do so. Now that monolithic 3D is practical, **it is time to augment dimension scaling with monolithic 3D-IC scaling**.

Chapter 2 - IEDM 2012 - The Pivotal Point for Monolithic 3D IC

by Zvi Or-Bach, the President and CEO of MonolithIC 3D Inc.

From our biased point of view we see the recent IEDM12 as a pivotal point for monolithic 3D. Here's why:

We start with the EE Times article IEDM goes deep on 3-D circuits, starting with "Continuing on the theme of 3-D circuit technology addressed in an earlier post about this year's International Electron Device Meeting, Rambus, Stanford University and an interesting company called Monolithic 3D will address issues related to cooling 3-D circuits. .." and follow with a quote from the abstract to IEDMs short course "Emerging Technologies for post 14nm CMOS" organized by Wilfried Haensch, of IBM's Watson Research Center:

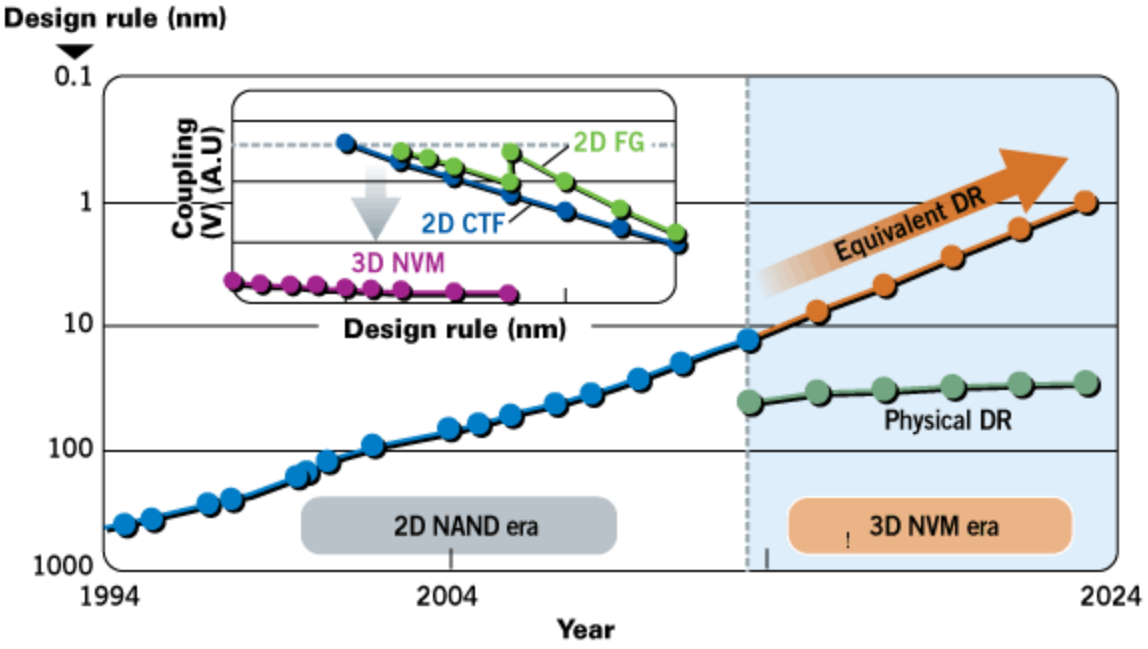
"Scaling the dimension was the key for the unprecedented success of the development of IC circuits for the last several decades. It now becomes apparent that scaling will become increasingly difficult due to fundamental physical limits that we are approaching with respect to power and performance trade-offs. This short course will give an overview of several aspects in this "end-of-scaling" scenario. ..."

We then continue with statements made by Dr. Howard Ko, a Senior Vice President and General Manager of the Silicon Engineering Group of Synopsys in his 2013: Next-generation 3-D NAND flash technology article:

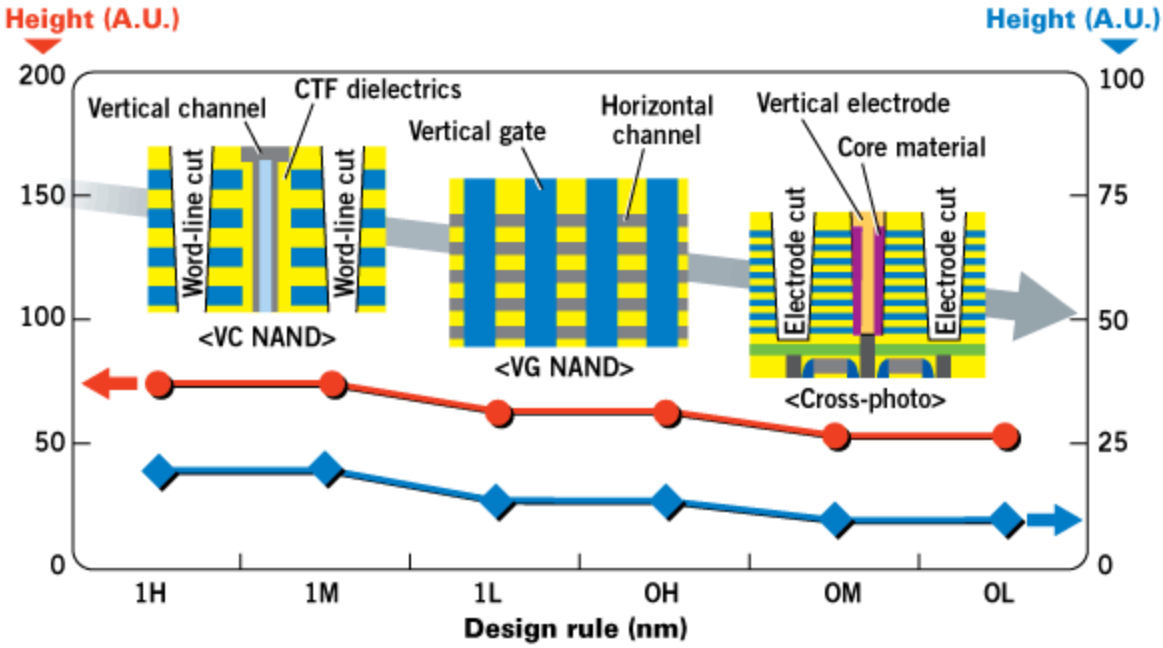
"Yet there are a variety of developments in another type of 3-D scaling that are likely to have a similarly large impact on semiconductors in the near future - 3-D devices for NAND flash.... And as in planar CMOS logic, NAND flash technology has been progressively scaled to smaller feature sizes, becoming the process leader in driving the smallest line-widths in manufacturing as evidenced by the current 1x-nm (~19-nm) process node. Yet, despite plans to scale down to the 1y-nm (~15-nm) and possibly 1z-nm (~13-nm) nodes, the traditional planar floating gate NAND flash architecture is approaching the scaling limit, prompting the search for new device architectures. Not to be upstaged by the planar to 3-D (FinFET) transition in logic devices, NAND flash has embarked on its own 3-D scaling program, whereby the stacking of bit cells allows continuous cost-per-bit scaling while relaxing the lateral feature size scaling."

In our recent blog 3D NAND Opens the Door for Monolithic 3D we discussed in detail the adoption of monolithic 3D for the next generations of NAND Flash. The trend was very popular subject of this year's IEDM and is nicely illustrated by this older chart:

In 2D-to-3D paradigm shift, challenges = opportunities



- By 2013, 3D NAND flash is going into commercialization. This will be the biggest paradigm shift in NVM business.
- To make 3D NAND flash happen, collaboration should be considered.



Source: Jungdai Choi, et al., (Samsung), VLSI 2011

And accordingly the updated ITRS 2012 present the change from dimension scaling to monolithic 3D scaling as presented in the following slide.

2012 Update: Non-Volatile Memory

Based on survey performed by Japan PIDS, completed in March 2012, together with market observations.

- Compared to 2011 Edition, half-pitch scaling is unchanged.
- Some revisions for FeRAM (cell size, switching charge density…).

NAND Flash														
Year of Production	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*
Uncontacted poly 1/2 pitch (nm)	20	18	17	15	14	13	12	11	10	9	8	8	8	8
Number of word lines in one NAND string	64	64	64	64	64	64	64	64	64	64	64	64	64	64
Dominant Cell type	FG	FG	FG/CT	FG/CT	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D
Maximum number of bits per chip (SLC/MLC)					128G / 256G	256G / 512G	256G / 512G	512G / 1T	512G / 1T	512G / 1T	1T / 2T	1T / 2T	1T / 2T	2T / 4T
Minimum array 1/2 pitch - F(nm) [15]					32nm	32nm	32nm	28nm	28nm	28nm	24nm	24nm	24nm	18nm
Number of 3D layers for array at minimum 1/2 array pitch [16]					8	16	32	32	64	64	98	98	98	128

ITRS Winter Public Conf/Dec. 5, 2012/Hsinchu, Taiwan 9

This year's IEDM brought up two of the driving forces behind the shift from dimensional scaling to monolithic 3D IC scaling, that we will detail below as #1 and #2.

The Current 2D-IC is Facing Escalating Challenges:

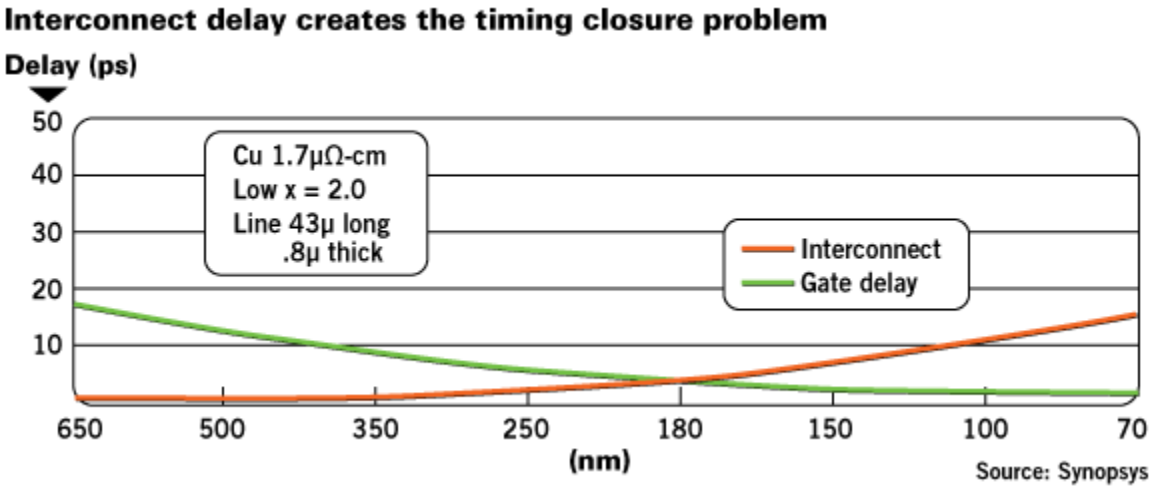
On-chip interconnect (#1)

- Dominates device power consumption
- Dominates device performance
- Penalizes device size and cost

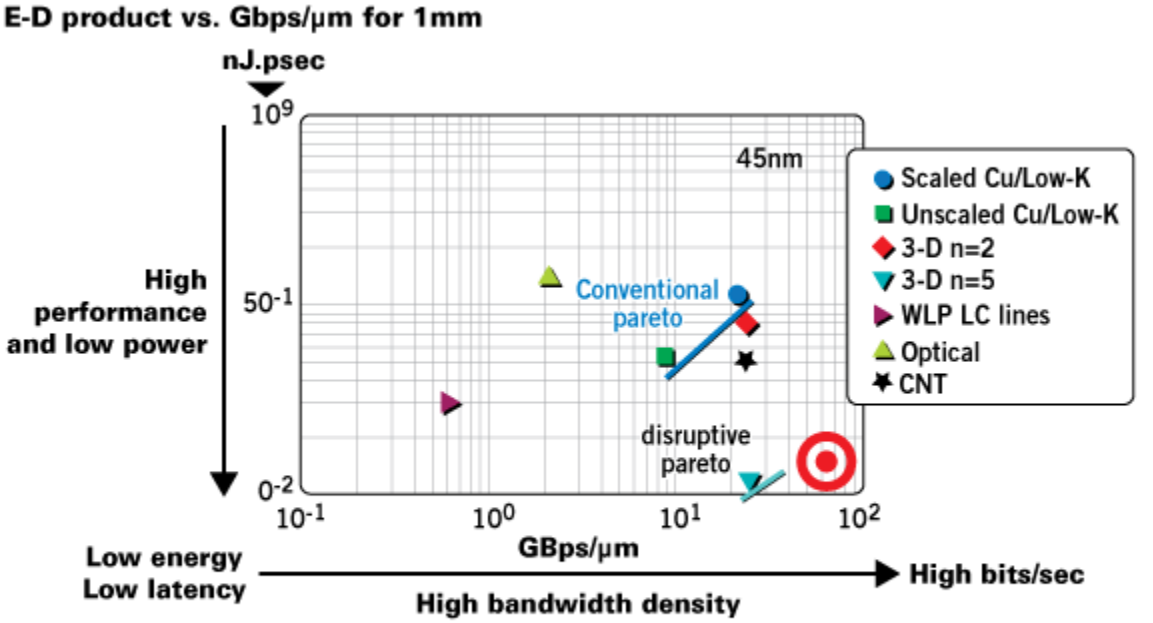
Lithography (#2)

- Dominates Fab cost
- Dominates device cost and diminishes scaling benefits
- Dominates device yield
- Dominates IC development costs

The problem with on-chip interconnect didn't start today. This vintage Synopsys slide below clearly indicates that on-chip interconnect started to dominate overall device performance a decade ago:



In response, the industry has spent an enormous amount of money to convert from aluminum to copper and to low-K inter-metal dielectrics. But now, we have very few additional options left (perhaps air-bridge?) as illustrated by the following chart:

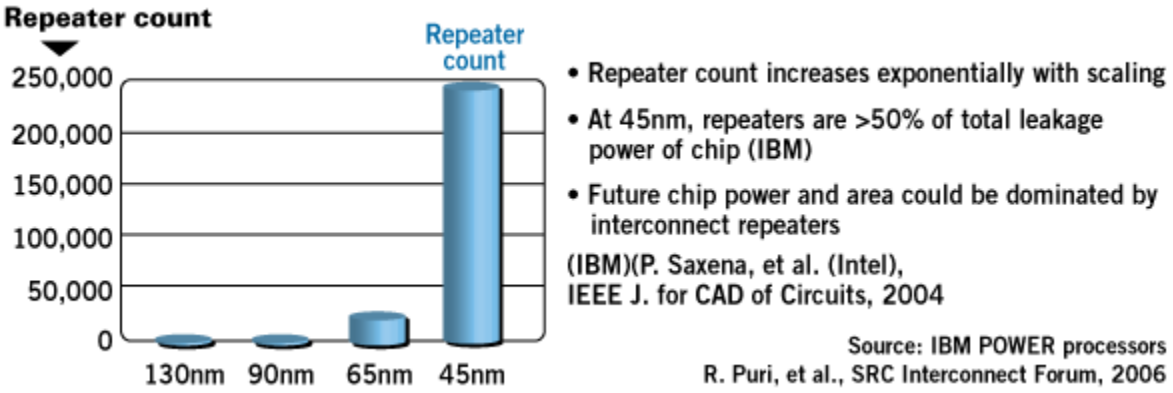


- 3-D with 5 strata clearly gives the highest bandwidth density for the lowest energy delay product.
- No other technology options give significant advantages over conventional scaled or unscaled Cu/Low k.

Source: Scott List, IMEC (M. Bamal, et al., IITC2006)

It shows that neither Carbon Nano Tube (CNT) nor Optical interconnect are better than copper, and that monolithic 3D still is the best path.

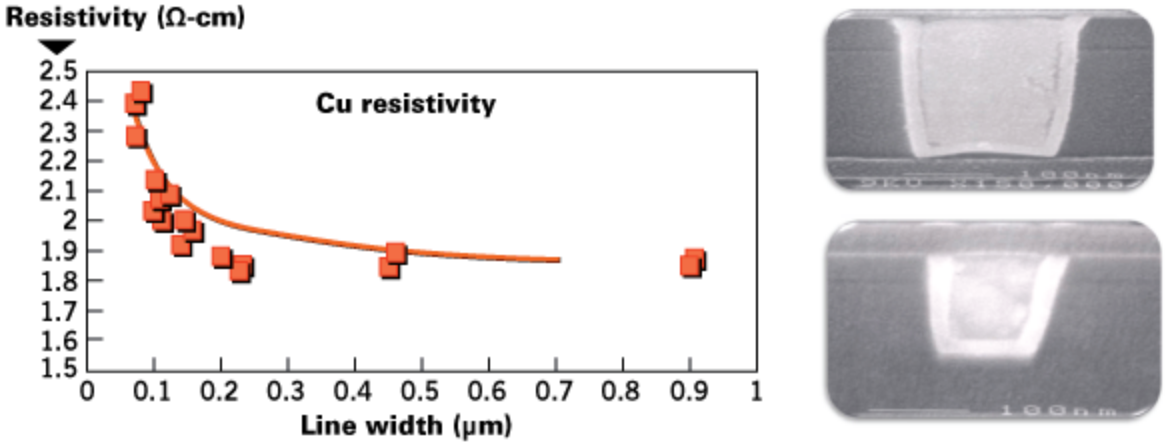
The practiced 'band-aid' fix so far has been throwing more transistors (they are getting cheaper, right? No longer. See father below) at the problem in the form of buffer and repeaters. But as we scale down we need exponentially more of these ban-aids as illustrated by the following:



Copper, however, is now reaching its inflection point as was articulated in a special session organized by Applied Materials attached to this IEDM, The 14 nanometer node is expected to be an inflection point. Quoting from the abstract:

"The 14 nanometer node is expected to be an inflection point for the chip industry, beyond which the resistivity of copper interconnects will increase exponentially and may become a limiting factor in chip design. On December 11, 2012, Applied Materials, Inc. will host an important forum in San Francisco to explore the path that interconnect technology must take to keep pace with transistor scaling and the transition to new 3D architectures." (emphasis added)

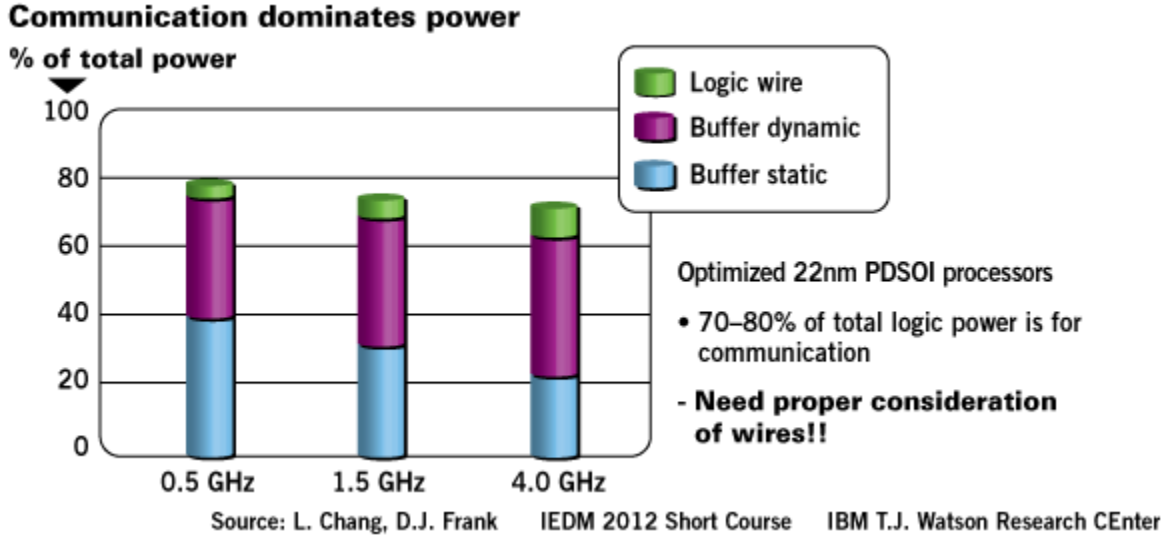
This had been illustrated before in the following chart



- Barrier layer takes a big part of the conducting area - increase resistivity
- Copper grain effect significantly increase resistivity

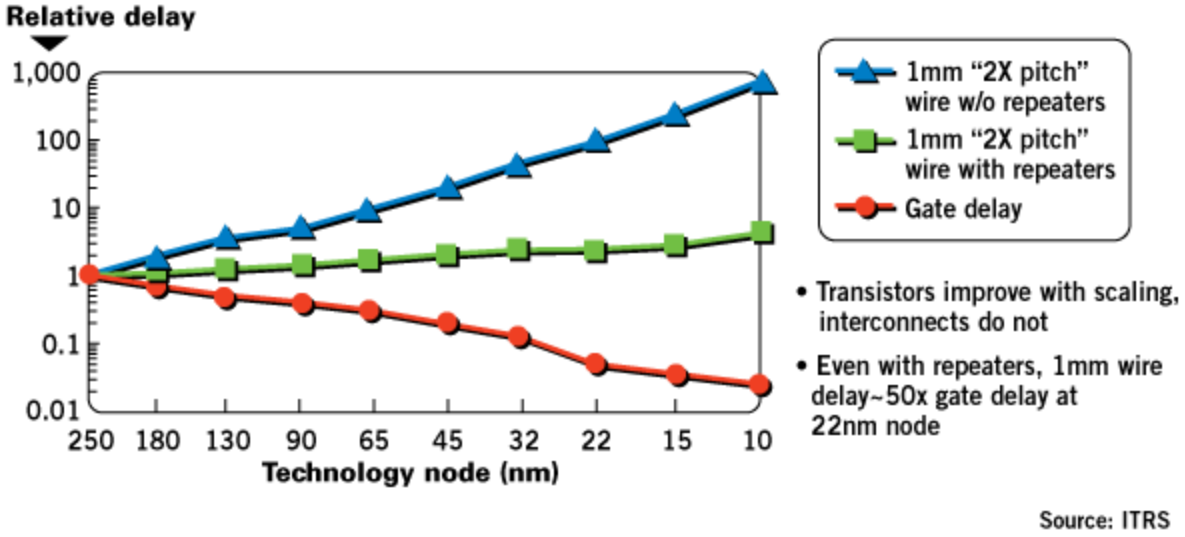
Source: Sam Naffziger, AMD Fellow at 2011 VLSI Symposium Keynote

And to make it crystal clear, IBM presented the following chart in its short course:

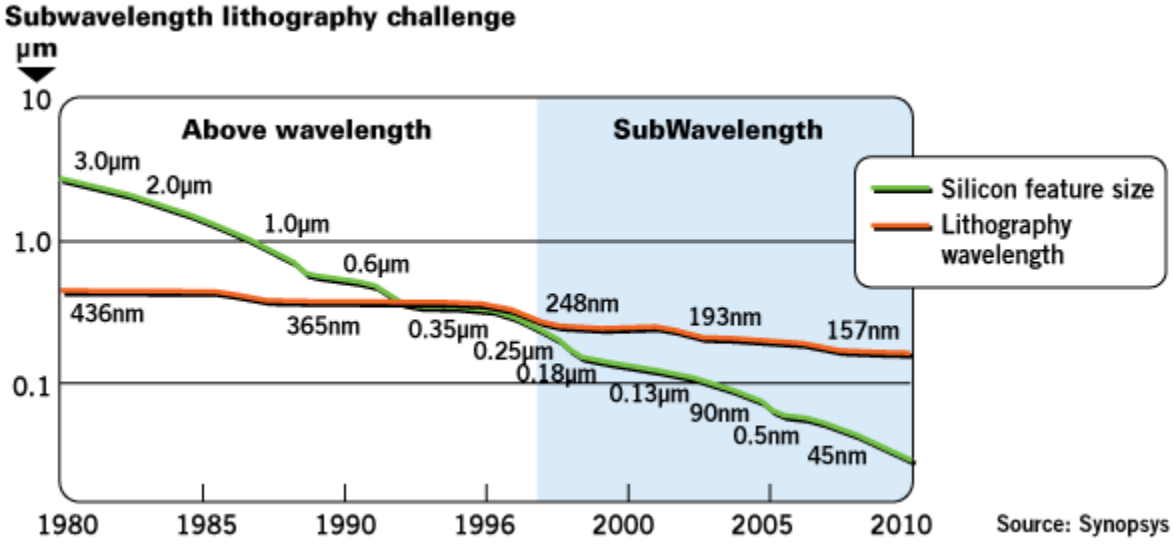


Source: L. Chang, D.J. Frank IEDM 2012 Short Course IBM T.J. Watson Research CEnter

Power is now dominating IC design and clearly dimensional scaling does not improve the interconnect's impact – see the following chart built from the ITRS Roadmap. The only effective path forward that addresses interconnect is monolithic 3D.

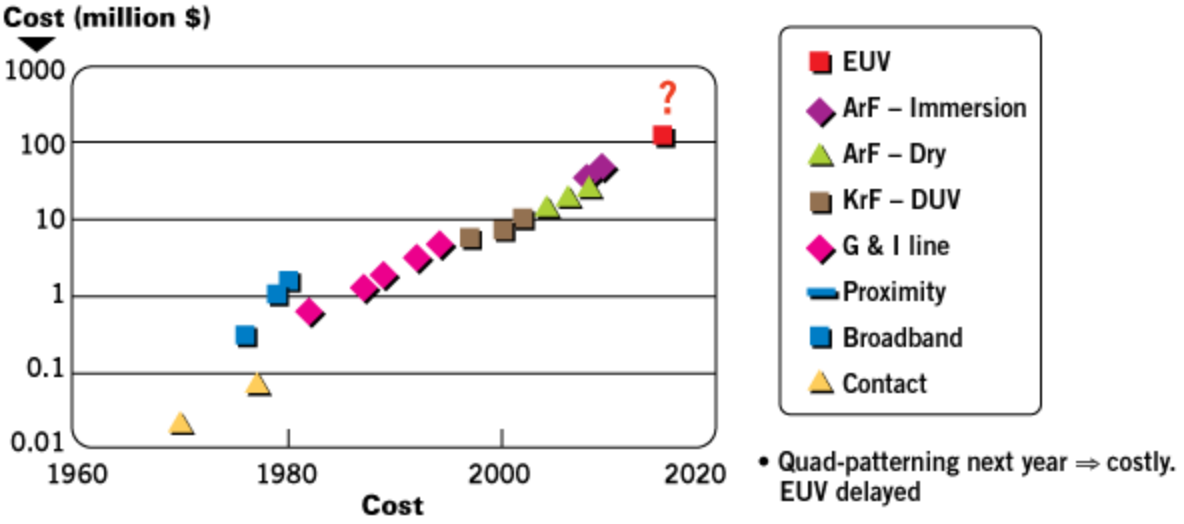


As for the second challenge – lithography – we start again with an old chart by Synopsys:

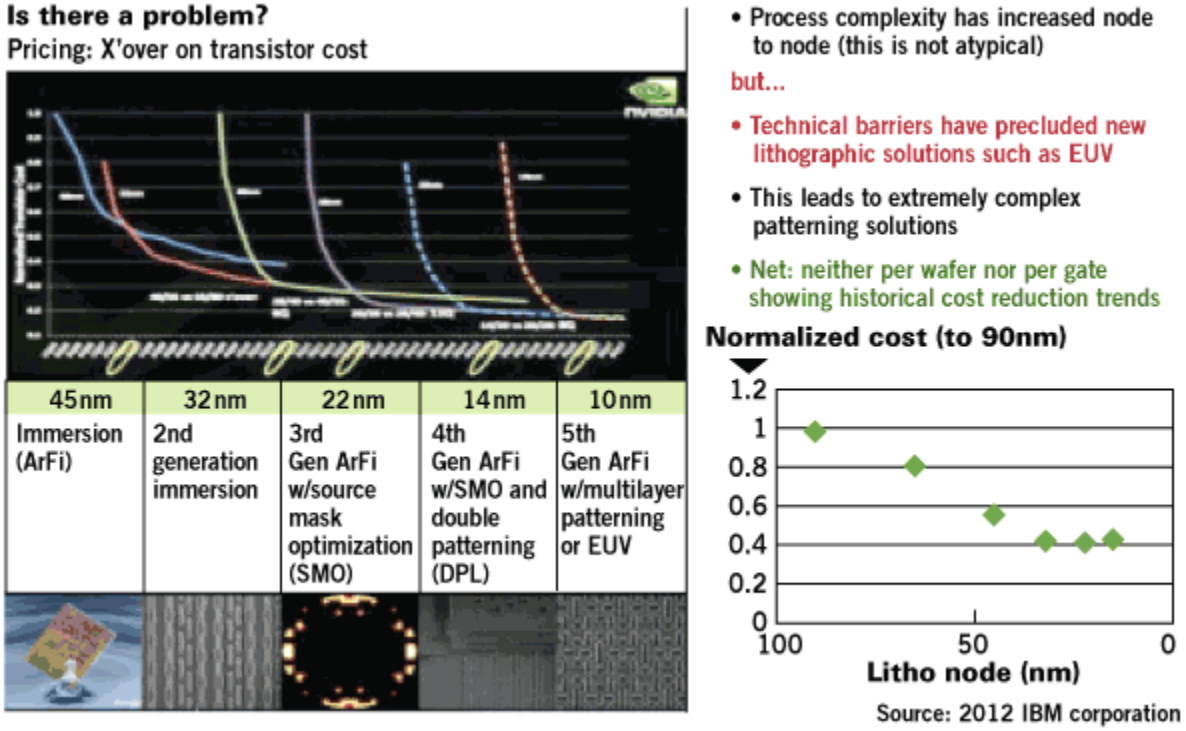


The implication is that any new node of dimensional scaling comes with escalating lithography costs; and sure enough, that's what is happening. When litho costs are plotted over time, it fits a log-linear scale….this is not a sustainable trend.

The following chart illustrates the lithography escalating cost of equipment which directly reflect the wafer cost.



This resulted in the following slide by IBM at the GSA Silicon Summit 2012:

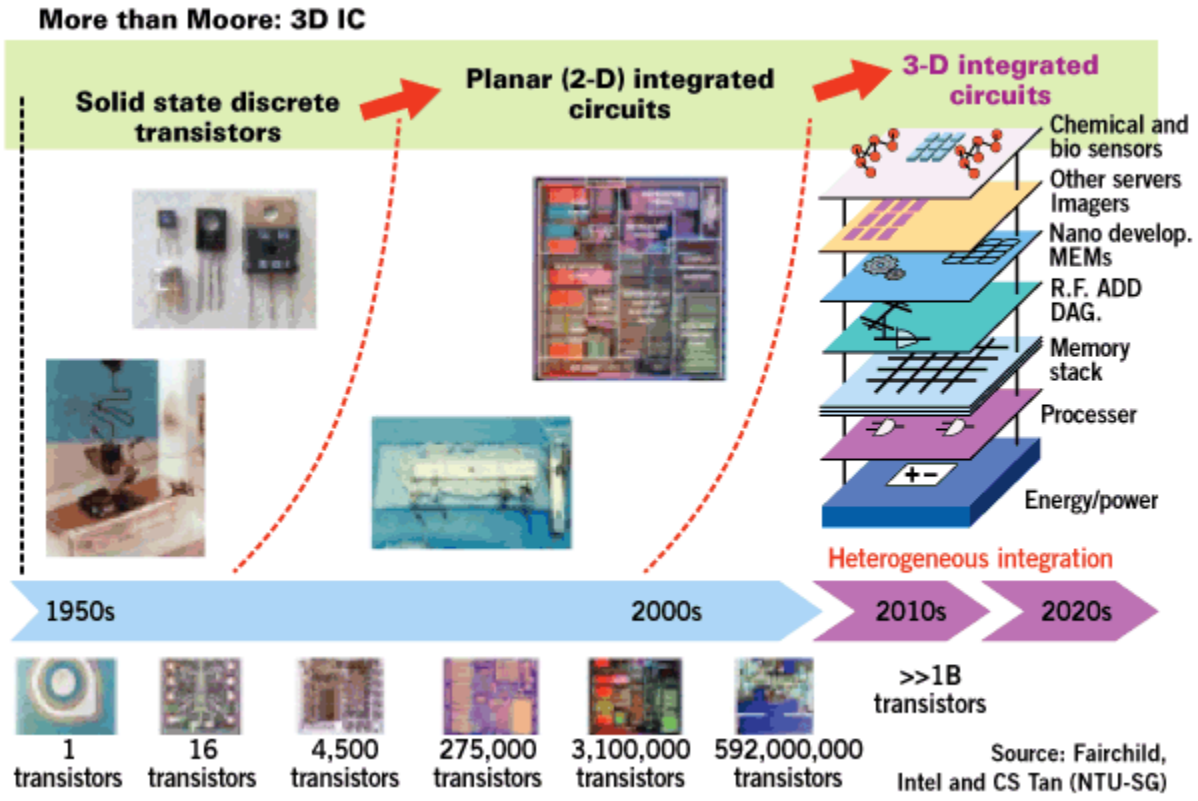


Quoting from the slide: "Net: neither per wafer nor per gate [are] showing historical cost reduction trends"

Another EE Times IEDM12 article covering a keynote given by Luc van den Hove, chief executive of IMEC, IEDM: Moore's Law seen hitting big bump at 14 nm, repeats the same conclusion. In fact, some vendors are already changing course accordingly. GlobalFoundries, in its recent 14nm announcement, disclosed that the back-end will be unchanged from 20nm. This suggests a similar die size and respective

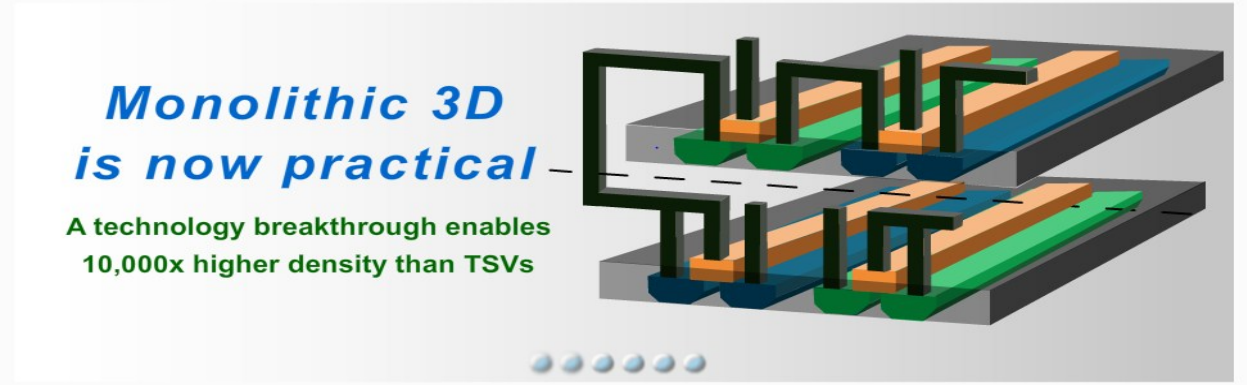
increase in per-transistor cost. Further, ST Micro in the Fully Depleted Transistors Technology Symposium (11 December, 2012) during IEDM12 week also acknowledged that their 14nm node will have a 20nm node metal pitch, and, just like GlobalFoundries, a similar die size and increase in per-transistor cost. So it would seem that also for lithographic reasons, the industry's next generation path, and the continuation of Moore's Law, would be achieved by leveraging the third dimension.

Now that monolithic 3D is feasible and practical, the time has come to move in this new direction, as has been nicely illustrated by this concluding chart below



Chapter 3 - The Monolithic 3D Advantage

by Zvi Or-Bach, the President and CEO of MonolithIC 3D Inc.



1. Introduction

Over the last 50 years we have seen tremendous technological and economic progress in semiconductors and microelectronics following what is known as Moore's Law. Accordingly about every two years the amount of transistors we can integrate on an IC doubles. This exponential increase in integration is achieved by scaling down the dimensions of the microcircuit by a factor of 0.7 at every technology node. For most of that half-century the scaling was relatively easy and was associated with about a 30% reduction of the transistor cost, a greatly improved performance, and markedly reduced power consumption. For most of us who have lived and worked this scaling - 'those were the days!'

However, recently the trend has changed dramatically, and it is now harder and harder (technically and economically) to achieve dimensional scaling; and as a result, there are diminishing improvements in transistor costs, power or performance. We discuss many of the details on our blogs:

IEDM: Moore's Law seen hitting big bump at 14 nm

Is the Cost Reduction Associated with Scaling Over?

Entanglement Squared

IEDM 2012 - The Pivotal Point for Monolithic 3D IC

A new form of scaling is shaping up as an alternative to maintain the exponential increase in integration. This new form is scaling up using monolithic 3D technology. The NAND Flash vendors are the early adopters of this new alternative scaling with multiple variations of products being developed that are scheduled to reach volume production in 2015.

In the following we will present "The Monolithic 3D" advantage. It is possible that this new technology could return us to the trend we had enjoyed before with reductions of cost, decreases in power consumption, and improvements in performance, and bring some new and compelling benefits.

Specifically, these are:

- Continuing reductions in die size and power
- Significant advantages for reusing the same fab line and design tools
- Heterogeneous Integration
- Processing multiple layers simultaneously, offering multiples of cost improvement
- Logic redundancy, allowing 100x integration at good yields
- Modular Platforms

2. Reduction in die size and power

A. Reduction in die size

Dimensional scaling has always been associated with increased wire resistivity and capacitance. Every node of dimensional scaling is associated with larger output drivers and more buffers and repeaters. The following charts illustrate the rapid increase of the number of transistors associated with the increased interconnect challenge.

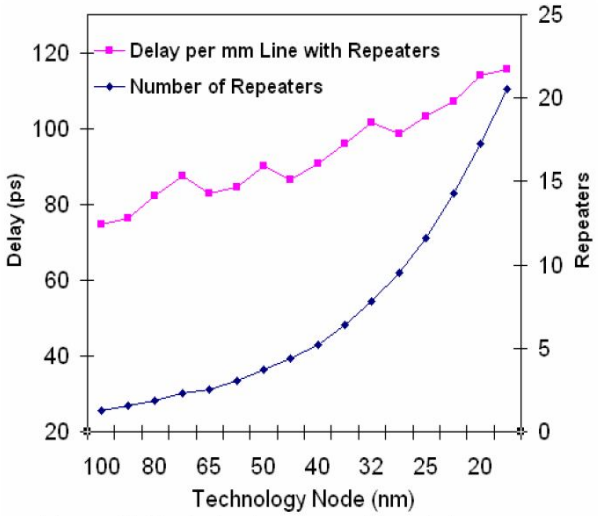
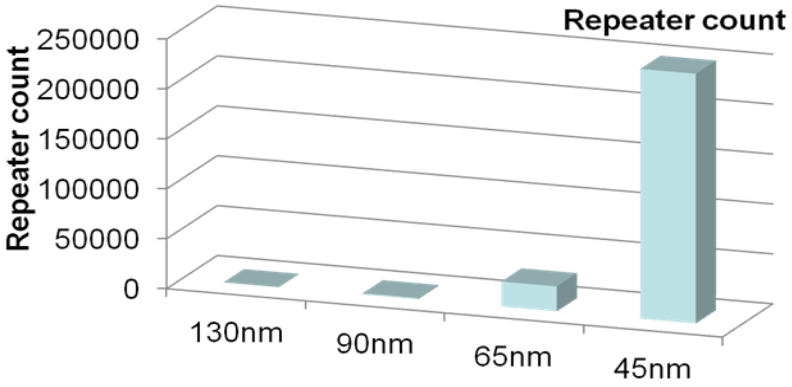


Figure 6. Global interconnection delay per mm length with repeaters.

Source: ISQED07 Alam

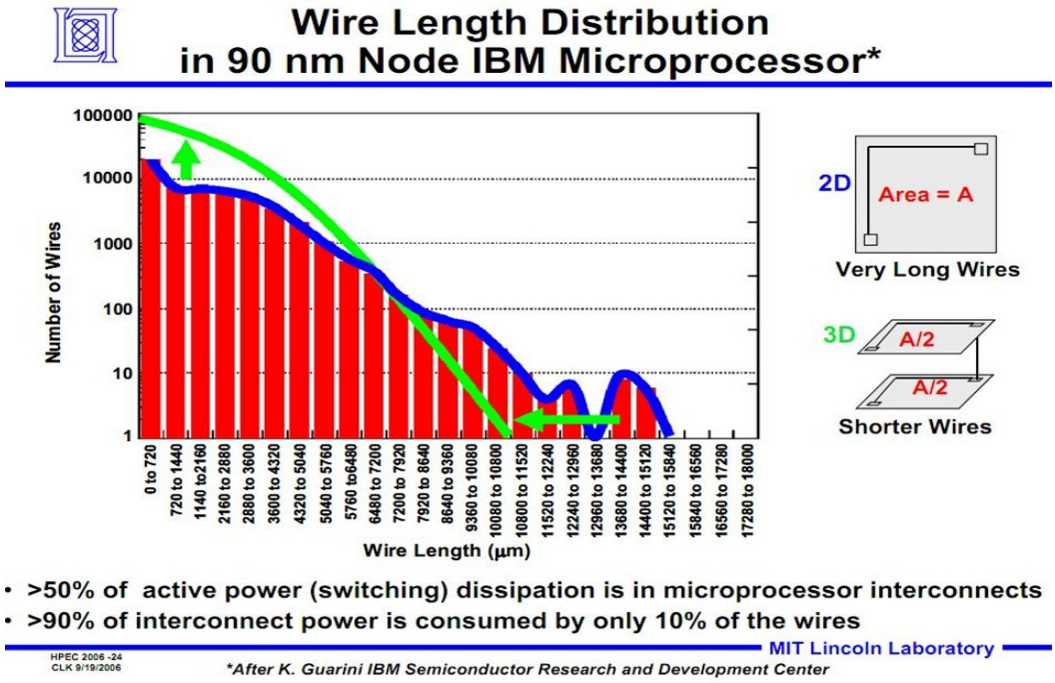


Source: IBM POWER processors
R. Puri, et al., SRC Interconnect Forum, 2006

- ***Repeater count increases exponentially with scaling***
- ***At 45nm, repeaters are >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, 2004]

Monolithic 3D enables the folding of a circuit, with the each stratum only about 1µ above or below its neighbor, combined with a very rich vertical connectivity between the strata. The following IBM/MIT slide illustrates the effectiveness of such a folding.



Further, the reduced silicon area generates an additional reduction of buffers and the average transistor size. MonolithIC 3D Inc. released an open-source high level simulator IntSim v2.0 to simulate a given design's expected size and power based on process parameters and the number of strata. More than 400 copies have been downloaded so far.

Using the simulator we can see in the following table that a 2D design of 50 mm2 area with an average gate size of 6 W/L, will only need an average gate size of 3 W/L and accordingly only 24 mm2 of total circuit area if folded into two strata (the footprint will be therefore just 12 mm2).

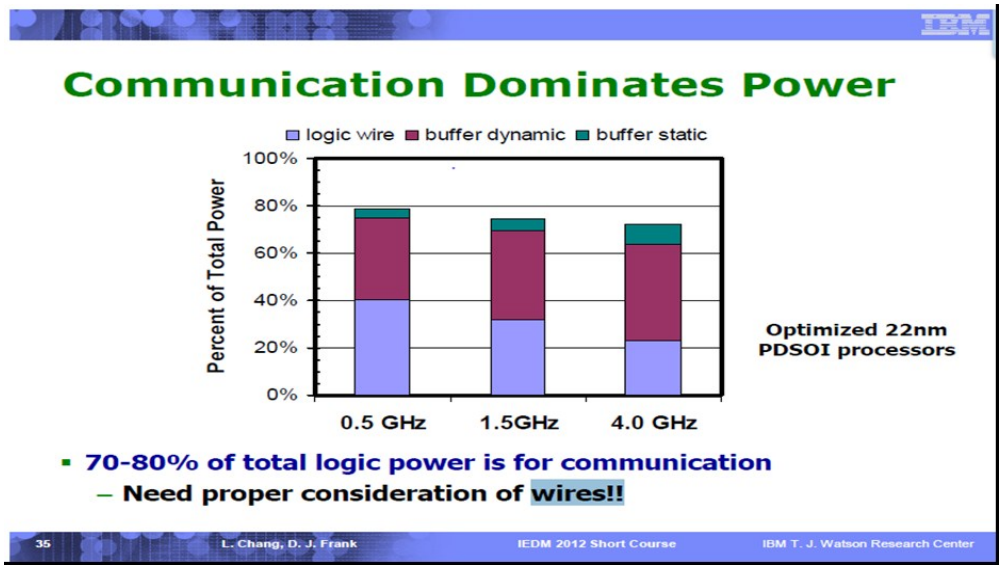
22nm node 600MHz logic core	2D-IC	3D-IC 2 Device Layers	Comments
Metal Levels	10	10	
Average Wire Length	6um	3.1um	
Av. Gate Size	6 W/L	3 W/L	Since less wire cap. to drive
Die Size (active silicon area)	50mm^2	24mm^2	3D-IC → Shorter wires → smaller gates → lower die area → wires even shorter 3D-IC footprint = 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	

These results are in-line with many other monolithic 3D research results.

=> Monolithic 3D 'folding' reduces the device silicon size by ~50% and leads to a similar reduction in transistor cost.

B. Reduction in power

The following chart illustrates that interconnect is now dominating the device power.



=>As every 'folding' effectively reduces the average wire length by about 50% it results in reducing the average power by 50%.

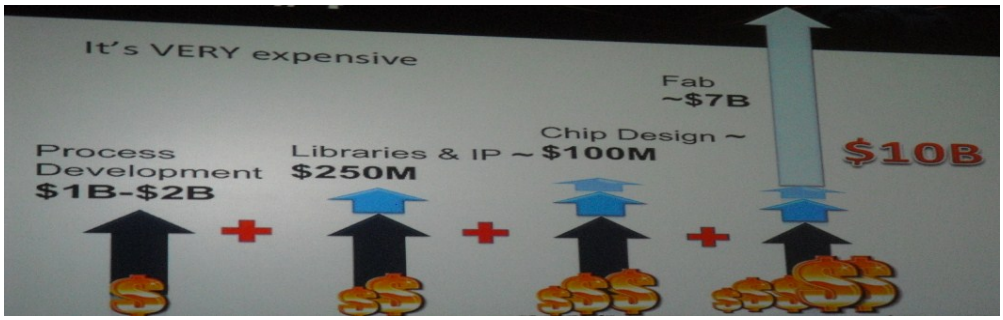
(Note: This assumes a proportional increase in complexity, which the industry has consistently done)

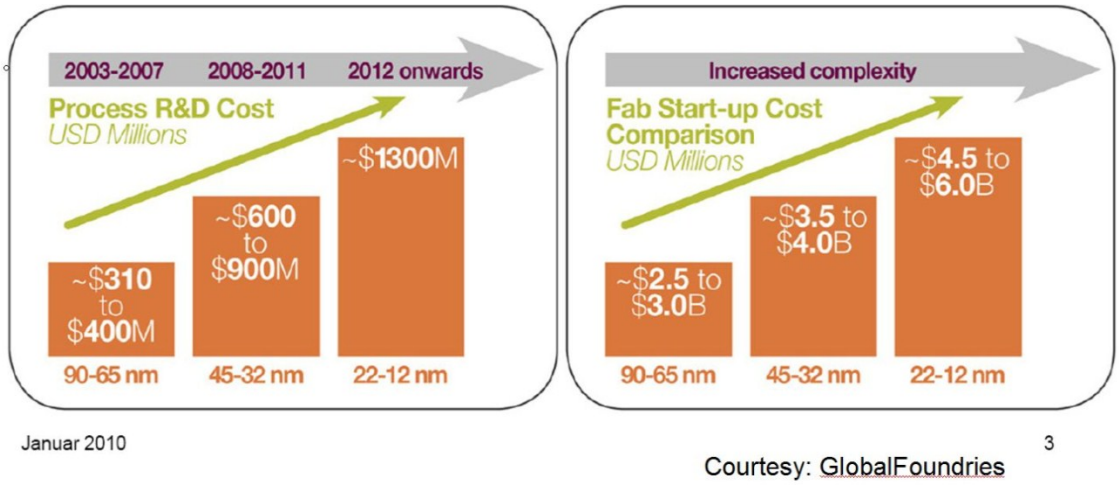
3. Significant advantages for using the same fab and design tools

A. Depreciation

With dimensional scaling every technology/process node requires a significant capital investment for new processing equipment, significant R&D spending for new transistor

process and device development, and the building of an ever more complex and costly library and EDA flow. The following charts illustrate this escalating cost trend:





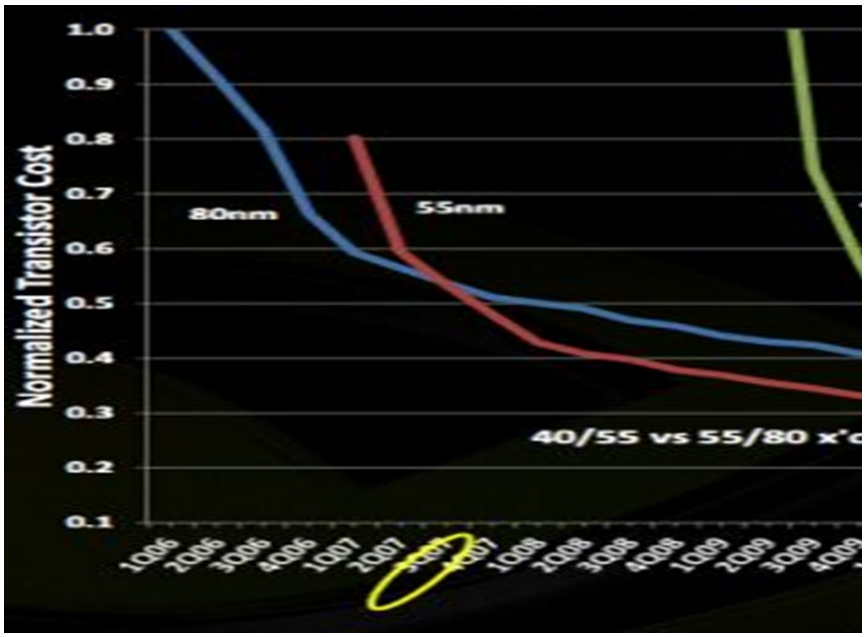
Januar 2010 3

Courtesy: GlobalFoundries

With monolithic 3D these costs are not required as dimensions are maintained for multiple generations and only the number of strata or layers is increased.

If the industry could use the same equipment and the same transistors and libraries for 4 years instead of 2, then all these costs could be depreciated over a longer time, with resulting significant cost benefits.

The following chart portion demonstrates the reduction of transistor cost per node as yield improves and equipment cost depreciates

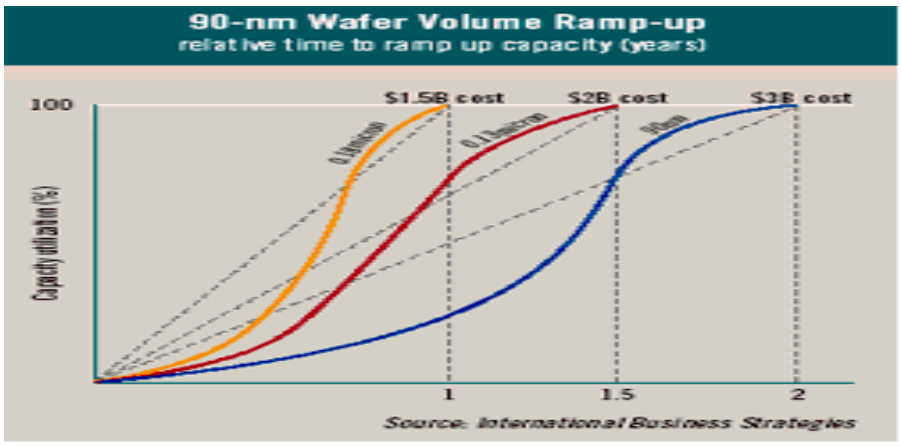


B. Learning Curve - Yield

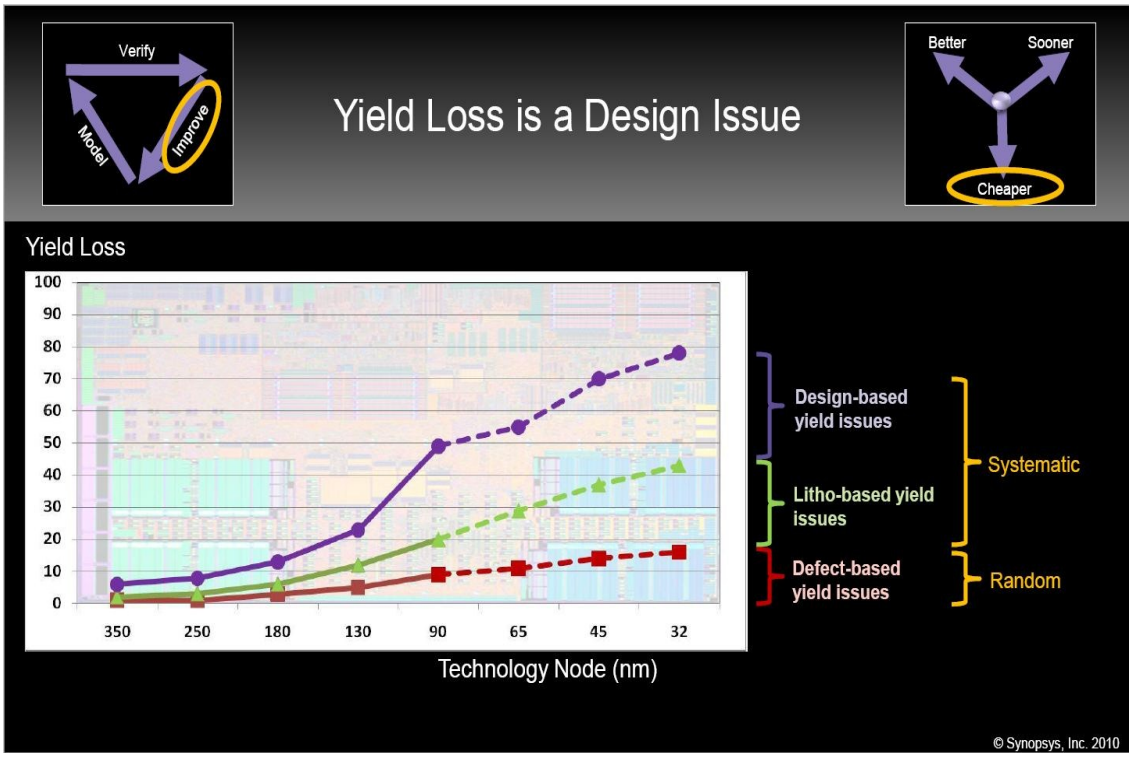
Using the same transistor tools and EDA has an additional important benefit. Learning curve equals yield improvement. With dimensional scaling we face the predicament that by the time we know how to manufacture a process node well, that learning quickly becomes obsolete as we are quickly moving on to the next node.

With monolithic 3D, the learning of the previous node stacking is directly utilized on the integration development of more strata, rather than on new materials, design tool issues, etc.

The following chart illustrates the dimensional scaling trend:



Each node of scaling is taking longer and costing more to get to mature yield ('ramped-up')



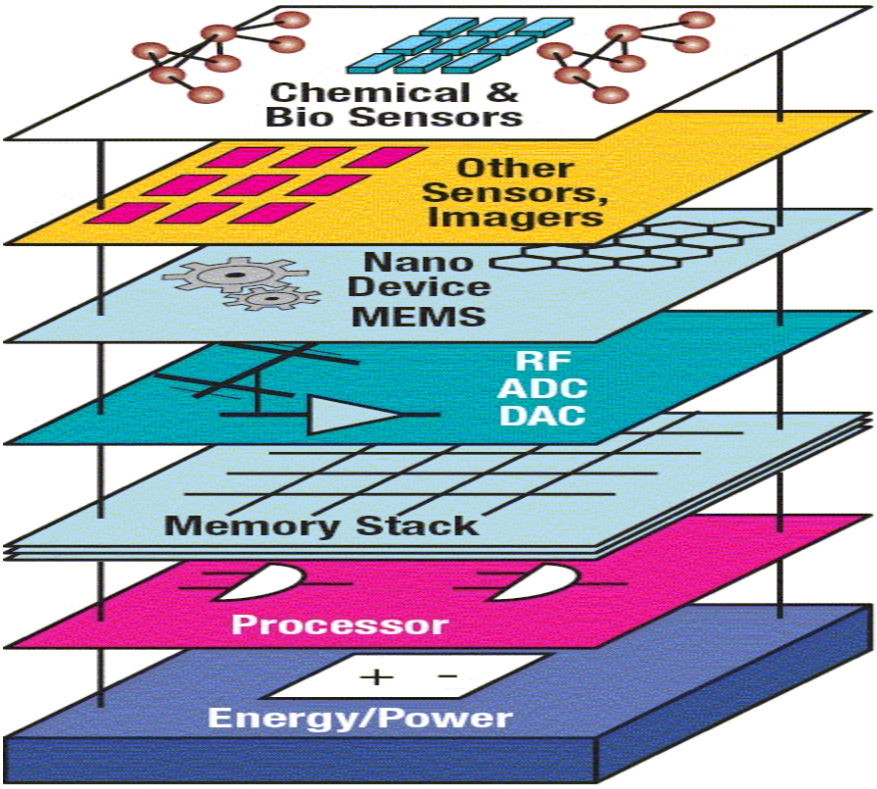
The design and litho based yield loss is growing quickly as the technology node gets dimensionally smaller.

4. Heterogeneous Integration

3D IC enables far more than an alternative for increased integration. It provides another dimension of design flexibility.

A well-known aspect of this flexibility is the ability to split the design into layers which could be processed and operated independently, and still be tightly interconnected - especially for monolithic 3D.

The following figure illustrates the ability to use different substrate crystal and different type of devices in such a heterogeneous integration.



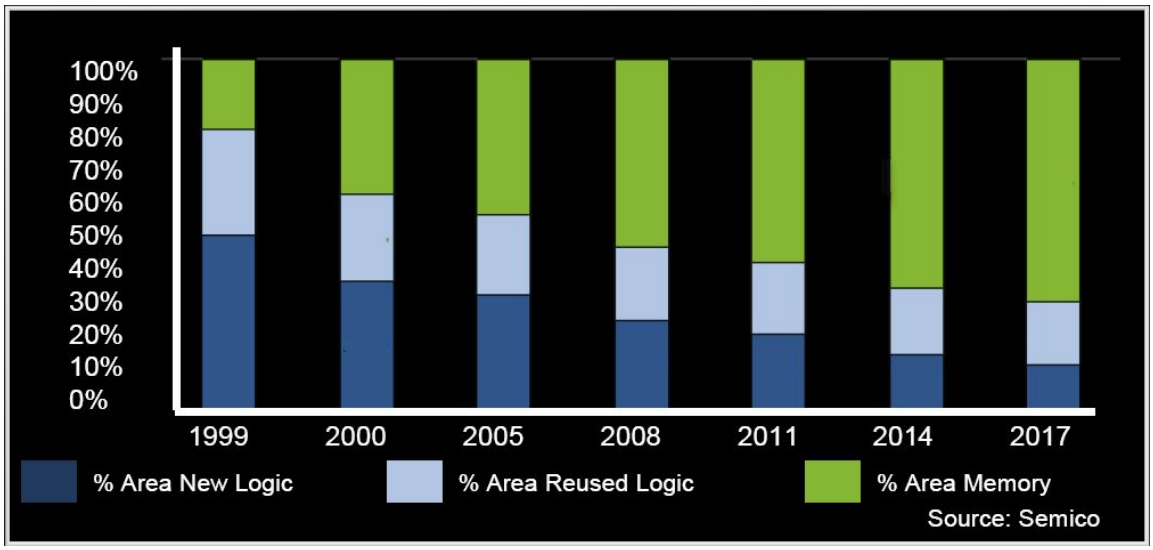
A. Logic, Memory, IO

Let's start with quoting Mark Bohr, in charge of Intel's process development:

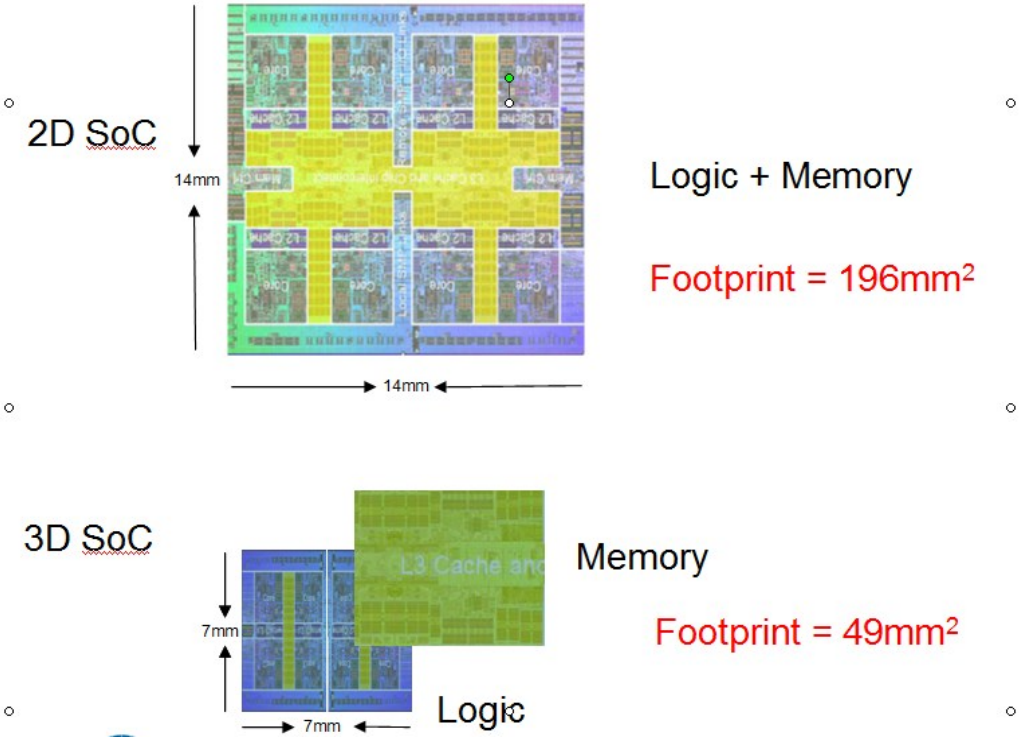
"Bohr: One important perspective is that chip technology is becoming more heterogeneous. If you go back 10 or 20 years ago, it was homogenous. There was a CMOS transistor, it was the same materials for NMOS and PMOS, maybe different dopant atoms, and that basic CMOS transistor fit the needs of both memory and logic. Going forward we'll see chips and 3D packages that combine more heterogeneous elements, different materials, and maybe transistors with very different structures whether they're for logic or memory or analog. **Combining these very different devices onto one chip or into a 3D stack—that's what we'll see**. It will be heterogeneous integration"

The most important market for semiconductor products is smart mobility. For this market the SoC device needs to integrate many functions, such as logic, memory, and analog. In most cases the pure high-performance logic would be about 25% of the die

area, 50% of the area would be memory, and the rest would be analog functions such as I/O, RF, and sensors.



In 2D all the functions need to be processed together and bear the same manufacturing costs. In a monolithic 3D-IC stack using heterogeneous integration each stratum is processed in an optimized flow, allowing for a significant cost reduction and no loss in optimized performance for each function type. The following illustration suggests the use of only two strata to build a device that in 2D would have a size of 196 mm2. By having one stratum for logic and one for memory, and by using DRAM instead of SRAM, the device could be reduced to 98 mm2 with footprint of 49 mm2. The device cost would be further reduced by the memory using only 3 or 4 metal layers. eDRAM on logic



B. Strata of Logic

The logic itself could be constructed better using heterogeneous integration. In many cases only portion of the logic need to be high performance while other portion could be better – and cheaper – done using older process node. Other scenarios could include designing different strata with different supply voltages for power savings, different number of metal interconnect layers, or other variations in the design space.

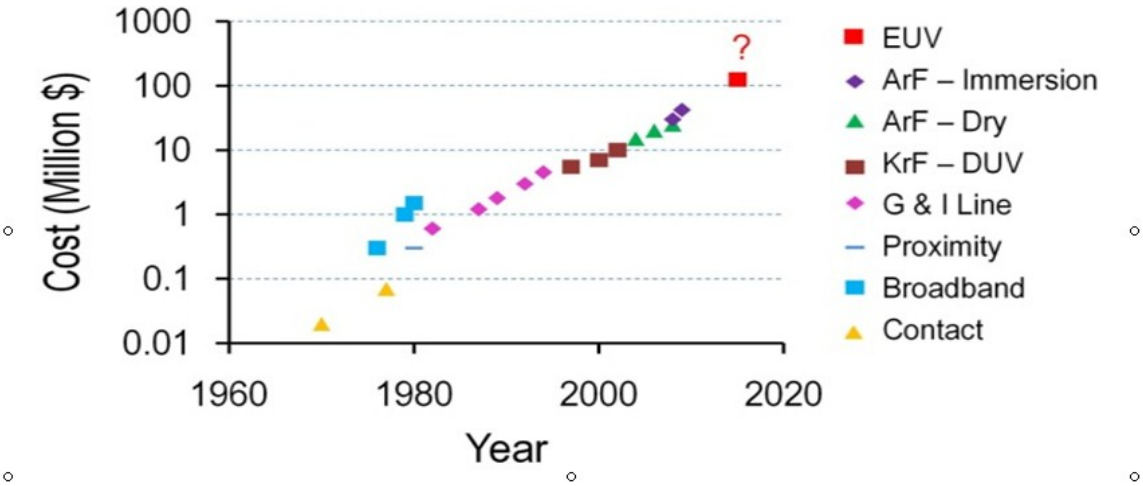
C. Strata of different substrate crystals and fabrication processes.

3D enabled heterogeneous integration could be used as illustrated in the beginning of the chapter. Some layers could utilize silicon while other might use compound semiconductors. Some layers could be image sensors or other type of electro-optic structures and so forth.

5. Multiple Layers Processed Simultaneously

An extremely powerful unique advantage of monolithic 3D is the option to process multiple layers in parallel following one lithography step. This option is most natural for regular circuits such as memory, but it is also available for logic circuits.

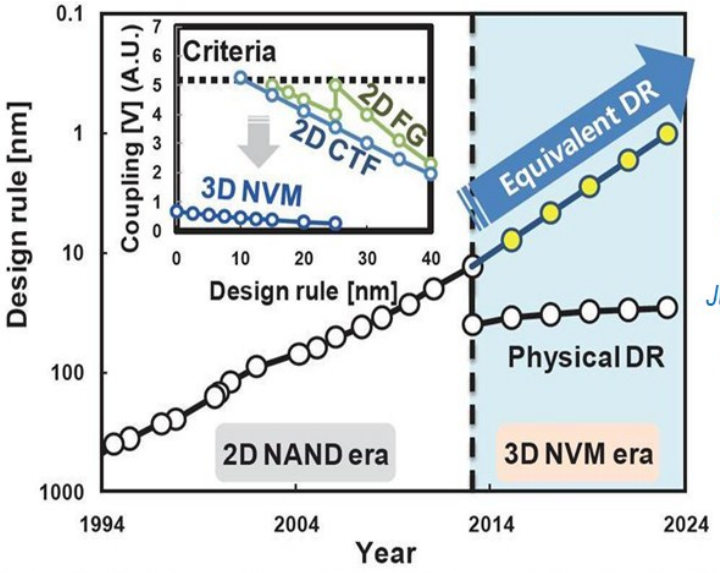
The driver for this option is the escalating costs of lithography in state of the art IC. The following illustration presents the impact of dimensional scaling on lithography costs.



- **Quad-patterning next year → costly. EUV delayed, costly.**

Currently the critical lithography steps dominate the end device production costs. Accordingly, if the critical lithography step could be used once for multiple layers rather than multiple times for each single layer, then the end device cost would roughly be reduced in proportion to the number of layers processed simultaneously.

The first merchants to recognize this option and who are moving to monolithic 3D are the NAND Flash vendors, as illustrated in the next figure.



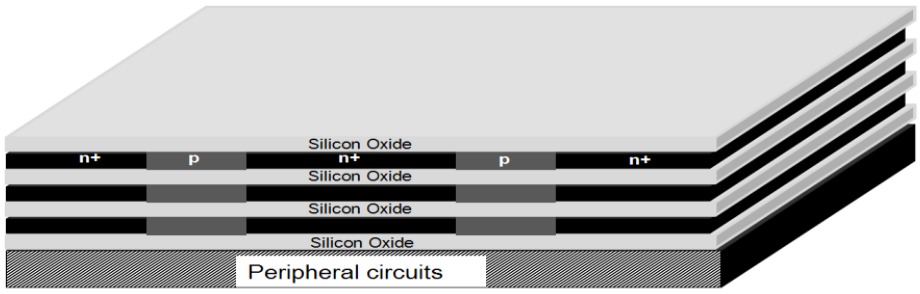
*2011 Symposium on VLSI Technology Digest of Technical Papers
Jungdal Choi and Kwang Soo Seol
Semiconductor R&D Center,
Samsung Electronics Co., Ltd.

The technical transition of NAND toward the 3D NVM era.

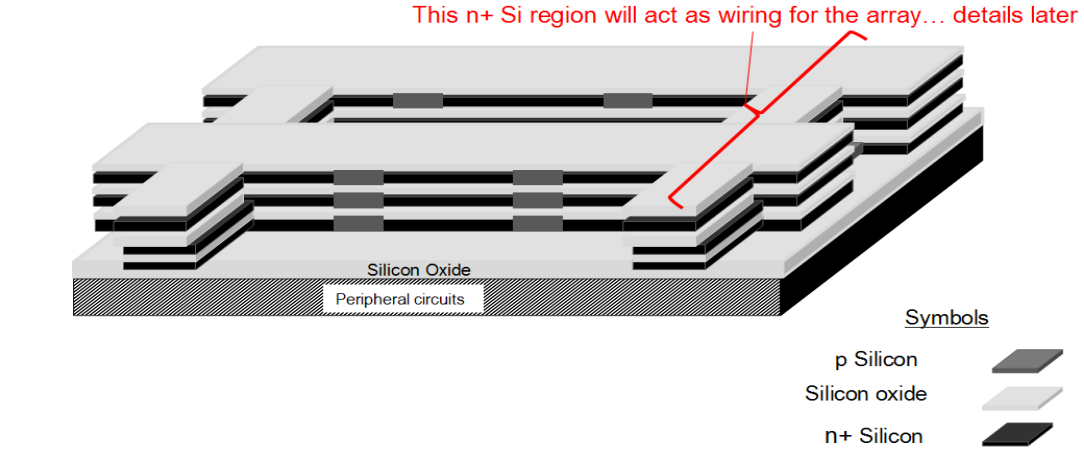
Using the proper architecture, multiple transistor layers could be processed together with a huge reduction in cost per layer. This could be applied to many different types of regular devices.

The following illustrates the concept applied to a floating-body DRAM:

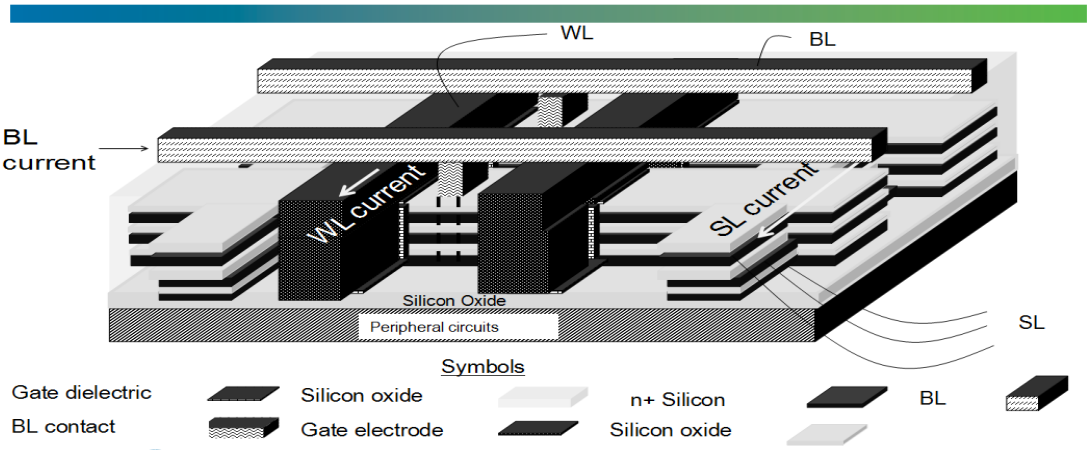
Process Flow: Step 6
Using methods similar to Steps 2-5, form multiple Si/SiO_2 layers, RTA



Process Flow: Step 7
Use lithography and etch to define Silicon regions



Process Flow: Step 11
Construct BLs, then contacts to BLs, WLs and SLs at edges of memory array using methods in [Tanaka, et al., VLSI 2007]



The MonolithIC 3D Inc. website presents more details for such a DRAM flow, and also related flows for RRAM and NAND Flash memories.

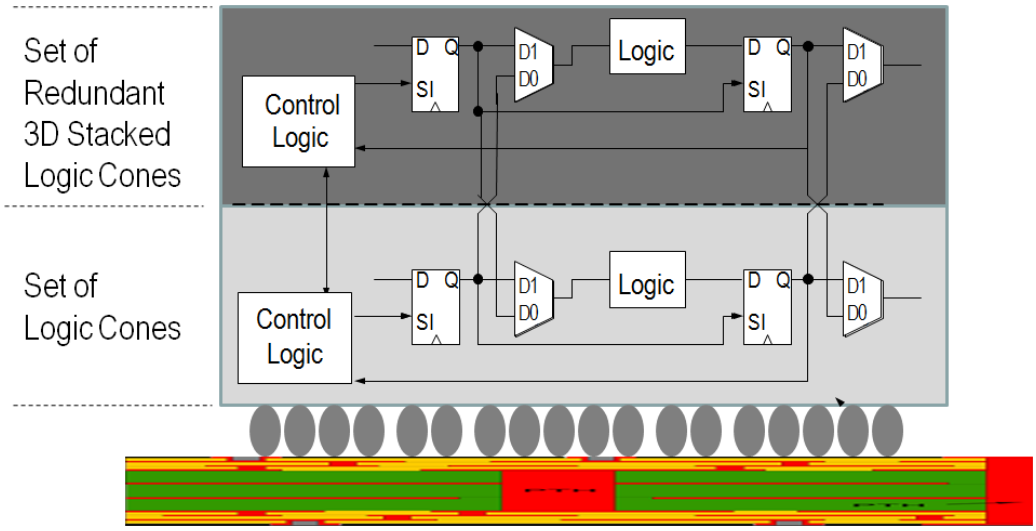
6. Logic redundancy allowing 100x integration with good yield

The strongest value of an IC is the integration of many functions in one device. This is and will be the most important driver of Moore's Law because by integrating functions into one IC we achieve orders of magnitude benefits in power, speed, and costs. At any given technology node the limiting factor to integration is yield. As yield relates strongly to device area, most vendors are trying to limit the die size to about 50mm²-100 mm². Some product applications require an extremely large die of over 600mm², but those are rare (and high value-add) cases because the yield goes down exponentially as die size grows.

While memory redundancy is prevalent in the IC industry, logic redundancy is only used in a few FPGAs – no solution has been found after the failure of Trilogy, where “Triple Modular Redundancy" was employed systematically. Every logic gate and every flip-flop were triplicated with binary two-out-of-three voting at each flip-flop. Quoting Gene Amdahl: “Wafer scale integration will only work with 99.99% yield, which won’t happen for 100 years.” (Source: Wikipedia)

An additional advantage of monolithic 3D is the ability to construct redundancy for circuits including logic, with minimal impact on the design process and while maintaining circuit performance.

The concept is illustrated in the following figure:



There are three primary ideas here:

- Swap at logic cone granularity.
- Redundant logic cone/block directly above, so no performance penalty.
- Negligible design effort, since the redundant layer is an exact copy.

The new concept leverages two important technology breakthroughs.

The first is the Scan Chain technology that enables a circuit test where faults are identified at the logic cone level. The second is the monolithic 3D IC which enables a fine-grained redundancy: replacement of a defective logic cone by the same logic cone that is only ~1 micron above.

Accordingly, by just building the same circuit twice, one on top of the other, with minimal overhead, every fault could be repaired by the replacement logic cone above. Such repair should have a negligible power penalty and a minimal cost penalty whenever the base circuit yield is about 50%. There should be almost no extra design cost and many additional benefits can be obtained.

This redundancy technique could be also used to repair faults throughout the device life-time, including in the field, which is a powerful advantage.

So the immediate question should be: how far can we go with such an approach?

A simple back-of-the-envelope calculation should start with the number of flip-flops in a modern design. In today's designs we expect more than one million F/F (and their logic cones). Consequently, if we expect one defect, then a device with redundancy layer would work unless the same cone is faulty on both layers, which probability-wise would be one in a million!

Clearly we have removed yield as a constraint to super-scale integration. We could even integrate 1,000 such devices!!!

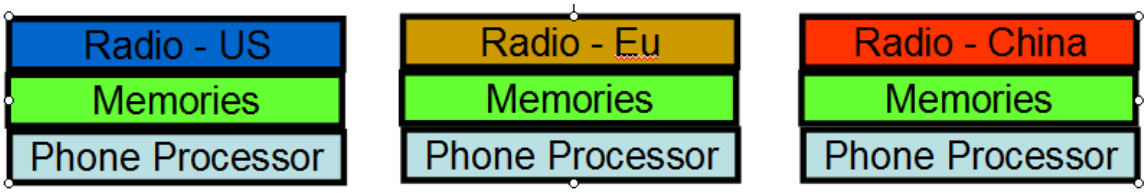
The ultra-integration value could be as much as:

- ~10X Advantage of 3D WSI vs. 2D @ Board Level
- ~10X Advantage of 3D WSI vs. 2D @ Rack Level
- ~10X Advantage of 3D WSI vs. 2D @ Server Farm Level

Overall, a ~1000x advantage is possible, all due to shorter wires. Instead of placing chips on different packages, boards and racks, we integrate on the same stacked chip.

7. Modular Platform

The 3D monolithic device would be a good fit to platform-based designs wherein some part of the device is used by all customers and others are tailored to a specific market/customer segment as illustrated by the following figure.



Such a system architecture could be inexpensively used in many market segments and with multiple variations. An interesting one could be in the FPGA sector where the same platform could come with many flavors of memories and I/O.

8. Stacked layers are naturally SOI

The upper layer or layers of monolithic 3D devices are naturally Silicon-On-Insulator (SOI). The advantages of SOI are well-established, increase with scaling, and include:

- 90% lower junction capacitance
- Near ideal sub-threshold swing
- Reduced device cross talk
- Lower junction leakage
- Effective back bias and multi-Vt options
- Multiple gate operation for superb electrostatic channel control

The recent developments of Fully Depleted SOI (FD-SOI) and SOI-FinFet has taken that advantage much further, and include:

- Lower manufacturing costs than bulk
- Less across-the-die transistor variation than bulk

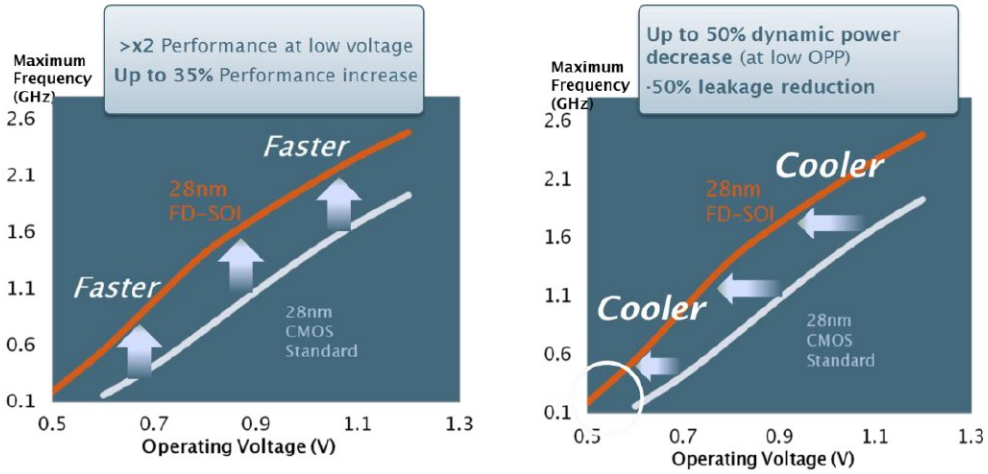


Figure 8: FD-SOI: faster and cooler

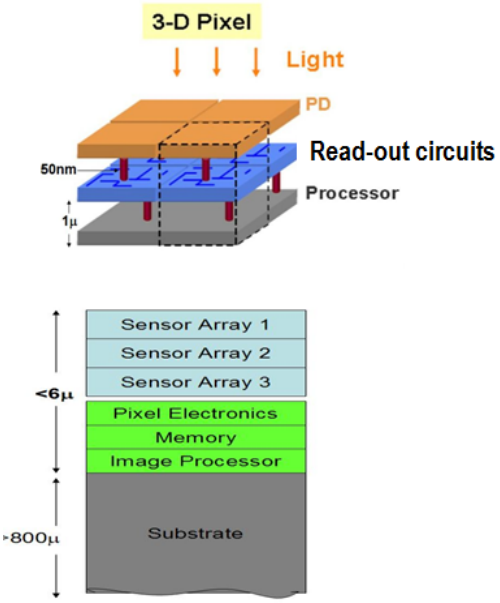
Source: ST-Ericsson < http://www.stericsson.com/technologies/FD-SOI-eQuad-white-paper.pdf>

9. Other ideas

There are other powerful advantages to monolithic 3D including those that we will discover in the future. In this chapter we present some specific applications where monolithic 3D provides significant advantages.

A. Image sensor with Pixel electronics

The image sensor industry has moved to back-side illumination to increase the image sensor area utilization. By adding the option for multiple layers many additional benefits could be gained as illustrated below:

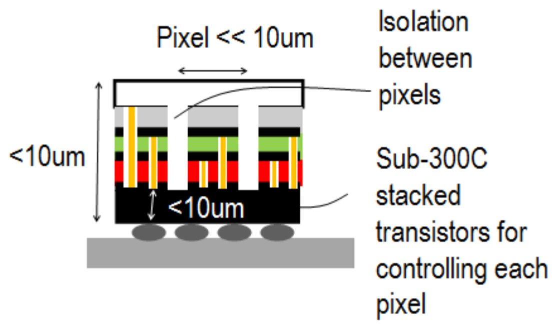


- Visible and infrared sensors integrated in a single stack → Day/Night capability in one stack.
- Multi-spectrum Imager
- Extremely high dynamic range
- Extremely high speed capture
- High resolution
- Dramatic reduction in power, size and cost

An interesting option is to build the pixel electronics behind every pixel and provide a very high dynamic range by counting and resetting individual sensors.

B. Micro-display

The display market is always looking to reduce power and size while increasing the resolution and brightness. Monolithic 3D could provide ultra-high resolution with extreme power efficiency and minimal size, by combining drive electronics with layers of different color light emitting diodes as is illustrated below.



Can control color of light using feedback circuits in silicon layer

- **A high-quality LED display without filters, polarizers, liquid crystals**
  - → **Avoids size and power penalty of these components**
  - → **1/10th power, much less weight than standard LCD display**
  - → **Brighter and more stable than OLED displays**
- **Can use as display, LED or communication device**

10. Summary

Monolithic 3D is a disruptive semiconductor technology. It builds on the existing infrastructure and know-how, and could bring to the high tech industry many more years of continuous progress. While it provides the advantages that dimensional scaling once provided, monolithic 3D offers many more options and benefits. And the best of all is that it could be done in conjunction with dimensional scaling.

Now that monolithic 3D is practical, **it is time to augment dimensional scaling with monolithic 3D-IC scaling.**

Chapter 4 - How can 3D be cheaper? Isn't it twice the cost?

by Brian Cronquist, VP of Techology and IP of MonolithIC 3D Inc.

An old CEO (John East of Actel) of mine kept drilling into our heads that “faster, cheaper, and easier to use” was the path to success in the IC industry, and that “Cheaper” was the key element of those three. Economics has always been a, maybe even **the**, key driver for scaling in specific [Moore's 1995 SPIE speech] and the industry in general.

So it was no surprise that when I have brought 3D-IC, and specifically monolithic 3D, into the potential solution space for combating the growing only-nations-can-afford-them costs of conventional (Dennard, etc.) scaling [IBM: Scaling dead], the first question out of their mouths (or keyboards) is: “Hey, doesn’t it cost twice as much to fold and stack it, so one gains nothing and perhaps even *loses* something due to the added costs of doing the 3D process (bonding, cleaving, connection)???” Well, when I started my monolithic 3D journey, I had the same first thoughts and questions. Here are a few of the answers….more will be forthcoming in future blogs and publications. Cost is a vast topic.

Die Size/Cost

“Hey, if I fold my chip over once, then I have the same silicon area (cost) as 2D but now double the processing, metal layers, etc., plus the costs of making and connecting the stack, right?” Well, that was my first impression too. But let’s take a deeper look. By placing about half the circuitry above the other half (i.e.: “folding”), not only do the long wires get shorter, but so does the average wire*. Hence, close to all of the logic gate to logic gate drivers and block to block buffers become smaller. Since they are smaller, then the circuitry moves closer to its neighbor; hence, the drivers can become smaller again. This *positive feedback mechanism* has been modeled by many people. Take a look at Davis, Zhou, and Synopsys, the references can be found at [Refs] as well as Meindl at MCISE 2003. This is a tractable problem for the universities, so there are many studies out there.

At Monolithic 3D Inc., we also have taken a close look at this to convince ourselves. Deepak took an older version of the IntSim tool he developed as part of his PhD thesis at Georgia Tech [Refs] and upgraded it to 3D. You will see more

publications on this tool and results, and it will soon be available on our website for you to try [3DSim]. Here is one result: The baseline was a 600MHz low power 2D logic core constructed at 22nm. A more complete description is at [Why Monolithic 3D], but the bottom line is: The monolithic 3D IC footprint is one fourth the die size of the 2D, and the total silicon area of the 3D chip is slightly less than half of the 2D chip (24 sq. mm vs 50 sq. mm).

22nm node	2D-IC	3D-IC 2 Device Layers	Comments
Frequency	600MHz	600MHz	
Metal Levels	10	10	
Average Wire Length	6um	3.1um	
Av. Gate Size	6 W/L	3 W/L	Since less wire capacitance to drive
Die Size (active silicon area)	50mm^2	24mm^2	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 capacitance to drive
	Total = 1.6W	**Total = 0.8W**	

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

Looks a lot like one nodal scale…..

But what about the added costs to bond, cleave, connect? See the next section…litho drives the wafer processing costs, mostly due to depreciation load. The strata to strata connect is only a 1 max 2 litho step adder to the 2x40+ total, using the same tools as a regular via.

Capital/Depreciation Cost

The majority of the cost of a die, assuming one is at yield maturity, is driven by the depreciation of the capital. The major capital cost of the modern wafer fab is the litho tools. And we all see the increasing costs and fears in this area Litho EETimes…100M$+ for a EUV machine EUV Cost. Also, as an old *fab-rat* and foundry guy, I can immediately relate to the fear many fab mangers and foundry execs have when they see the IBS trend [IBS 2010] and the ASML/AMAT price lists: How can I

keep up? Well, as explained by Israel a few blogs ago [Israel] by scaling UP with 3D and hence using the same litho tools, etc. to make the 3D stack, the wafer cost becomes much cheaper than the scale down wafer, the scale down wafer being subject to new litho tool process and depreciation cost. For monolithic 3D the only incremental capital would be for the wafer bonder/cleaving, implanter, and CMP machines, which are in the single digit M$ per machine costs, not the triple digit M$ per copy of NGL. The strata to strata via connect will look and act and process like a regular inter-metal via. We will be detailing this in upcoming publications, utilizing the Sematech COO framework.

Lots more to talk about (like lower mask costs), but I'll stop here for now. It's a big area. Ripe for many savings, and, like anything new, has the potential for unforeseen costs too….yes, yield and repair/redundancy mitigations will be a future subject. When you have a paradigm shift, as Zvi talked about last week [Zvi] and Deepak talked about in Monday's blog [Deepak], there can be many interesting opportunities to make chips faster, better, and cheaper….

What are YOUR questions and comments about 3DIC and cost? What do you think?

One more thinking question, especially for those who have *not* **lived** *in a wafer fab* (yup, I had a cot behind the diffusion furnaces when we started CSM Fab-1): With all the 'goodness" promised by 3DIC, doesn't it make sense to put 3D into the well-known and proven batch economics of the wafer fab?

*This is one of the key differences between TSV 3DIC and monolithic 3DIC: The long wires get shorter for both, but the remainder and greater number of wires only get shorter for the Monolithic 3D case due to vertical connectivity being approximately equal to horizontal connectivity.

Chapter 5 - Obtaining Monocrystalline Semiconductor Layers for Monolithic 3D

by Israel Beinglass, CTO of MonolithIC 3D Inc.

The idea of Monolithic 3D where several layers of devices are built, has been around for a few years. Many approaches were taken to create one or several layers of transistors on a completed first device (Transistor and interconnect).

Saraswat in www.jbkempf.com/~jb/Post-CMOS/Stanford/Saraswat.ppt depicted the concept of multilayer Monolithic 3D with several "transistor levels" (Figure 1). Furthermore, he suggested to 1.nucleate and 2.crystallize amorphous silicon, forming the second level of transistors followed by another set of interconnect layers (Figure 2). Nucleating and crystallizing amorphous silicon turns to be a very difficult task especially when the chip has millions of transistors per level and when elevating the temperature could be detrimental. An alternative way is using TFTs on the second level of silicon, the problem with that is of course very poor performance of the transistors even after crystallization of the amorphous silicon to polycrystalline materials, as well as the need to generate S/D junctions at high temperature that will adversely affect the underlying devices.

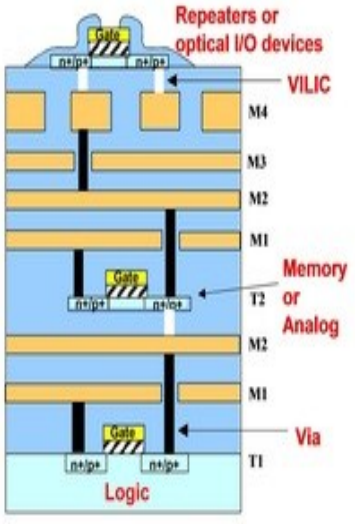


Figure 1

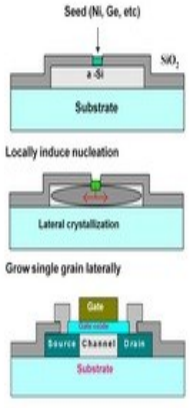


Figure 2

Another way which was suggested is low temperature Ge epitaxial over growth from "windows" in the silicon substrate and laser annealing the structure (Figure 3). This technology was developed by P. Griffin from Stanford and graduate students J. Feng,

M. Kobayashi and G. Thareja

(http://nanodevice.stanford.edu/3dworkshop/docs/8_Griffin-TEL3DWorkshopNov07.pdf). They reported some limited success on growing epitaxial Ge. However fully integrating the technology seems to run into insurmountable difficulties of process control, as well as integrating Ge transistors on a full advanced CMOS process.

The other approach is integrating thin layer transfer onto a fully processed wafer, by that creating a second level of transistors, followed by a set of interconnect layers. The layer transfer is similar to the "smart-cut" process described by SOITEC in http://www.soitec.com/pdf/SmartCut_WP.pdf (Figure 4).

Applying layer transfer technology along with MonolithIC 3D Inc.'s IP portfolio is a new

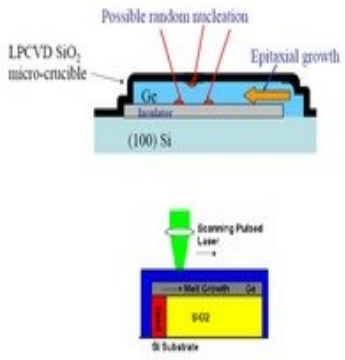


Figure 3

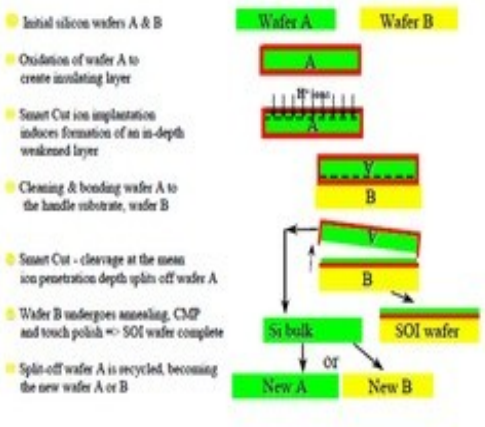


Figure 4

and fresh way to build the next generation of 3D device integration.

Chapter 6 - Low Temperature Cleaving

by Brian Cronquist, VP of Technology and IP of MonolithIC 3D Inc.

Thanks to everybody who came by our booth at SemiconWest SemiconWest 2012 this second year! We really enjoyed talking with you about all the exciting possibilities for new products and processes that are enabled by monolithic 3D IC.

For those who could not make it, here is what our booth looked like:

Nice tie again Zvi! You can still visit us at www.monolithic3d.com.

The most common area that you asked us was about low temperature (less than 400°C) bonding and low temperature cleaving processes. The two topics are quite inter-related: One must make the bond stronger than the energy it takes to cleave at the plane you want, rather than cleave at th at fresh bond. In October last year I wrote a blog about the many low temperature bonding techniques and strategies available and their respective bond strengths. Today, I would like to briefly address some of the low temperature cleaving methods available. Generally they involve either a mechanically induced (blade, gas jet, water jet) method, a lower temp thermal (co-implantation, microwave, etc.) cleaving/layer-transfer method, or a combination of both.

Here are a few papers, with some industrial announcements at the end.

One of the earliest methods published is co-implantation by Q.Y. Tong et al. of Duke University at the *1997 IEEE SOI Conference*. Tong could greatly affect the kinetics of the hydrogen blister formation by co-implantation of Boron. They were able to transfer a 0.4um silicon layer onto a quartz substrate with a 150°C exposure to the quartz by pre-annealing the co-implanted silicon for 10 minutes at 250°C.

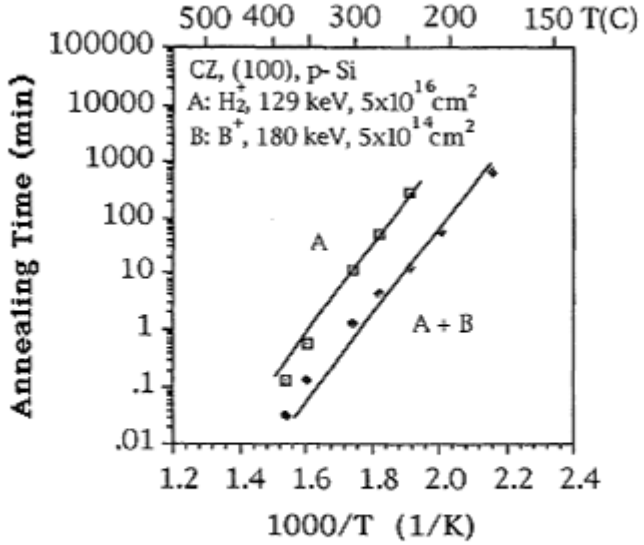


Fig.1 B+H co-implant effect on times required to form optically detectable surface blisters in hydrogen implanted Si as a function of annealing temperature

Tong with colleagues at the Max-Planck-Institute followed up with more co-implantation

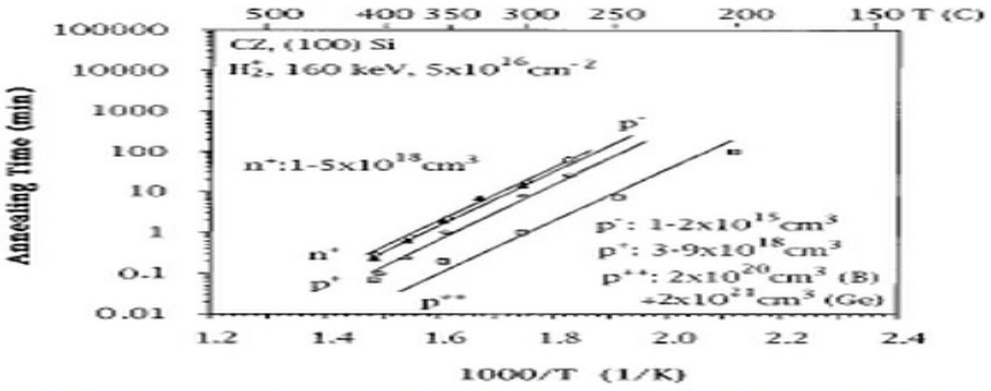


FIG. 1. Time required to form H implantation-induced optically detectable surface blisters on H-implanted, uniformly doped Si wafers (with phosphorous or boron concentrations as indicated) as a function of inverse absolute temperature.

kinetics data in a *2008 Applied Physics Letter*. They again demonstrated a 200°C silicon cleave.

In *1998 App. Phys. Lett.*, Agarwal et al. showed that He implanted with the H could lead to a significant decrease in the total implant fluence (and hence cost) necessary to achieve Si layer transfer. The total implantation dose can be three times smaller than that which is necessary using H alone.

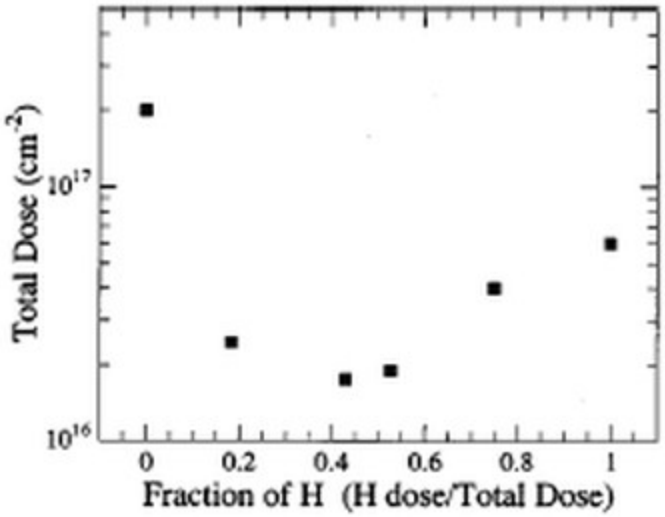


FIG. 2. Total (H^+ + He^+) implantation dose necessary for blistering as a function of the fraction that is H^+.

Nguyen et al. of Soitech/CEA-Leti, at the *2003 IEEE SOI Conference* showed that He co-implantation could be used to control the kinetics, so time, dose and temperature trades could be made.

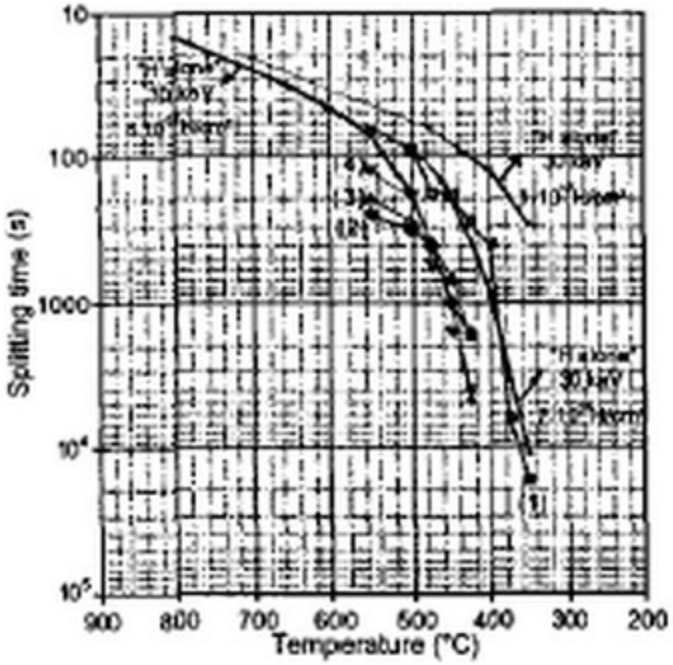


Fig. 2: Splitting kinetics in the various cases of H/He co-implantation (1, 2, 3, 4) compared with cases of H implanted alone

Ma, et al. showed in *Semcond Sci. Technol. 2006* that a co-implanted cleave has a smoother surface than a hydrogen-only implanted cleave.

Table 1. The RMS surface roughness of SOI fabricated by B^+/H^+ co-implantation and H^+-only implantation.

Dose and energy of implantation	RMS surface roughness
Co-implantation: B^+: 180 keV, 1×10^{15} cm^{-2} H^+: 60 keV, 4.5×10^{15} cm^{-2}	2.1 nm
H-only implantation: H^+: 140 keV, 6×10^{15} cm^{-2}	>10 nm

In *2000 App. Phys. Lett.*, Henttinen et.al showed mechanical cleaving, blade or N2 gas, on low temperature bonded silicon wafers (ox-ox bond). Depending on the H dose, Henttinen could

TABLE I. Effects of the hydrogen implantation dose and the bond annealing temperature on the mechanical exfoliation. The "+" sign denotes that the Si layer is successfully transferred, whereas the "−" sign indicates that the layer is not transferred. The bond annealing time is shown in parentheses.

	Implantation dose (H_2/cm^2)		
Bonding temperature	4.0×10^{16}	4.5×10^{16}	5.0×10^{16}
200 °C	−(2 h)	−(2 h)	+(2 h)
250 °C	−(2 h)	+(2 h)	+(30 min)
300 °C	+(2 h)	+(2 h)	+(2 h)

cleave the silicon wafers at 200°C or 300°C. Henttinen et.al followed up later in 2002 in *J. Nucl. Instr. and Meth. in Phys* with fundamental mechanistic studies and also demonstrated that with enough B doping one can enable H-implanted layer exfoliation below 200°C.

Cho et al., in *2003 App. Phys. Lett*. reported that full wafer layer transfer could be achieved with a mechanical cleave (edge initiated crack propagation) after a 250°C annealing that enabled the bonding strength at the acceptor/donor interface to exceed the required cleave energy at the hydrogen implant plane.

En, et al., of Silicon Genesis, described a room temperature H implant using PLAD (Plasma Immersion Ion Implantation), plasma assisted oxide to oxide bonding, and a room temperature mechanical cleave process at the *1998 IEEE SOI Conference*.

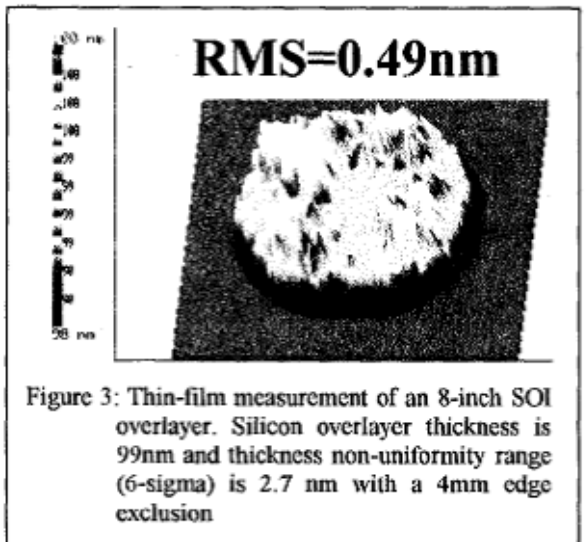


Figure 3: Thin-film measurement of an 8-inch SOI overlayer. Silicon overlayer thickness is 99nm and thickness non-uniformity range (6-sigma) is 2.7 nm with a 4mm edge exclusion

Table 1: PIII machine specifications

Spec Description	Spec Value
Wafer Size	4"-12"
Implant Time	60-120s (wafer size independent)
Area energy non-uniformity	+/- 5% range
Area dose non-uniformity	+/- 10% range

Table 2: Genesis Process SOI wafer specifications

Spec Description	Spec Value
Wafer Size	4"-12"
SOI layer thickness (tSOI)	50-250nm
tSOI uniformity	< 3nm range
Buried oxide thickness (tBOX)	No restrictions
tBOX uniformity	< 3% range
Surface roughness	< 1.5A R_{rms} (2x2 μm)

Current, et al. of Silicon Genesis, showed a wafer separation tool in *MRS 2001* where they utilized a pressurized N2 jet to cleave silicon bonded pairs at room temperature.

Recently from the industrial side:

Soitec announced at *SemiconWest 2012* the availability of a room temperature smart cut:

"Soitec's low-temperature Smart Cut process uses oxide-oxide molecular bonding and atomic-level cleaving to transfer mono-crystalline silicon films as thin as 0.1 micron onto partially or fully processed wafers. On this new material layer, a second level of devices can be processed and this integration can be repeated in an iterative mode. Transferring an extremely thin layer enables higher interconnect density, higher signal throughput and simpler TSV processing. Benefits include increased computing bandwidth, lower overall manufacturing cost, and power savings due to the reduced wiring distance between connected devices. This final benefit is well suited for producing advanced memory or CMOS logic 3D IC systems." See: http://www.soitec.com/en/news/press-releases/article-346/

SiGen (Silicon Genesis) has tools (some shown above) available that will bond and cleave at or near room temperature: http://www.sigen.net/semi_debondCleave.html

References:

TONG, Q.-Y., et al., "Low Temperature Si Layer Splitting", Proceedings 1997 IEEE International SOI Conference, Oct. 1997, pp. 126-127

TONG, Q.-Y., et al., "A ‘‘smarter-cut’’ approach to low temperature silicon layer transfer", Applied Physics Letters, Vol. 72, No. 1, 5 January 1998, pp. 49-51

AGARWAL, A., et al., "Efficient production of silicon-on-insulator films by co-implantation of He+ with H+'" Applied Physics Letters, vol. 72, no. 9, March 1998, pp. 1086-1088.

NGUYEN, P., et al., "Systematic study of the splitting kinetic of H/He co-implanted substrate", SOI Conference, 2003, pp. 132-134

MA, X., et al., "A high-quality SOI structure fabricated by low-temperature technology with B+/H+ co-implantation and plasma bonding", Semiconductor Science and Technology, Vol., 21, 2006, pp. 959-963

HENTTINEN, K. et al., "Mechanically Induced Si Layer Transfer in Hydrogen-Implanted Si Wafers," Applied Physics Letters, April 24, 2000, p. 2370-2372, Vol. 76, No. 17.

HENTTINEN, K. et al., "Cold ion-cutting of hydrogen implanted Si," J. Nucl. Instr. and Meth. in Phys. Res. B, 2002, pp. 761-766, Vol. 190.

CHO, Y., et al., “Low Temperature Si Layer Transfer by Direct Bonding and Mechanical Ion Cut,” Applied Physics. Letters., vol. 83, no. 18, November 2003, pp. 3827-3829.

EN, W. G., et al., “The Genesis ProcessTM: A New SOI wafer fabrication method”, Proceedings 1998 IEEE International SOI Conference, pp. 163-164 (Oct. 1998).

CURRENT, M. I., et al., “Atomic-layer Cleaving and Non-contact Thinning and Thickening for Fabrication of Laminated electronic and Photonic Materials”, 2001 Materials Research Society Meeting, April 16-20 2001, Paper I8.3.

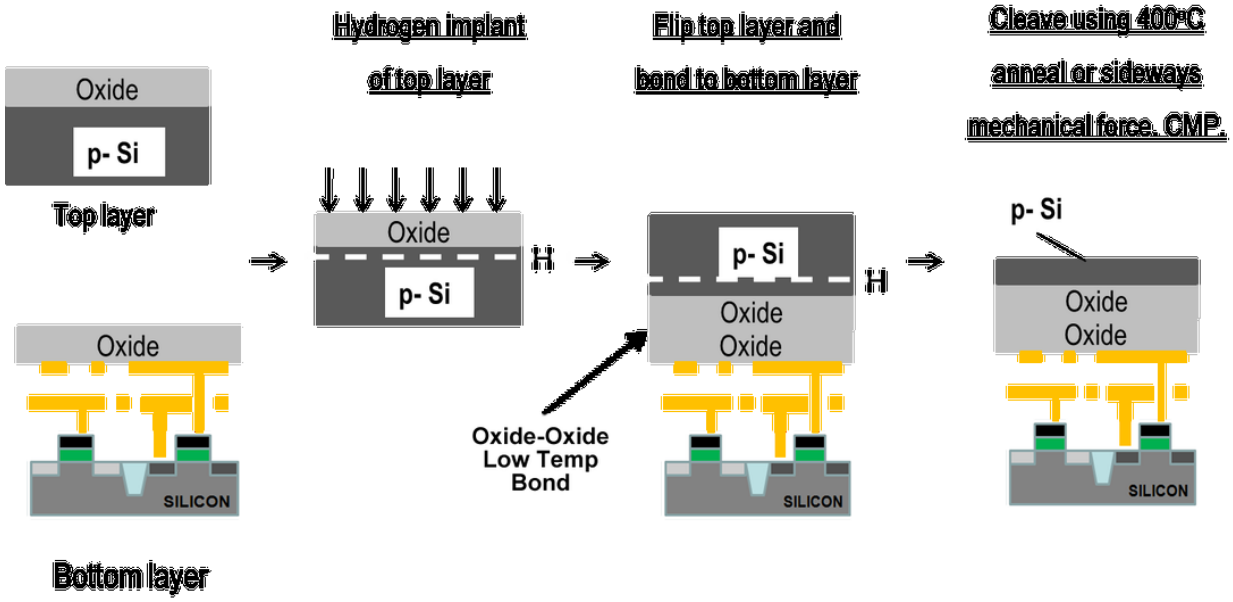
Chapter 7 - Low Temperature Wafer Direct Bonding

by Brian Cronquist, VP of Technology and IP of MonolithIC 3D Inc.

Sometimes we get questions about a particular aspect of the monolithic 3DIC flow. Here I would like to talk about Low Temperature Wafer Direct Bonding, where an important concern is the strength of the wafer to wafer oxide to oxide bond. Can it survive the subsequent transistor formation or wafer thinning processing, whether that processing entails the shear forces of a CMP, the thermal gradients of a low temperature deposition, or the stress release of a plasma or wet etch?

Direct wafer bonding is both desirable and required for low cost high yield monolithic 3D integration. “Direct” meaning that an extra layer, an intermediate layer, such as an adhesive, is not used. The bonding between the surfaces only involves the chemical bonds between the two surfaces. The simplest case for a conventional wafer fab, which has the highest probability of achieving high yield & low cost direct bonding, is oxide to oxide bonding. Oxide to oxide wafer bonding has the added advantage that a through layer via connection may not need an isolation liner, and is part of a process integration strategy that delivers a Thorough Layer Via (TLV) with processing ease and characteristics similar to a conventional BEOL metal to metal via.

Another enabler for monolithic 3D integration is a direct bonding process that has thermal exposures to the underlying layer or layers that does not exceed 400°C. This allows the use of conventional metallization and low-k dielectrics such as copper & carbon containing low-k oxides BEOL, rather than difficult to manufacture high temperature metals such as tungsten. Two additional advantages of low temperature bonding are avoiding any wafer deformation due to thermal expansion effects (greatly helps across the wafer alignment precision), and minimizing the thermal effects on the lower layer transistor hi-k metal gate stacks and junctions.



The generally accepted strength threshold metric of a wafer to wafer bond that would enable thinning, such as CMP, and other processing (<400°C), is between 1.0 and 1.2 J/m2. DiCioccio talks about 1.14 J/m2 bond strength as *sustaining processing such as silicon thinning (backgrind and CMP)*. Dragoi shows that surface energies above 1.2 J/m2 *allow bonded pairs to survive even harsh processes as grinding or lapping*. Radu found that a bonding strength *of more than 1 J/m2 has been sufficient to sustain post-processes such as silicon back thinning using coarse and fine grinding*.

Many investigators, groups and companies have developed pre-bonding surface conditioning and post-bonding thermal treatments to control and optimize the bonding strength within the constrained thermal budget window (<400°C) and have achieved excellent bond strength's greater than 1 J/m2. A sampling of the literature follows:

DiCioccio et al. at ICICDT 2010 [CEA-LETI-Minatec, Grenoble, USA] showed acceptable bonding strengths from bonded wafers with 5um copper pads that cover 20% of the area, the remainder is oxide to oxide, after a 2 hour 200°C or 400°C post bond anneal. The surfaces were carefully prepared with CMP.

Annealing temperature	Bonding toughness ($G=r_0+r_p$)
200°C	1.14 J/m^2
400°C	6.6 J/m^2

Table 1: Bonding toughness (G=r0+r) of the bonding pair as a function of the post bonding annealing temperature. The annealing step was 2h long.

Radu et al. at the 2010 3DIC conference [Soitec Bernin, CEA-LETI-Minatec, Soitec USA], showed bonding energy data obtained from 200mm wafer bonding of Cu/Cu full sheet, SiO2/SiO2 full sheet, Cu/SiO2 full sheet, and pattered 5um Cu pads at 20% density. Oxide to oxide bonding at 200°C produces over 1 J/m2 bonding energy. The surfaces were carefully prepared with CMP.

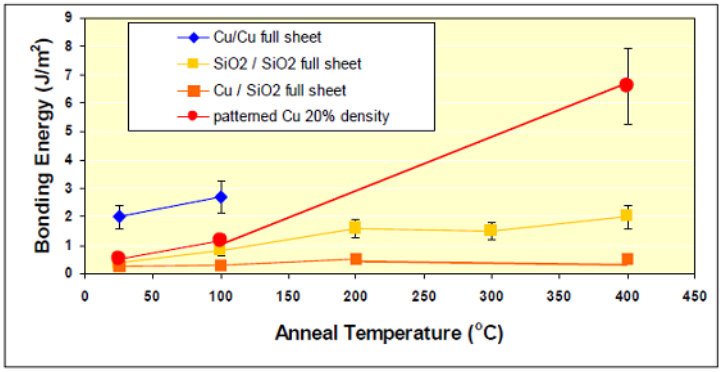


Figure 2 Bonding energy evolution with temperature of different type of interfaces (Cu-Cu, Cu-Ox and Ox-Ox)

Gaudin et al. at 3DIC 2010 [Soitec Grenoble, Soitec USA, IBM Albany, IBM East Fishkill] utilized 300mm wafers with a backend CMOS process and deposited oxide layer acting as the bonding layer. Bonding surfaces were prepared with an optimized CMP process and post-bond annealing, thinning and grinding were successfully performed. Gaudin studied one TEOS-based oxide and two different condition sets for silane-based PECVD oxide. Silane condition B was certainly superior and exceeded the 1 J/m2 metric at both 200°C and 400°C.

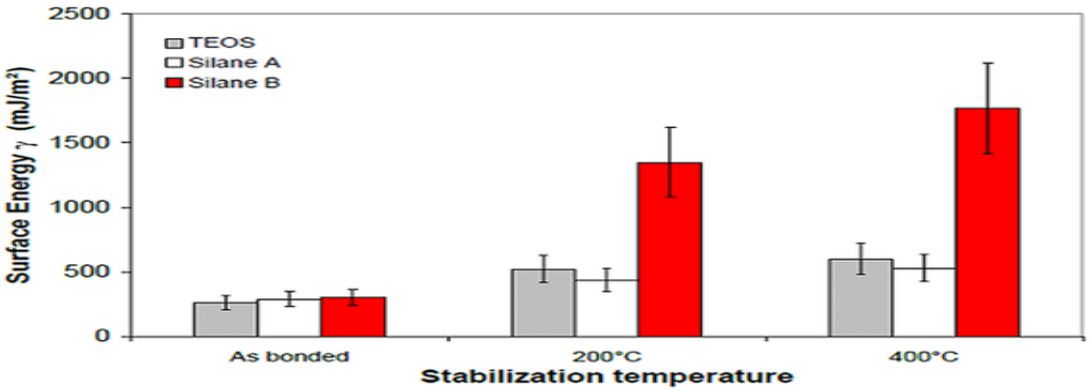


Figure 3 Surface energy (γ) evolution with temperature of oxide bonding stacks using 3 different deposited oxides

Gaudin further influenced the bonding quality by conditioning the surface with wet chemical processing (Process I) and dry plasma processing (Process II).

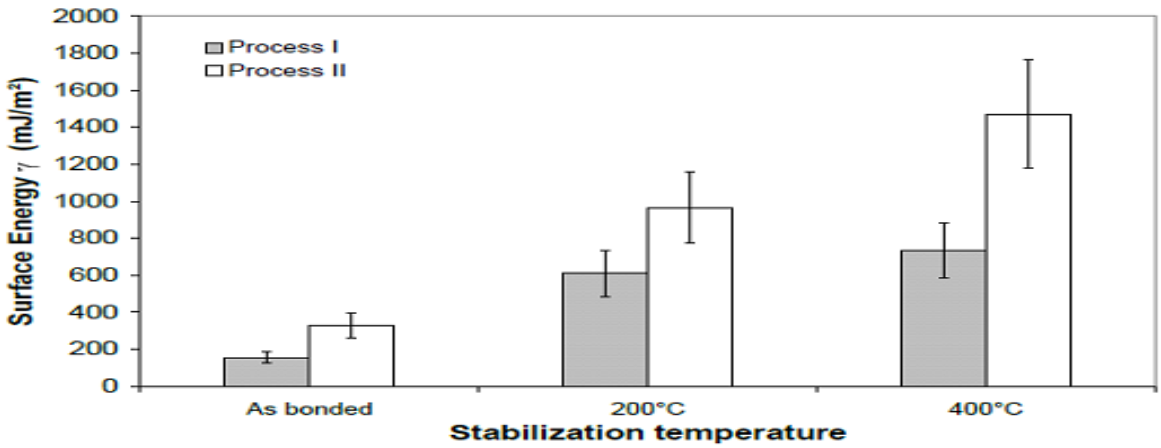


Figure 4 Surface energy (γ) evolution with temperature for 2 surface preparation process options on a TEOS based oxide

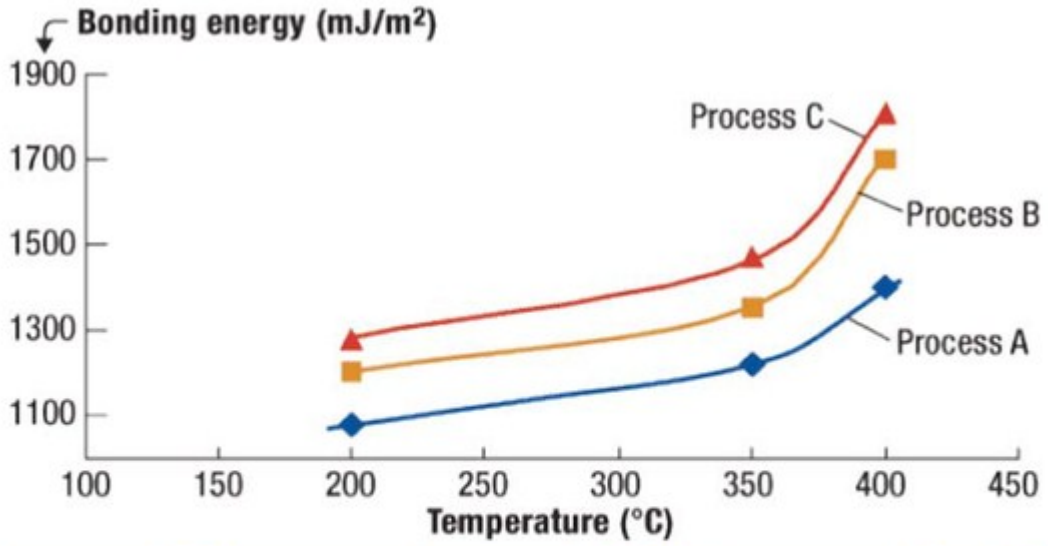
Dragoi et al. at SPIE 2007 [EV Group] showed blank wafer data where a PECVD oxide was deposited, outgassed in a vacuum anneal at 300-400°C 1-3 hr anneal, CMP polished, nitrogen plasma activated, megasonic cleaned, vacuum bonded with 5kN force, then annealed for 1 hour at 300°C.

Sample no.	Measurement no.	Surface energy and measurement error (J/m^2)
1	1	$2.116 \pm 0.556\ J/m^2$
	2	$2.342 \pm 0.627\ J/m^2$
2	1	$2.89 \pm 0.807\ J/m^2$
	2	$2.598 \pm 0.71\ J/m^2$
3	1	$2.342 \pm 0.627\ J/m^2$
	2	$2.598 \pm 0.71\ J/m^2$
4	1	$2.598 \pm 0.71\ J/m^2$
	2	$2.598 \pm 0.71\ J/m^2$
5	1	$2.598 \pm 0.71\ J/m^2$
	2	$2.89 \pm 0.807\ J/m^2$

Tabel 2 Bond strength measured for sample 1

Dragoi successfully applied the process on 200mm Si on CMOS bond pairs.

Sadaka et al. in electroiq.com (2010) [Soitec USA, CEA-DRT-LETI] showed 3 different processes (CMP/surface conditioning/planarization/cleaning). With 200°C, 350°C or 400°C post bond anneals, the target of 1 J/m2 was achieved.



Test	Condition	Pass/fail
Operating life test (endurance)	125°C/2000h	√
High temperature storage	150°C/1000h	√
Temperature cycling	-55°C, +125°C, 15°/min	√
Moisture resistance	65°C, -10°C, 90-100% RH-10 days	√

Ziptronix talks about their DBI (Direct Bond Interconnect) technology as utilizing RIE surface cleans & porosity enhancement, NH4OH surface treatments, CMP to 0.5nm RMS, and obtaining bond energies >1 J/m2 at room temperature.

Henttinen et al. in Applied Physics Letters April 2000 [VTT Electronics, Finland; UC San Diego] demonstrated oxide to oxide bonding of silicon wafers with various plasma or RCA clean pretreatments, and post-bond thermal anneals.

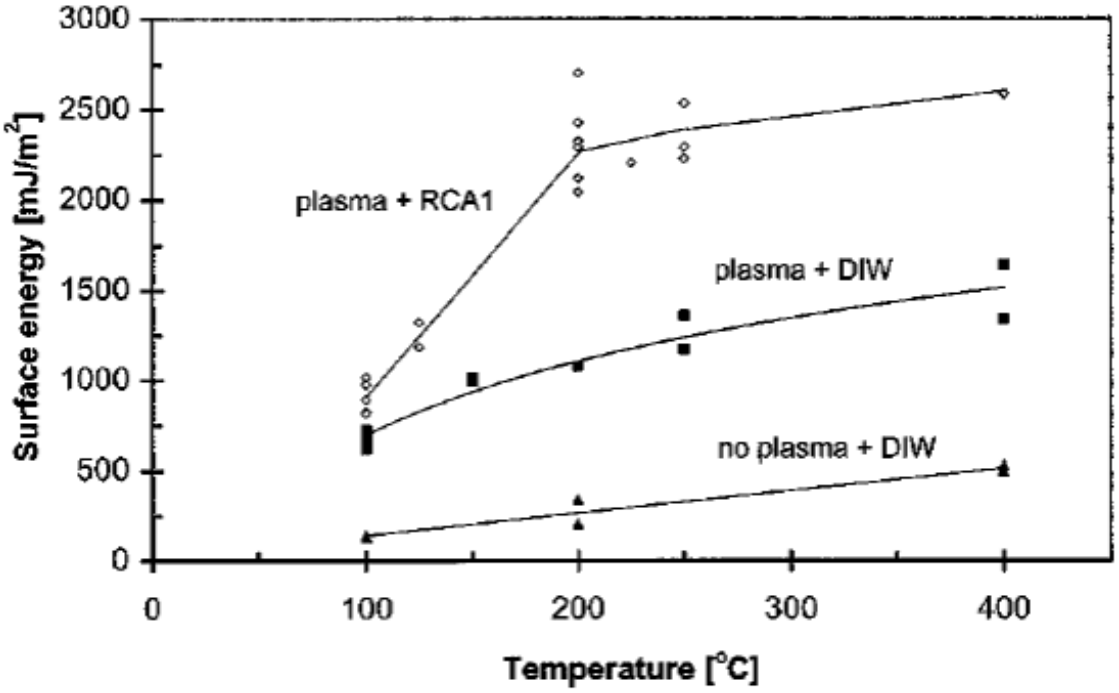


Figure 6 Bond strength of the bonded interface as a function of bond annealing temperature. The annealing time varied between 30 min and 24h.

SiGen Corporation reported in 1999 and 2000 the use of a plasma activated pre-bond step to achieve >1 J/m2 bonding strength.

In summary, a variety of investigators have shown processes capable of providing excellent wafer to wafer bond strengths.

References:

DICIOCCIO, L., et al., "Direct bonding for wafer level 3D integration", ICICDT 2010, pp. 110-113.

DRAGOI, et al., “Plasma-activated wafer bonding: the new low-temperature tool for MEMS fabrication”, Proc. SPIE, Vol. 6589, 65890T (2007).

RADU, I., et al., "Recent Developments of Cu-Cu non-thermo compression bonding for wafer-to-wafer 3D stacking", IEEE 3D Systems Integration Conference (3DIC), 16-18 Nov. 2010.

GAUDIN, G., et al., "Low temperature direct wafer to wafer bonding for 3D integration", 3D Systems Integration Conference (3DIC), IEEE, 2010, Munich, 16-18 Nov. 2010, pp. 1-4.

SADAKA, M., et al., "Building Blocks for wafer level 3D integration", www.electroiq.com, August 18, 2010.

www.ziptronix.com, DBI fact sheet

HENTTINEN, K. et al., "Mechanically Induced Si Layer Transfer in Hydrogen-Implanted Si Wafers," Applied Physics Letters, April 24, 2000, p. 2370-2372, Vol. 76, No. 17.

I.J. MALIK, et al., “The Genesis Process: A general layer transfer method for electronic applications,” Spring 10999 MRS Symp. Tech. Proc., 1999
F.J. HENLEY, et al., European Semiconductor, 25, Feb. 2000.

Chapter 8 - How much does ion-cut cost?

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

Ion-cut, the process used for manufacturing SOI wafers for the past 15 years, is the most popular method to form c-Si layers for monolithic 3D-ICs. Here, I'll share cost estimates for ion-cut, and explain why even price-sensitive markets such as solar are adopting it.

For monolithic 3D, it is often required to form single crystal silicon above copper wiring layers at temperatures lower than 400C. Fig. 1 shows the ion-cut process, which is the most popular method of achieving this objective. Hydrogen is first implanted into a "top layer wafer" to create a defect plane. This "top layer wafer" is then flipped and bonded onto a "bottom layer wafer" having transistors and copper wiring. After this, the structure is cleaved at the defect plane using a 400C anneal or a sideways mechanical force. Finally, a CMP is done to get a good surface.

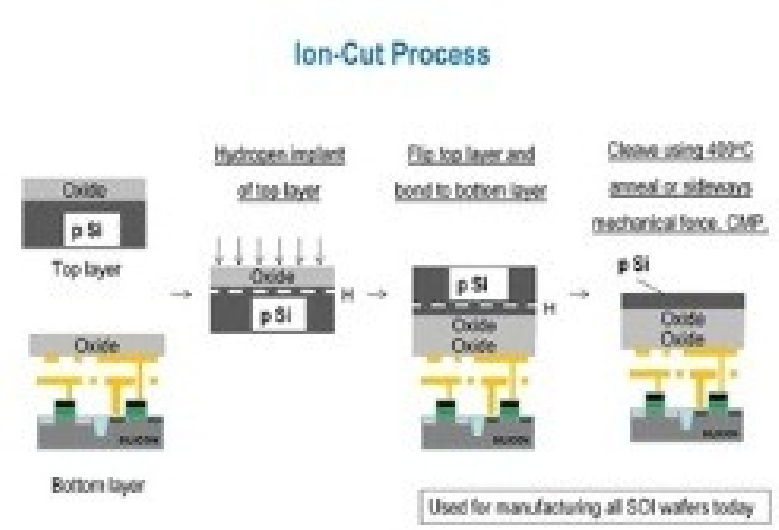


Fig. 1

The previous paragraph explained how ion-cut can be used for stacking single crystal silicon layers for 3D-ICs. For forming a SOI wafer using ion-cut, the "bottom layer wafer" in Fig. 1 is a blank silicon wafer instead of a processed one with transistors and wires. As many of you know, ion-cut is the standard process used for high-volume manufacturing of SOI wafers today.

Cost-of-Ownership Analysis

Fig. 2 shows cost calculations for ion-cut using a Sematech Cost-of-Ownership framework. Tool prices and throughputs are obtained from equipment manufacturers who provide tools for these ion-cut process steps. The "top layer wafer" in Fig. 1 is re-used, as is typical in an ion-cut process. **The total cost per wafer for a single ion-cut is $58**, which is close to estimates that ion-cut practitioners in the industry have provided us. The number seems reasonable... this is what you'd expect of a process that doesn't involve any litho steps. In addition, with passage of time, one would expect throughput of various steps to improve significantly, bringing the price down further.

Cost-of-Ownership for an Ion-Cut Process

Step	Tool throughput	Tool Cost	Consumables	Cost per wafer
Oxidation	40wph	1.4M	$1	$3
H implant	50 wph	4.5M		$5
Bond	30 wph	3M		$6
Cleave	50 wph	2M	$1	$3
CMP	35 wph	3M	$4	$9
Substrate re-use				$22
Other steps				<$10 (?)
Total cost per wafer for a 20k wspm fab				**$58**

Fig. 2

Companies such as Twin Creeks Technologies and SiGen are using ion-cut for the solar industry today (Fig. 3). As you'd know, the solar industry is a lot more cost-sensitive than the semiconductor industry... this application is possible mainly because these vendors are reaching costs similar to Fig. 2.

Fig. 3

Hmmm... If the additional cost per wafer is $58, why are SOI wafers considered "costly" today?

This is because of business issues with SOI wafer manufacturing (see Fig. 4).

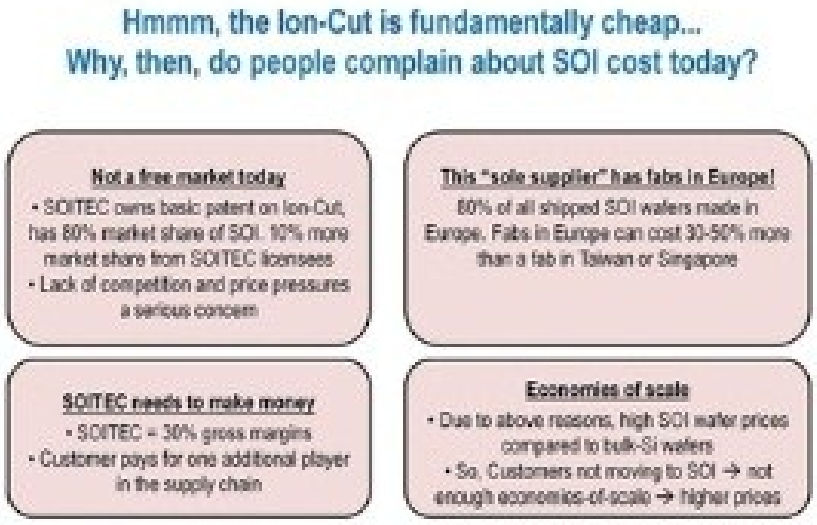


Fig. 4

- SOI wafer manufacturing is not a free market now: One player, SOITEC, **controls 90% of the SOI market** today since it owns the basic patent on ion-cut from **Michel Bruel**. Markets dominated by a single supplier typically have high prices due to lack of competitive pressures.

- This "sole supplier" makes ~80% of its SOI wafers in Europe: Believe it or not, around 80% of SOI wafers are made in Europe today! Its incredibly expensive to have a fab in Europe - that's why all manufacturing is moving to the Far East. A rule of thumb I've heard is that a fab in Taiwan or Singapore is 30-50% cheaper than one in the US or Europe. This is mainly due to government incentives such as tax breaks, lower building costs and lower labor costs. For example, one company I know got a deal from a Far Eastern nation to have a 10 year tax holiday and the government paid 60% of the company's capital expenditure! I also hear SOITEC built its latest fab in Singapore to tackle some of these issues, but that fab only provides only ~20% of its total output now and is running at 10% of its maximum possible capacity. All of us know an under-utilized semiconductor fab is expensive... (Note that some of the numbers in this paragraph are things I heard from industry sources, they are not official estimates)

- Additional player in the supply chain: A company providing SOI wafers today buys a bulk silicon wafer, does the ion-cut process on it and then sells the finished wafer to foundries and IDMs. You're essentially adding an additional player in the supply chain here, with his own margin requirements. Ion-cut manufacturers such as SOITEC have 30% gross margins, so the customer pays extra for this.

- Not enough economies-of-scale: Due to the business constraints listed above, the SOI wafer price overhead is significantly more than the $58 we calculated above. So, SOI adoption has not proceeded as fast as expected, and one cannot reach high enough economies-of-scale. This, in turn, keeps price high compared to bulk Silicon wafers, which hinders adoption. This chicken-and-egg problem (high prices --> low adoption --> not enough economies of scale --> high prices) is a concern.

How do we deal with the business challenges of ion-cut?

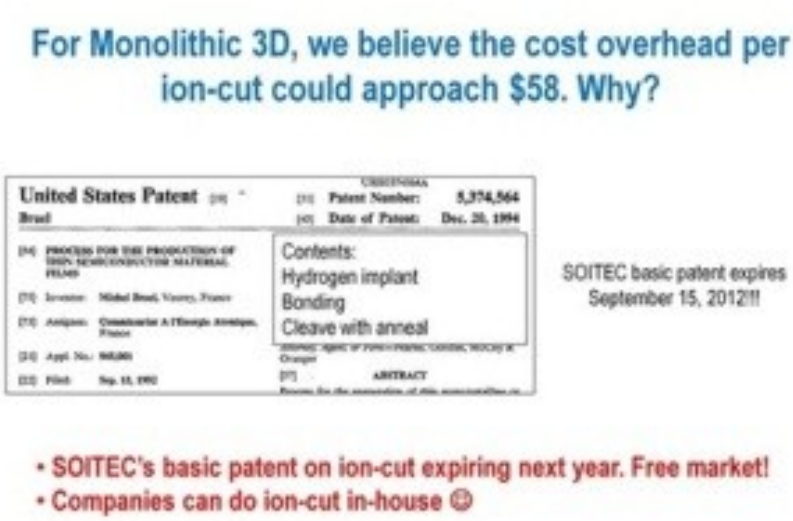


Fig. 5

- Many people in the ion-cut community believe the business situation for ion-cut will change on September 15, 2012. Why? Because the basic patent from Bruel describing ion-cut expires that day. Check out patent number **5374564 at the US Patent Office Website**. It talks about all the technologies described in Fig. 1: the atomic species implant, bonding, cleaving with anneal, surface cleans, etc. See Fig. 5 for more details. Once the ion-cut becomes a public-domain technology, we believe a free market situation will arise, benefiting everyone. Competition will lower prices which will boost adoption significantly.

- For Monolithic 3D applications, we feel the best way forward is for each company (eg. TSMC, Intel, ST, Toshiba, Fujitsu, Samsung, Micron, etc) to do the ion-cut in-house. So, these companies would place equipment for H implant, bond and cleave in their own fabs and run this process themselves. This will keep costs down since the problems described in Fig. 4 can be avoided, and this will be possible after 2012. One could approach the $58 price per ion-cut that I showed in Fig. 2.

What's the bottom line?

The price per ion-cut could be as low as $58, which is miniscule compared to wafer cost of a logic wafer (~$4000), NAND flash memory wafer (~$1500) or DRAM wafer (~$2000). This is encouraging for the monolithic 3D application, since ion-cut is the most popular technique to get stacked single crystal silicon layers. Once these stacked single crystal silicon layers are obtained, one can use MonolithIC 3D Inc.'s innovative device architectures to build high-quality 3D chips.

Chapter 9 - Is MonolithIC 3D-IC less risky than scaling or TSV?

by Brian Cronquist, VP of Technology and IP of MonolithIC 3D Inc.

I recently saw this great 5 minute video by Applied Material's Richard Lewington **[AMAT 3D Blog Video]** where three types of 3D-IC construction are demonstrated. Note that the first two 3D-IC options he shows (with those plastic blocks) are monolithic. Only the third option is TSV based.

What's going on here? Why is this major equipment vendor talking about monolithic 3D when it *seems* that most of what the industry is talking about these days are scaling, interposers, and Thru-Silicon-Vias(TSVs)? Let's take a look.

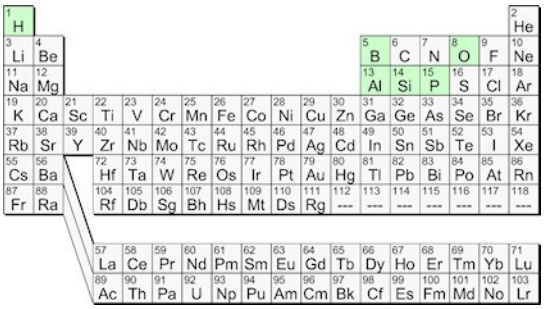
Being a fab-guy (built parts of and worked in Chartered Fab-1 & Fab-2, Sierra Semi's fab inside National Semi's Bldg#4, AMI Poci Fab-4, Synertek Fab-3, etc.) I am going to approach this from a process/fab-rat perspective. Because this is a key point to what monolithic 3D is about: it is supposed to bring 3D-IC back into the wafer batch economics of semiconductor processing. No piece part handling expense, TSV/interposer reliability & cost issues, or OSAT troubles (I applaud TSMC for trying to remedy this OSAT part, but am surprised that Global Foundries did not do it first….they could have beaten TSMC to the punch here).

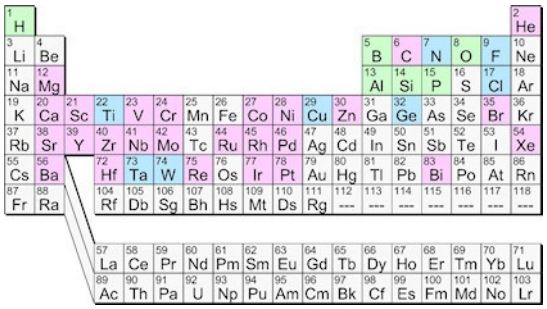
The major rule for wafer fabs is *Take no Risks*….. Everything you do is focused on control: understanding, eliminating, controlling variables. Protect and preserve that huge capital investment so you can pay it down. By definition & nature, fab managers are very conservative. But scaling forced us to do dramatically different and risky things. That's a major reason why it takes 10+ years for new process/technologies to get into a large production fab. Think about HKMG, Cu BEOL, CMP, strain, plasma metal etching rather than wet (caused lots of corrosion issues/mousebites), to name a few. Even platen cooling (instead of aluminum mask layers) for high current implantation took a long time. Changing from flats on the starting material wafers to the notch took about 10yrs too.

At its root, many of these changes took new machines, new chemistries, and/or new process methods (think APCVD, LPCVD, UHVCVD, PECVD, SACVD, ALCVD, MOCVD, RTCVD,…..) Another large risk factor with scaling has been the use of more

elements of the periodic table to solve scaling challenges. We did not just alter the form or compound of a known element (bad enough risk-wise); we changed to and added new elements to our expensive wafer fabs. (In fab parlance, all this "newness" added up to what is called the *Sphincter Effect*)

When I started in the industry we used only six elements from the periodic table:

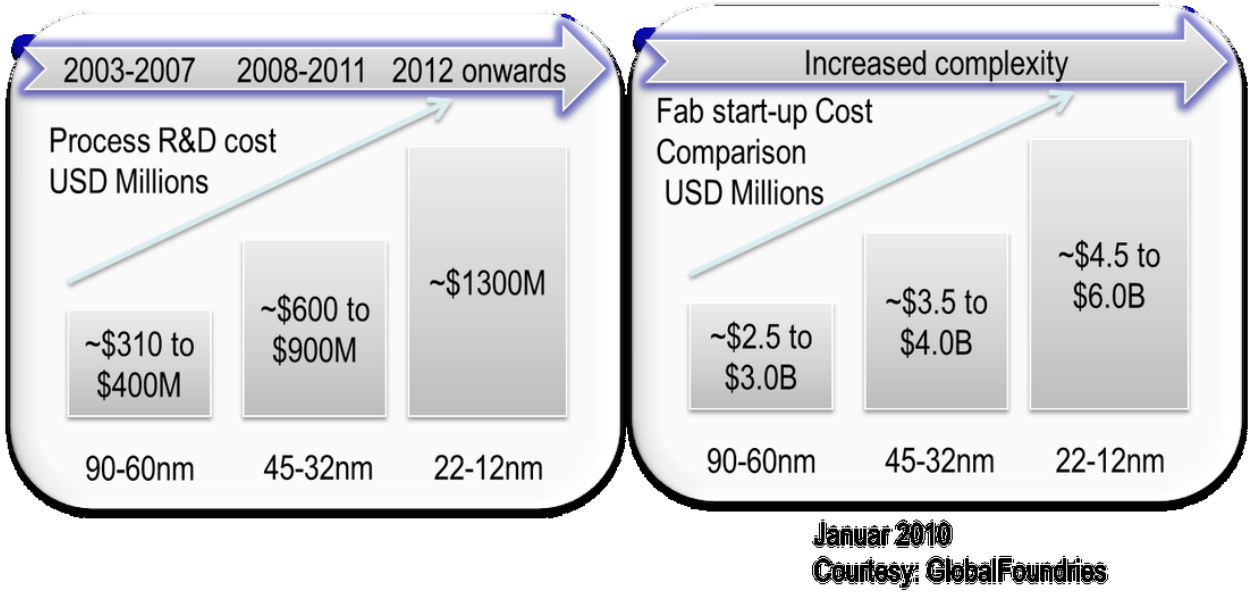




Yet, all of us scientists and engineers, as well as fab managers, solved the problems caused by relentless scaling, and the industry grew…we had a lot of fun, we were supremely challenged, and we solved those challenges. But we also grew grey hair and permanently pinched sphincters.

At what cost? (remember, low cost is crucial to successful manufacturing!)

Here's what Global Foundries showed about costs:



So, now we have now included the investment and banking communities into our *Sphincter Effect.*

Enough! This is the road to ruin; well, at least to vastly diminishing returns (think Handel Jones' chart **[ElectroIQ link to ISS12 Day 2]**on how transistor cost is no longer going down…)

3D-IC is the solution. OK, so…. monolithic or TSV or interposer? Above I already mentioned a few of the risks and costs to a TSV/interposer solution. Look at all the new processes and machines that had to be developed to etch and fill such deep holes at least somewhat economically. And the integration issues are significant because of the novelty and the architecture & flow: Cu/silicon stresses, keep out zones, liners, new reliability fail modes, etc. As usual, these issues will likely be solved; hence, TSV & interposers will be useful for obtaining some cost and functional/architectural gains from its limited vertical connectivity. But they are not the endgame. To get fully back onto the economic scaling path we need rich vertical connectivity.

What about monolithic 3D-IC risks & costs? Fab equipment and unit processes exist. No new elements from the periodic table are necessary. And the gains resulting from this dense vertical connectivity keep us on a scaling equivalent path (no need to spend space here…lots has been written about this). Let's instead look at the process details:

Oxides for ox-ox direct bonding: Deposited oxides are well understood and cheap. No new equipment or elements are needed. Lots of manufacturing proven techniques to get there: PECVD, SACVD, etc.

H Implant: Can be done on current models. No new equipment needed. Done by SOI manufacturers for 20 years. H in silicon is well understood.

Bonding: Two well-known equipment vendors (EVG & SUSSMicroTec) with low temp oxide to oxide bonding capability and significant sales of machines (mostly to BSI sensor folks at this time). A recent third new entry (MHI-Mitsubishi Heavy Industries) with **room temp**ox-ox bonding. I recently blogged on this topic too. **[BC LT direct bonding]**

Cleave: Lots of methods proven for SOI manufacture, sensors, and solar. Simplest is thermal … just use a furnace or RTP. We made a short movie clip showing how simple cleave is with the AG RTP at Stanford.

Monolithic 3D-IC uses existing wafer-fab equipment, needs no new elements from the periodic table, and utilizes well-known unit processes and chemistries.

What's the catch? It's the integration. Integration work (*blood, sweat, and tears*) will always be there, even with no new elements, machines, chemistries, etc. Always. However, those who have done new process introductions know that integration is **significantly** less risky (= costly) and faster to market without than with the elements/machine/chemistry changes. New modes of defect generation are always generated from integration, but there are a lot less of them if all the unit processes are standard accepted practices, than if those unit processes are totally new.

If you look very very carefully at the MonolithIC 3D Inc's process flows, you notice we were single mindedly focused on making it simple. For example, the nm-scale thru layer vias (TLVs) are always made thru the STI (Shallow Trench Isolation); hence, no dielectric liners, minimum stress, conventional etch and fill, nothing high aspect ratio about it. Make the TLV look and feel like a regular metal to metal via.

This shows in the costs. Deepak Sekar did a SEMATECH based cost estimate and talked about it in a blog. **[Deepak Blog ion-cut cost]** Here's his summary chart for 300mm wafers.

Step	Tool throughput	Tool Cost	Consumables	Cost per wafer
Oxidation	40wph	1.4M	$1	$3
H implant	50 wph	4.5M		$5
Bond	30 wph	3M		$6
Cleave	50 wph	2M	$1	$3
CMP	35 wph	3M	$4	$9
Substrate re-use				$22
Other steps				<$10 (?)
Total cost per wafer for a 20k wspm fab				**$58**

Validation of Monolithic 3D

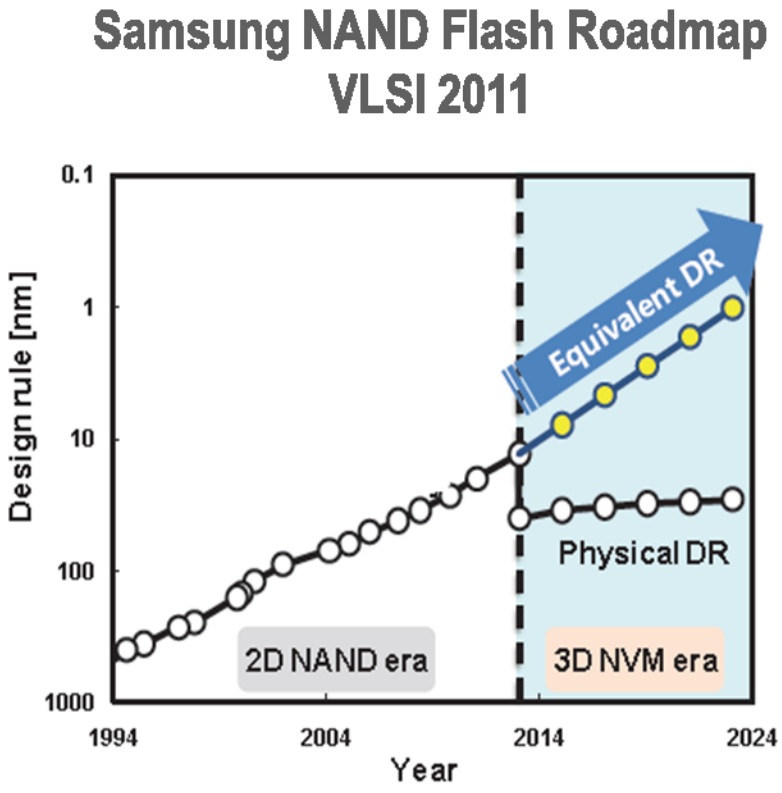
One may make the argument that validation of a nascent & new game-changing technology is impossible, or at least very nearly so. However, for monolithic 3D-IC there

are at least two important data-points to consider. And I hope that you will be convinced that monolithic 3D-IC is neither so nascent nor new.

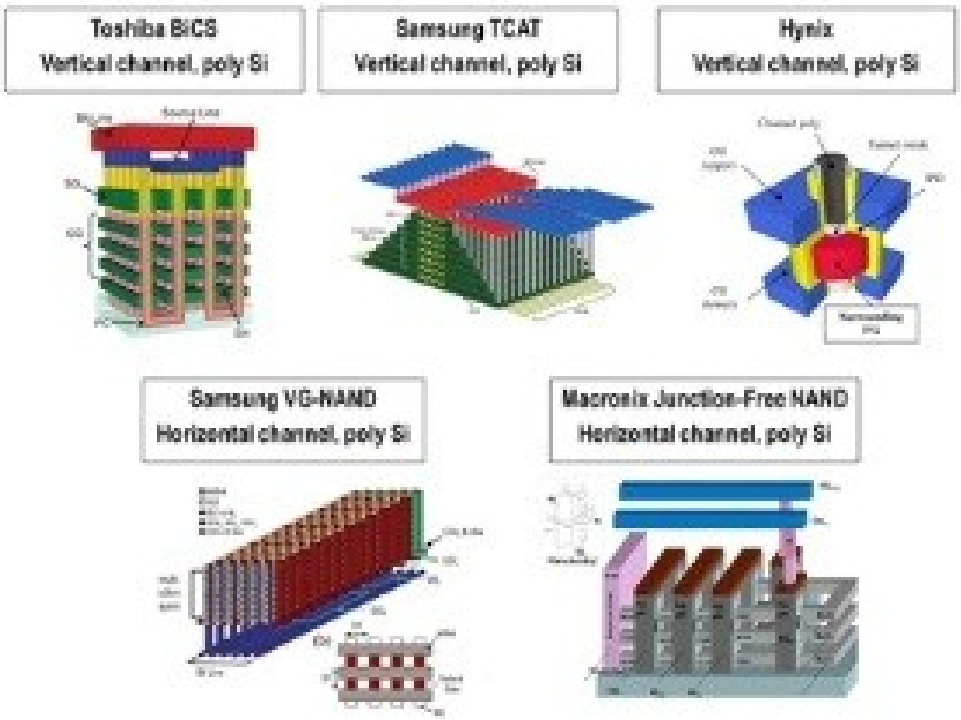
NAND Memory Makers going 3D: People such as David Lammers of Semiconductor Manufacturing & Design Community **[Lammers July 2011]** have pointed to validation evidence that the time of monolithic 3D-IC is near: the bleeding edge NAND memory makers are already moving to monolithic 3D-IC.

"The advent of 3D NAND memories may be only two or three years away, speakers said at Semicon West in San Francisco. By 2013 the major memory companies developing 3D NAND, including Hynix, Samsung, and Toshiba, may be ready with pilot lines, moving to volume production a year or so later. Taiwan-based Macronix International also has been developing a 3D NAND solution."

At the recent (2011) VLSI Symposium J. Choi of Samsung showed their view of how they will keep on making cheaper bits … by going 3D monolithically.



Deepak Sekar has also talked in detail about this 3D monolithic push by the NAND industry (Sekar hails from flash maker SanDisk) in his recent blog **[12/11/2011: where-is-the-nand-flash-industry-heading]**.



Second, the global semiconductor equipment leader, AMAT, has talked about sales into that market **[SemiconWest2011-new products including 3D architecture support]**

[OptivaCVD for BSI] and even has a video (Richard Lewington's blog video noted above) to promote it.

When both manufacturers and equipment suppliers are talking about, committing to, and executing on a specific technology change, you know that the economics are attractive and not just niche. Think back to how HKMG and copper BEOL came to production.

The chicken and egg are *out the window*….it's happening now. The risks are contained. Others are going for it.

Whether polysilicon or monocrystalline silicon based monolithic 3D, jump in and be a part of this next important evolution of our great industry.

Don't miss out.

Chapter 10 - The Future is the Interconnect: IITC

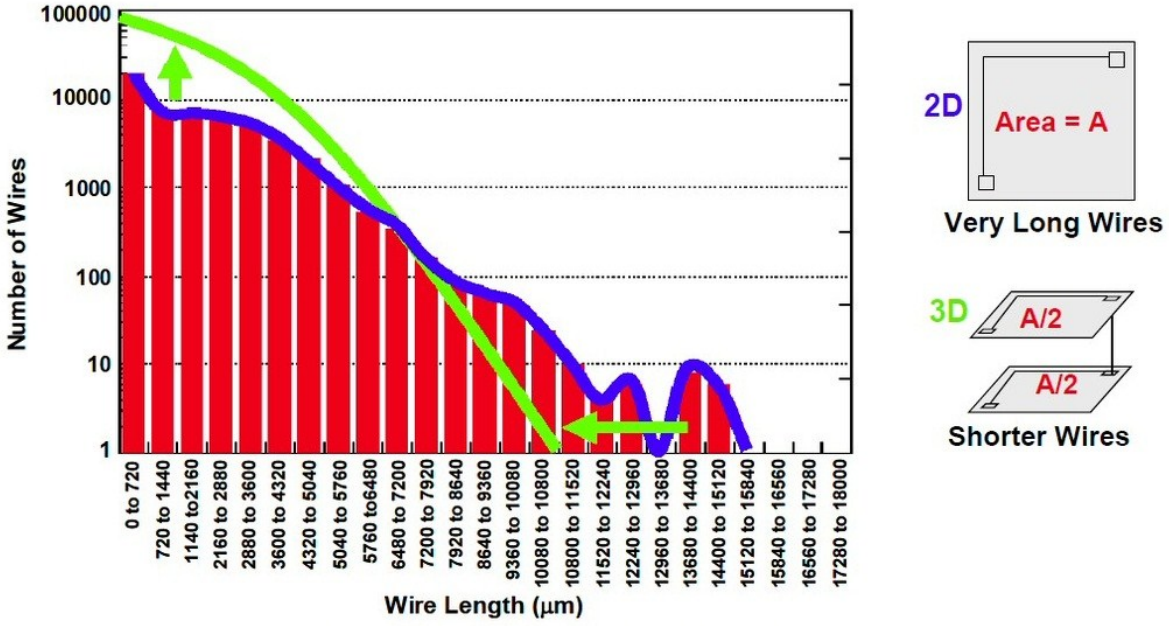
by Ze'ev Wurman, Chief Software Architect of MonolithIC 3D Inc.

Does Size Matter?

The next International Interconnect Technology Conference (**IITC 2012**) will be held in San Jose in a couple of weeks (June 4-6). This is a good opportunity to recall that, in some sense, the reason for scaling silicon down has changed in recent years from packing more transistors in a square (or cubic) millimeter to increasing functionality and performance at reduced power. An ever higher fraction of the power dissipation resides in the interconnect – both in the net switching itself as well as in the ever-increasing number of repeaters required to re-power more and more "long" nets.

Estimates of the area dedicated to repeaters as technology shrinks vary but even if the early predictions of 70% cells being dedicated to repeaters at 32 nm may have not come to pass (Saxena, TCAD 2004), a large fraction of chip power is now dissipated by interconnect structures. This is particularly true in FPGAs where the interconnect share of routing-related dynamic power may easily reach 2/3 of the power, but even non-programmable devices have been reported to have half of their power dissipated in the wires already at 90nm. The following slide is from the 2006 High Performance Embedded Computing workshop.

Wire Length Distribution in 90 nm Node IBM Microprocessor*



- >50% of active power (switching) dissipation is in microprocessor interconnects
- >90% of interconnect power is consumed by only 10% of the wires

MIT Lincoln Laboratory

HPEC 2006 -24
CLK 9/19/2006

After K. Guarini IBM Semiconductor Research and Development Center

Last year IITC included a paper from Georgia Tech (Dae Hyun Kim, et al., *Impact of Through-Silicon-Via Scaling on the Wirelength Distribution of Current and Future 3D ICs*) that explores the impact of 3D on the average wire-length of deep submicron ICs. This paper differs from many others in that it explores the impact as a function of TSV size, and it models TSVs from the currently feasible 5 micron, with a 5:1 aspect ratio for the corresponding 25 micron thick silicon layer, down to a futuristic 100 nm, with a 50:1 aspect ratio for a 5 micron thick layer. Such futuristic TSV actually gets close to a monolithic process, which can achieve silicon thickness of one micron and below. Here is a key chart from this paper:

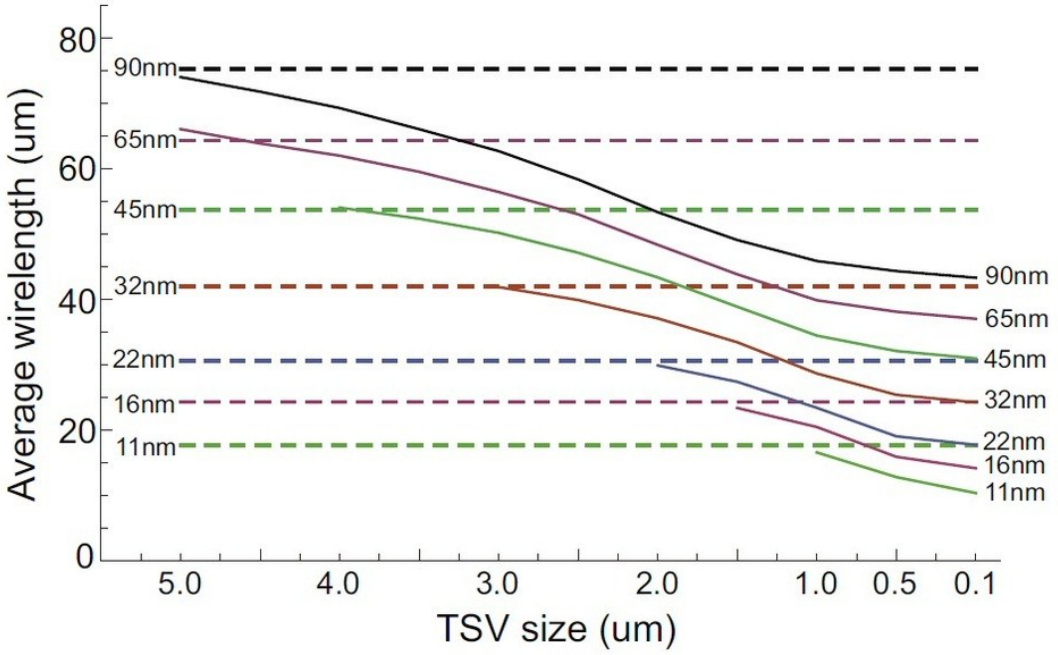
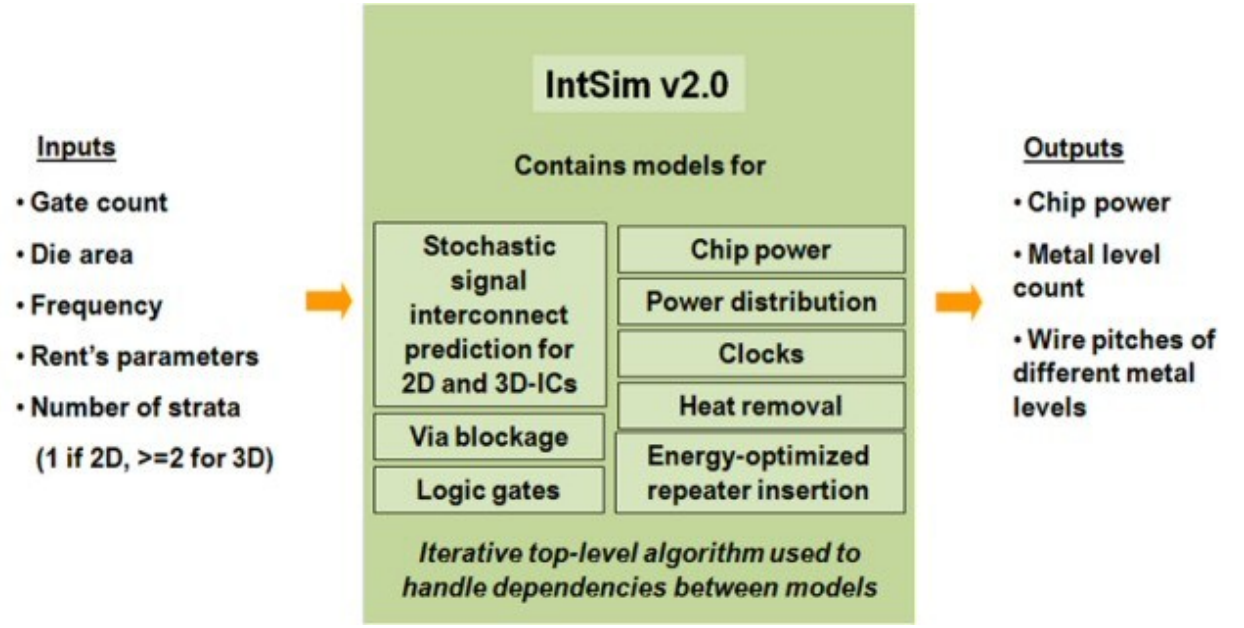


Fig. 6. Cross comparison among various 2D and 3D technologies. Dashed lines are wirelengths of 2D ICs. # dies: 4.

As we can see, a small-sized TSV can significantly reduce the average wire-length by up to 50%, and reflects an improvement equivalent to two or three technology generations. In other words, a 4-way stacked 32nm chip with monolithic-style vertical connectivity can have wire-length distribution as good as a 16nm cutting edge technology, with the associated reduction in power and increase in performance, but using a relatively inexpensive and depreciated fab line.

Yet there is a fly in this ointment – TSVs with aspect ratio of 50:1 are not likely to happen, and using nanometer-TSV with extremely thin silicon layers to maintain AR below 10 creates problems of its own. Just recently **IMEC reported** stress issues at 25 micron thickness and "*found that increase in the die thickness from 25 to 50 um resulted in a stress reduction of 3X. Final conclusions were that 50 um thickness die were currently much better option for scalable manufacturable process*." In other words, the road to nanometer-scale vertical connections does not go through scaling down TSVs but through **monolithic** process and layer transfer.

I find all this a nice illustration of the importance of the monolithic stacking approach that is also easily visible using **our free simulator**, IntSim.



Transformation to 3D monolithic stacking is much more than simply saving on a footprint by slicing and stacking the same design. The rich vertical connectivity offered by monolithic stacking significantly reduces the average distance between source and destination and therefore improves performance, saves power, saves total area, and allows players to continue using older process fabs to achieve cutting edge results at a cheaper cost. The chart below illustrates such savings at 22nm technology:

22nm node 600MHz logic core	2D-IC	Fine-Grain 3D 2 Device Layers	Comments
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
Optimal 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**	

The future of Moore's Law and the continued well-being of our industry is in the small nanometer-sized TSV, not in the big micron-sized TSVs used today **that are so**

hard to manage. And let's hope that the upcoming IITC will be at least as interesting as last year's.

Chapter 11 - Can Heat Be Removed from 3D-IC Stacks?

by Brian Cronquist, VP of Technology and IP of MonolithIC 3D Inc.

Thanks to everybody who came to **IEDM** this year, and especially to those I met and who came to paper 14.2, delivered by **Hai Wei** of**Stanford** University. You can find the meeting paper and slides **here**.

One of the big challenges facing 3D-IC is how to remove the heat dissipated on the upper layers to keep a high performance chip temperature within the system and reliability constraints and prevent hot spots. Most existing proposed techniques rely on arrays of TSVs and thick (xxum) silicon layer to conduct and spread the heat laterally and vertically. We propose that properly designed PDNs* (Power Delivery Networks) can significantly contribute to heat removal in both parallel (think TSV and xx um thick Si layers) and monolithic/sequential (think 100nm Si layer) 3D-ICs.

We investigated both parallel and monolithic in the paper. Here, I will, of course, focus more on the monolithic challenges and solutions, but I will make some important comparisons to parallel at the end.

Since the 130nm node, we have entered an era in our industry where we are not only using new materials, but also new device structures. I have **written previously** about the risk associated with this, and (hopefully…) made a case for monolithic 3D technology being the best way for the industry to move forward, still enjoying Moore's Law type economics (i.e., lower cost) but with a much lower development risk.

Life is getting thin and narrow in our business….so, how best to take advantage of this nanometer and angstrom era and avoid the economic (think EUV at 110+M$ a pop, or double/quad patterning) and atomistic (think 7 nm) brick walls coming? Monolithic 3D stacking technology is the answer: keeping the next evolutionary step of our industry in the wafer fab, where the batch economics of the silicon wafer can be enjoyed, and avoiding the costly piece-part assembly processes of TSVs.

One of the basic tenets of monolithic 3D is the ability to have thin (preferably monocrystalline) silicon layers that enable very small vertical interconnect

manufacturing, and hence a large (>1 million/cm2) layer to layer vertical interconnect density in the stack. This opens up the possibility for powerful new architectures and devices, such as Amdahl's wafer scale computer (see **blog**, **website**,**technology**) and cost effective MLC 3D **memories**.

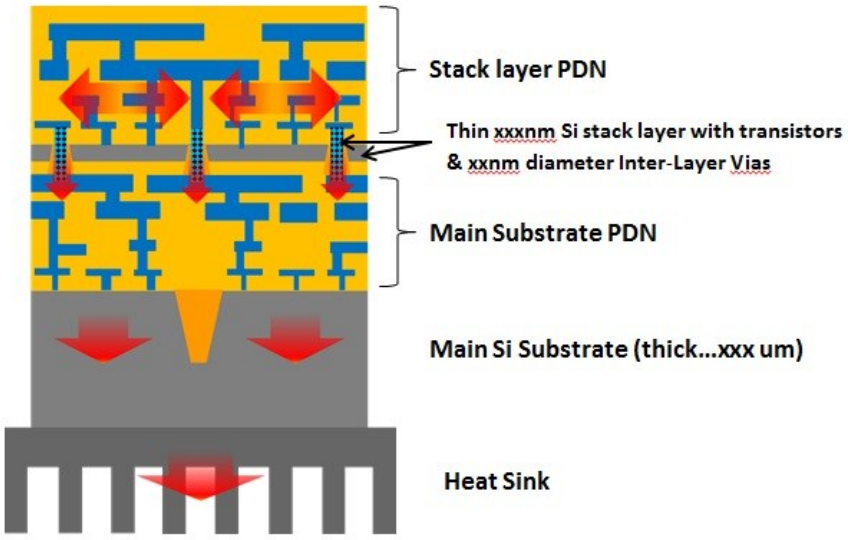
Two implications arise from the thin (on the order of 100nm or less) silicon layer stacking. First, that fully depleted (FD) devices, and hence silicon islands floating in an insulator such as silicon dioxide, will be the norm. Second, taking full advantage of a manufacturable aspect ratio etching (5:1 to 10:1), we will end up with a large density of very small layer to layer vias (of 1-2 lambda diameter), where vertical interconnect density rivals the horizontal density of interconnect that we have enjoyed thru the many cycles of **Dennard** scaling. FD devices are soon to be the norm in 2DICs; for example, the thin **UTBBOX** of STMicro/GlobalFoundries and the narrow **FinFets** of **Intel**/**TSMC** (incidentally, at IEDM12, Intel was criticized for doping the fins…).

Both of these implications, FD devices in islands of Si and very dense vertical interconnect, play a significant role in how we propose to solve a major challenge in 3D stacking.

Since the stacked layers are not in direct contact with the heat sink:
How do we ***get the heat out*** of the stacked layers???

In short, the answer is to take the heat out of each silicon island with the power delivery network, move it laterally in the metal interconnect of that stack layer (just *as if* we had a thick silicon layer underneath), and then vertically move the heat to the heat sink with that large density of interlayer vias (which we can now make due to the thin stacked layer being very thin).

Here's a picture of what we are doing:



Sounds at least plausible, right?

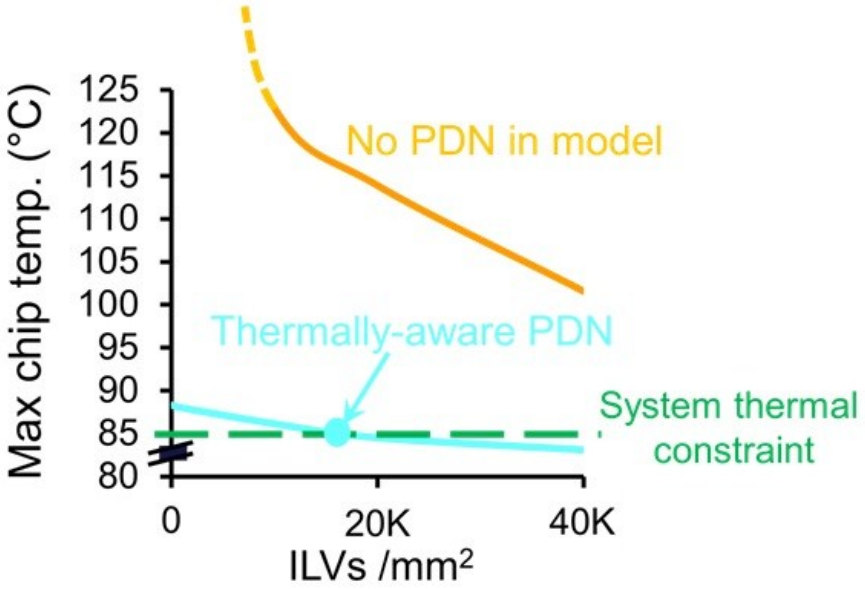
Well, that's what we set out to show, with the heavy lifting done by our friends at Stanford. Hai Wei & Tony Wu of Professor **Subhasish Mitra's group**, Professor Mitra, and Professor **Fabian Pease**, were the drivers in creating the simulation approach and engine to see if this works as we thought it might. It did, and then ended up developing a tool that may be very useful for future 3DIC design work.

Hai and Tony describe in the paper and the presentation the details of the simulation approach, engine, assumptions, and methodologies developed. Quite a nice piece of work! They have built an analysis framework that can be adapted for exploring technology-circuit-application interactions for a wide variety of 3D technologies, cooling options, and PDN designs. Types of **3DIC** technologies modeled are conventional **TSVs**, called parallel 3D integration by many in the industry, and monolithic 3D integration, a type of sequential 3D integration. Cooling options range from conventional air cooling of the heat sink (2 W/K·cm2) to external liquid cooling (10 W/K·cm2) for high power systems. PDN designs studied ILV densities from 0 to 4 million/cm2.

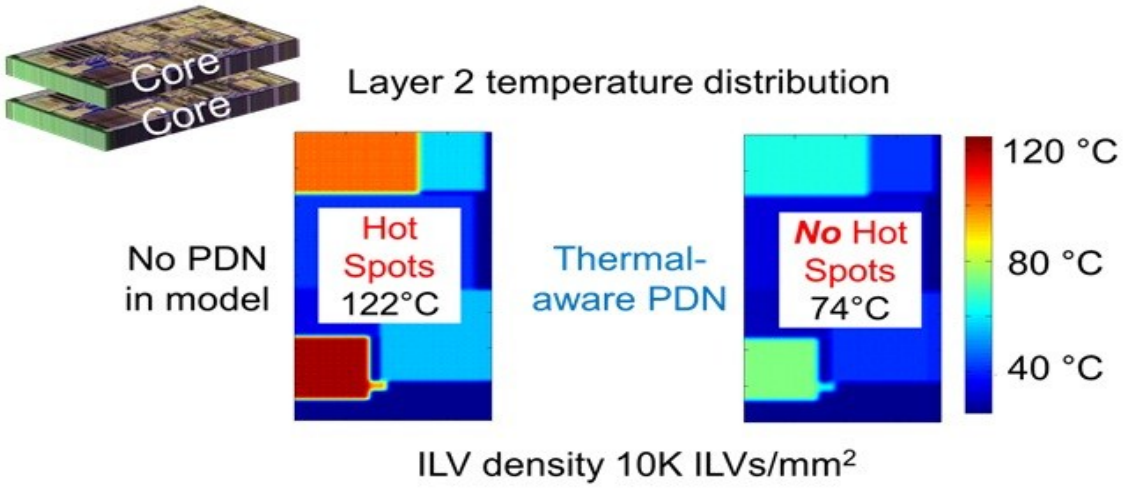
That said, what are the essential takeaways?

First, the cooling benefits of PDNs are essential to achieve monolithic 3D integration. Without accounting for PDNs in the 3DIC thermal model, it will be next to impossible to achieve the desirable thermal characteristics and result of a 3D IC stack.

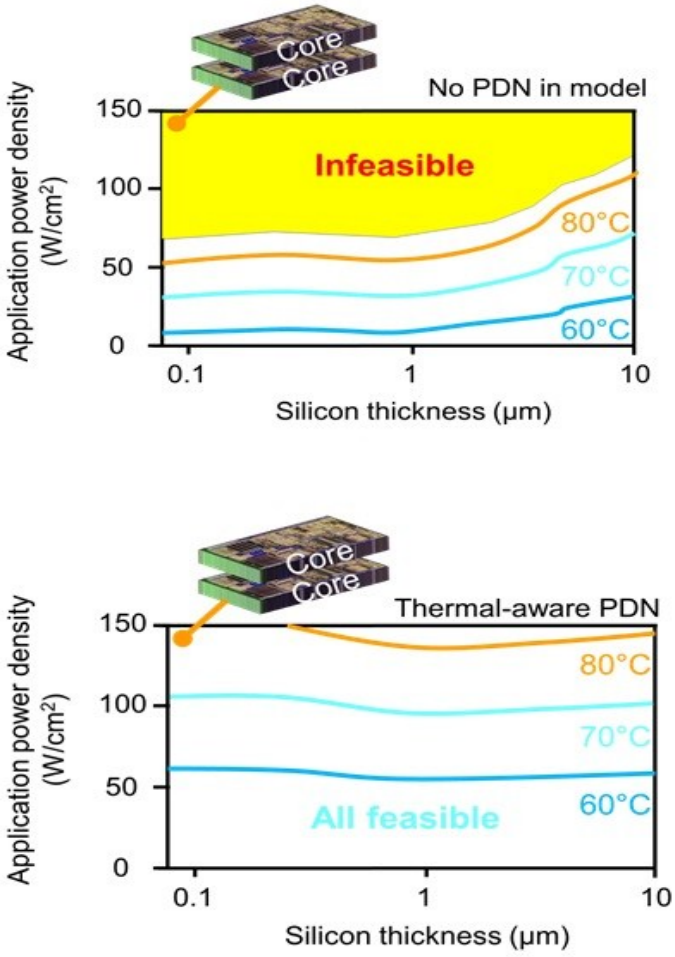
Further, the density of ILVs is important to achieving the system thermal constraint. In the 100nm thick Si example below, the desired maximum chip temperature is 85°C or less.



Second, a processor can be effectively cooled, with no hot spots, using PDNs in a monolithic 3D configuration. Hai and Tony's thermal analyses of core-on-core and memory-on-core designs, utilizing the **OpenSPARC T1** industrial multi-core design operating running an 8-threaded program that solves the **Black-Scholes** application (i.e., hot), showed significant improvement and no hot spots. The top silicon layer is 100nm thick and the hottest parts of the chips were operating at 138 W/cm2. Those hottest parts, the EXU units, were stacked directly on top of each other to show the worst case.



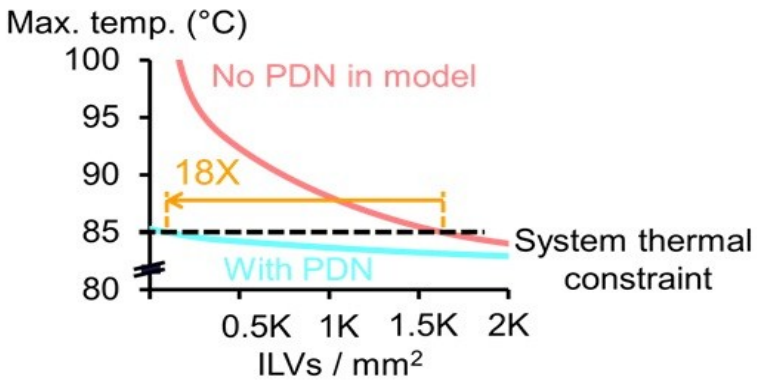
Combining these two seems to indicate that no PDN in the model versus designing and optimizing with thermal-aware PDNs makes the difference between being able to run the design (processor on processor in this example) at only 1/3 of the full power density or at a full power.



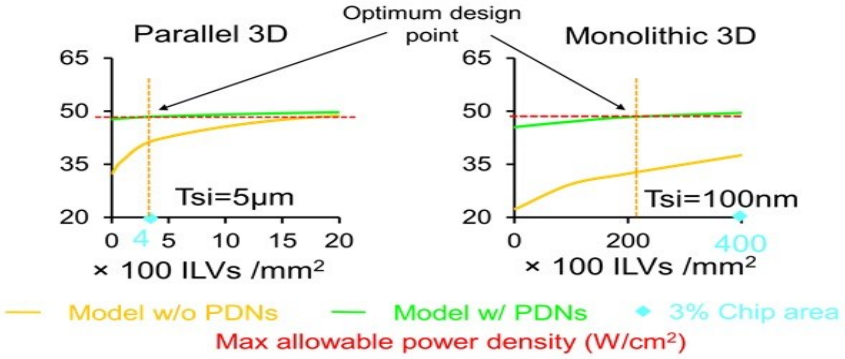
That's the essential take-away for monolithic. Mimic the lateral heat conduction of thick silicon with the PDNs of the thin silicon stack layer, and then get that heat vertically to the heat sink with the dense network of vias provided by the monolithic 3D integration.

For the parallel 3D integration case, the 5um thick silicon greatly helps with the lateral heat conduction to the TSVs. With a properly designed PDN; however, there can be a significant savings in the number of TSVs (ILVs on chart below) used to vertically conduct the heat away, and thus offers a significant area savings by eliminating many of those big TSVs and Keep Out Zones (KOZs). (Note: for both the parallel and monolithic cases, Hai made the KOZ twice the ILV diameter as a conservative choice)

- Parallel 3D IC: area benefit



Moreover, by use of a properly designed PDN and an optimized density of TSVs, the maximum power density of the top layer in can be increased considerably …. from 35 to 50 W/cm2 for the parallel 3D case.



It is worth noting an important point from these graphs: At the optimum design point, where the density of ILVs coupled to the PDN satisfies the desired 50W/cm2 max allowed power density, the required number of TSVs to effectively conduct the heat costs about 3% of the chip area. For the monolithic case, the chip area cost is about half that.

A high density of small vias not only makes possible some powerful product architectures such as logic-cone level redundancy, but is also key to producing area efficient vertical heat conduction networks.

BC

*Patent Pending technology

Chapter 12 - 3D NAND Opens the Door for MonolithIC 3D

by Israel Beinglass, the CTO of MonolithIC 3D Inc.

NAND technology, which is a subset of NVM (Non Volatile Memory), was invented by Fujio Masuoka of Toshiba back in 1984. Flash memory was presented at IEDM1984 by Dr. Masuoka and his colleagues [1]. The following is a short quote from the original paper "the cell is programmed by a channel hot carrier injection mechanism similar to EPROM. The contents of all memory cells are simultaneously erased by using field emission of electrons from a floating gate to an erased gate in a FLASH (Hence the name FLASH)".

Masuoka came back to the IEDM in 1987 and suggested a Flash NAND structure [2].
Intel created the first commercial NOR type of Flash chips in 1988. For the next few years some major developments occur in the Flash arena:

- In 1989, Samsung and Toshiba created a NAND flash memory.
- In 1994, Compact Flash was invented and introduced by SanDisk.
- In 1999, the SD memory card was released by a combination of SanDisk, Toshiba and Matsushita.
- In 2001, the world's first 1 Gigabit Compact Flash card was introduced.

From 2006 onwards, NAND became the most scaled of devices beating out the microprocessor devices (see Figure 1). The current state of the art is 20nm (2x) technology, as the world's appetite for storage is still strong. Flash Cards, SSD, Smartphone and Tablets are the leading growing applications.

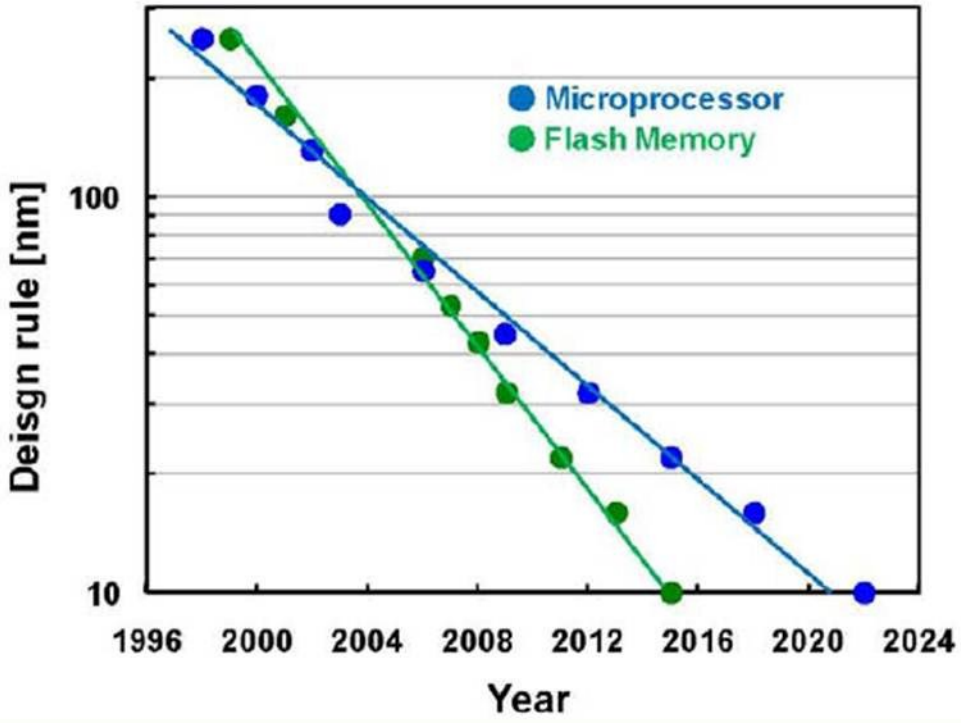


Figure 1: Flash Vs. Microprocessor design rules cross over

NAND memory as a true cross point array with the control gate on top of the floating gate and only one contact for a whole string of cells has the smallest memory cell size as shown in Figure 2 In addition, when one adds with the capability of MLC (Multi Level Cells) to NAND devices, the bit density dramatically increases.

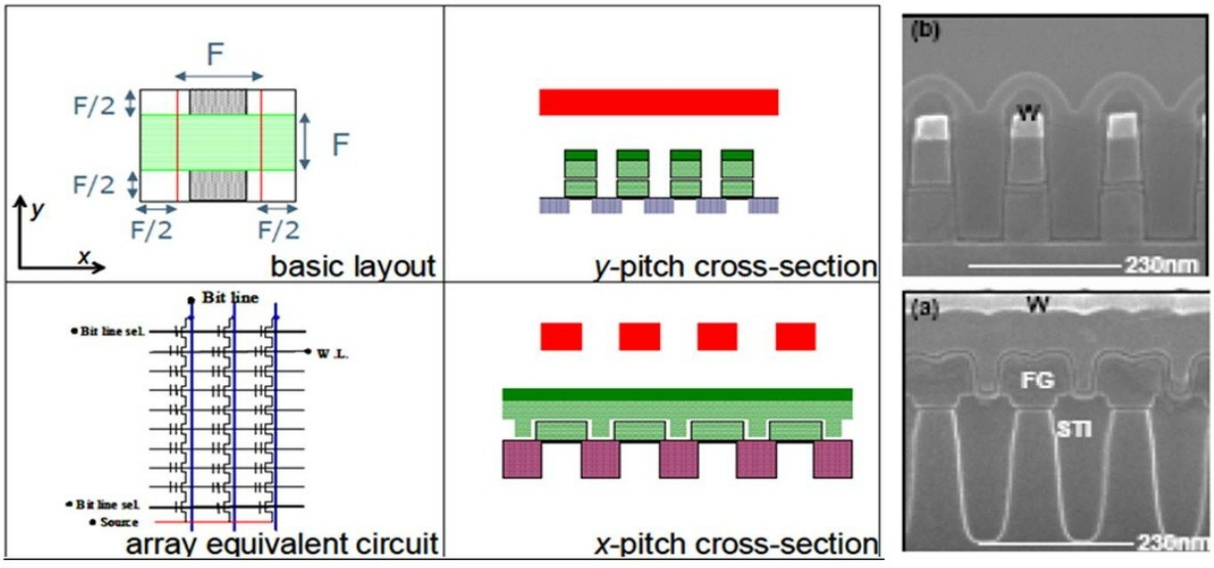


Figure 2: NAND, circuit diagram and SEM pictures in x and y directions.

The NAND market has been continuously growing for the last several years. Figure 3 shows the NAND revenue and Gigabytes increase since 2008 and the forward projection for the years 2012-2016.

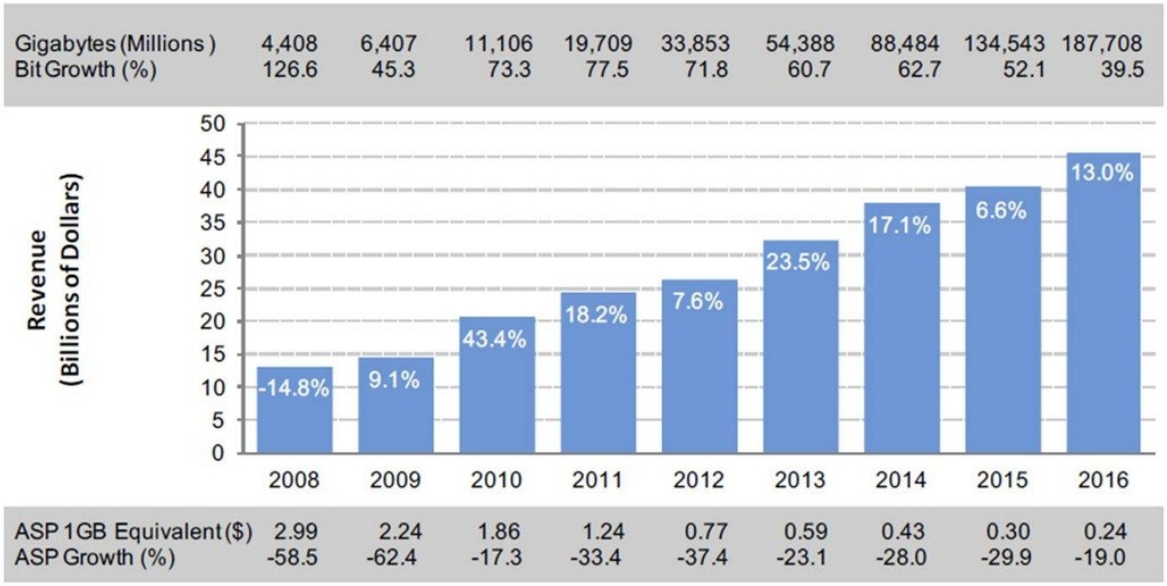


Figure 3: NAND Revenue and Gigabytes growth

As the NAND technology has been moving to smaller and smaller process nodes some serious problems, physical and electrical surfaced:

Physical Limitations:

- Pattern scaling - lack of EUV is a major issue
- Structure formation, Figure 4 depicts a 27nm NAND cell that shows how close the cells are getting to each other, and how much the aspect ratio is getting out of hand. This is a limiter to obtaining high yield.

Electrical Limitations:

- There is an increase in cell-to-cell interference in the word lines.
- Capacitive coupling ratio has decreased
- Dielectric leakage has increased

The number of electrons on the floating gate has decreased dramatically so much so that a small fluctuation in the number on the floating gate can make a huge effect on the cell function. Figure 5 describes the scaling induced phenomenon.

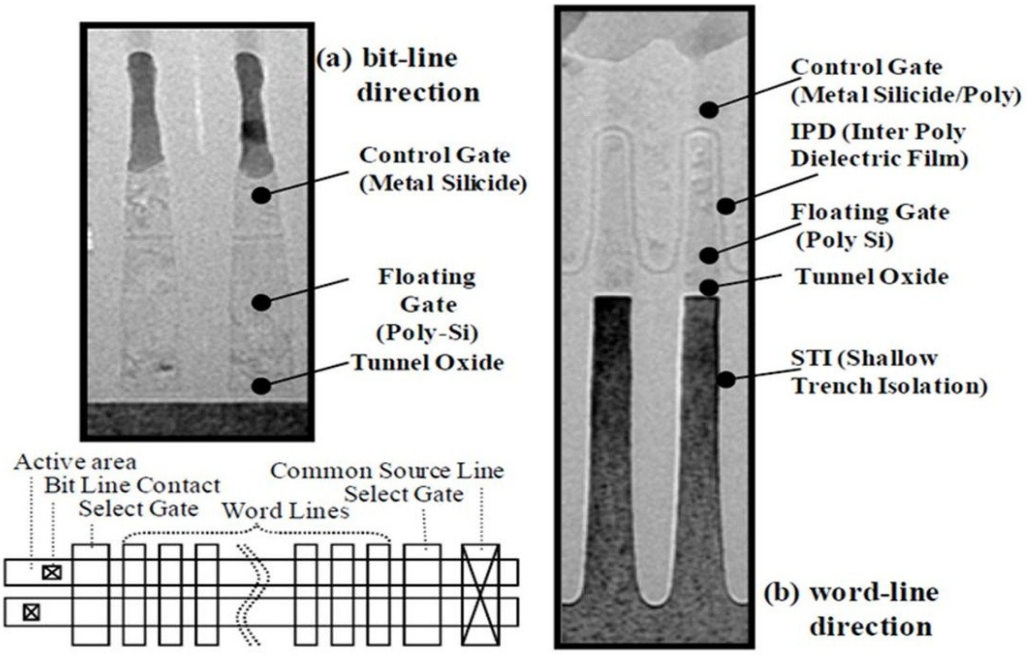


Figure 4: A 27nm NAND cell structure

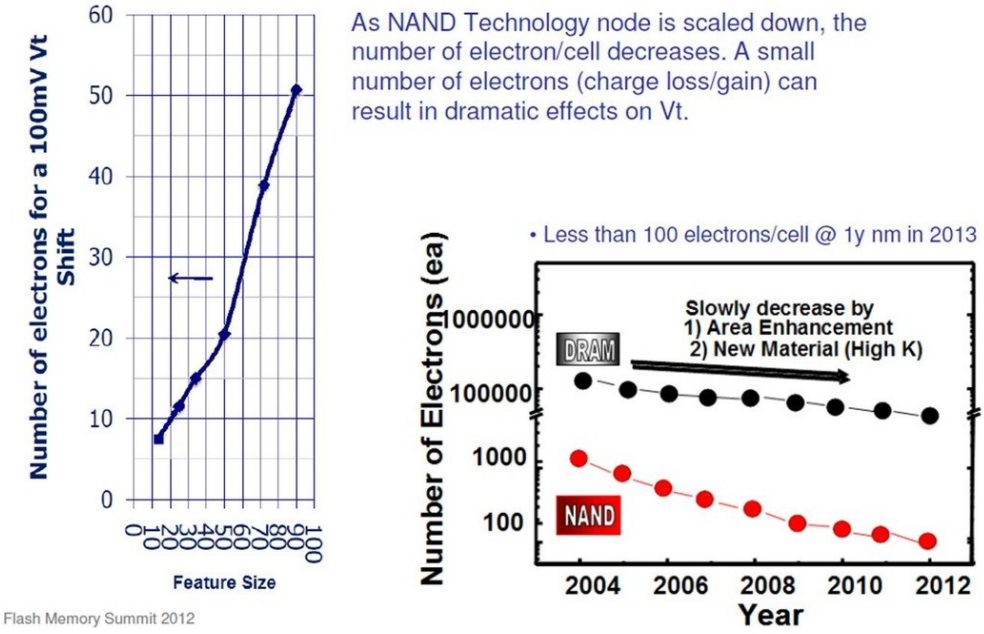


Figure 5: Number of electrons on the FG decreases for advanced NAND technology nodes

It is a common understanding among the experts that the current NAND technology will not be able to be scaled down to the 10nm node.

The solution for this dilemma is the 3D NAND, which was initially proposed by

Toshiba at the 2007 VLSI Symposium [3]. Toshiba unveiled its Bit Cost Scalable (BiCS) technology. BiCS makes use of a “punch-and plug” structure and charge trap memory films. Toshiba has fabricated a prototype 32-Gbit BiCS flash memory test array with a 16-layer memory cell using 60nm design rules, see Figure 6. Hynix, Samsung and Macronix have also come with their versions of the 3D NAND.

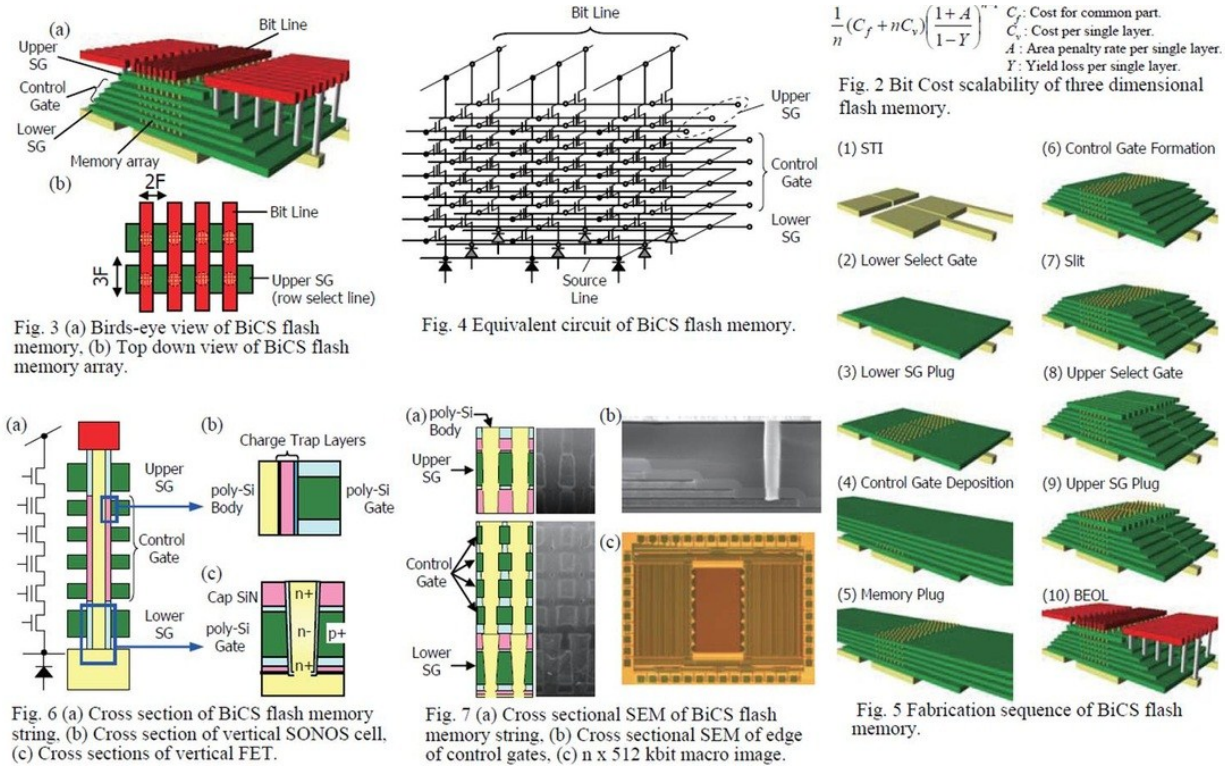


Fig. 3 (a) Birds-eye view of BiCS flash memory, (b) Top down view of BiCS flash memory array.

Fig. 4 Equivalent circuit of BiCS flash memory.

Fig. 2 Bit Cost scalability of three dimensional flash memory.

Fig. 6 (a) Cross section of BiCS flash memory string, (b) Cross section of vertical SONOS cell, (c) Cross sections of vertical FET.

Fig. 7 (a) Cross sectional SEM of BiCS flash memory string, (b) Cross sectional SEM of edge of control gates, (c) n x 512 kbit macro image.

Fig. 5 Fabrication sequence of BiCS flash memory.

Figure 6: 3D NAND process steps, as described by Toshiba

The following are the key advantages of the 3D NAND:

- With 3D NAND, scaling is no longer driven by lithography. The gate length is defined by deposition
- The key steps to 3D NAND are

  - Build a multitude of oxide/nitride or oxide/doped polysilicon stacked layers
  - Fill the deep memory holes or trench slits. The top foreseeable challenges are ultra-high-aspect ratio (>40:1) conductor etch and dielectric etch with high etch selectivity to the hard mask

- 3D NAND is relatively straightforward for a DRAM maker since it has stacked SiO2 and polysilicon layers like a stacked capacitor DRAM, and trenches like a trench cell DRAM.

- 3D NAND is evolutionary, not revolutionary.
- The good news is continued cost reduction, smaller die sizes and more capacity.
- Installed NAND toolsets in the wafer Fabs can, for the most part, be reused, thereby extending the useful life of Fab equipment.
- 3D NAND technology is still basically NAND with all its inherent limitations of data reliability and performance: hence, generally well understood (evolutionary).

At this point all the NAND companies are putting a lot of effort to bring this process to high volume manufacturing; the current expectations are that in 2014-2015 it will be ready for prime time. 3D NAND will be a technology that will take us between the 2D planar NAND and whichever post-NAND technology emerges in the future.

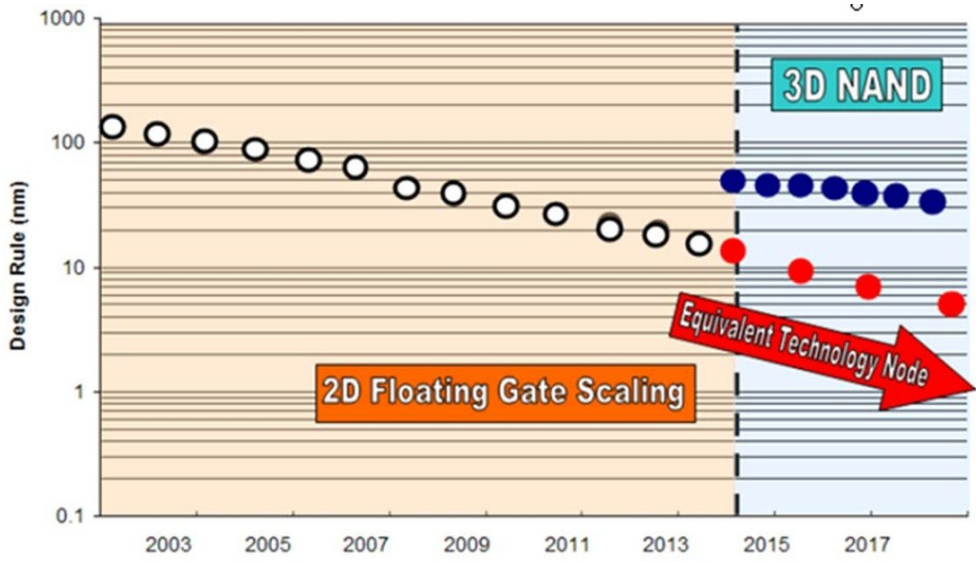


Figure 7: 3D NAND effect on design rules

Figure 7 describes the essence of the advantage of moving from 2D to 3D NAND. The adoption of 3D NAND technology will remove the burden from the Litho (and hence EUV) into the much easier process steps (deposition). Of course there are other advantages as described above.

It is not too difficult to see the similarity between the up and coming 3D NAND and the Monolithic 3D approach. As we describe in our web site (**www.monolithic3d.com**) the advanced technology patented by MonolithIC 3D Inc. enables the fabrication of Monolithic 3D Integrated Circuits with multiple stacked transistor layers and ultra-dense *vertical* connectivity. Thus, it appears *monolithic 3D-ICs with 2 device layers provide benefits similar to a generation of conventional scaling*. Furthermore, just as conventional scaling reduces feature sizes every generation,

monolithic 3D opens the road for many years of continuous scaling by 'folding' once, twice, and so forth *without necessarily reducing feature sizes*.

1. F. Masuoka et. al IEDM 1984 pp464-467
2. F. Masuoka et. al IEDM 1987 pp552-555
3. H. Tanaka et al., Symp. on VLSI Tech. Dig., pp 14-15, 2007

Part 2: 3D-CMOS: Monolithic 3D Logic Technology

Chapter 13 – The Way and How of Fine-Grain 3D Integration

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

The Silicon Valley IEEE Components, Packaging and Manufacturing Technology (CPMT) Society invited me to give a talk on "Fine-Grain 3D Integration" last week. In case you're not familiar with this IEEE chapter, they host speakers from around the Valley periodically. Check out **their website** if you get a chance - they have some nice talks lined up for the future. Now, let me describe the stuff I presented there.

Introduction

As many of you know, 3D technologies in the marketplace today have huge TSVs. For example, TSMC's 28nm technology has 6um diameter TSVs with 5um keep-out zone. Other manufacturers are offering similar TSV sizes too. When you start comparing these with on-chip feature sizes (28nm), you'll understand why I use the term "huge" to describe these TSVs. In contrast, fine-grain 3D technologies are defined as those having TSV pitches smaller than 500nm.

Why Fine-Grain 3D Integration?

There are many applications that benefit from small TSV sizes. Fig. 1 describes the basic motivation - wires consume a lot more energy than transistor-based computation today, and 3D can reduce lengths of these wires. Micron-scale TSVs can reduce chip-to-chip wire lengths, but smaller TSVs are needed to reduce on-chip wire lengths.

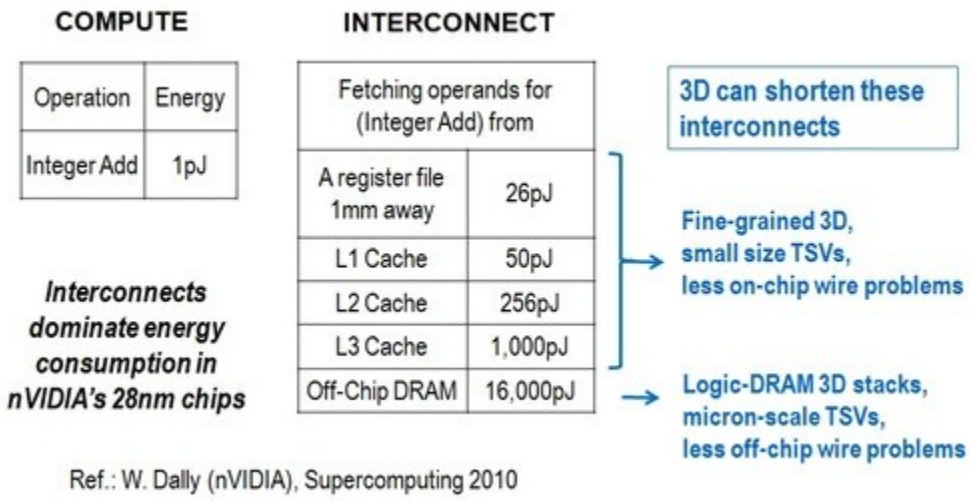


Figure 1: Situation in nVIDIA's 28nm chips.

Below are some uses for fine-grain 3D. Note that small TSV sizes (around minimum feature size) are required for some of these applications:

- Short on-chip wires in logic cores and SoCs: Components within a single logic chip can be stacked atop each other to shorten on-chip wires. This leads to smaller gates, since these gates need to drive less wire capacitance. The result is reduced power and die size. Analyses show that a 2x reduction in power, a 2x reduction in silicon area and a 4x reduction in chip footprint may be possible by doubling the number of 3D stacked layers (**link**).

- Logic-SRAM stacking: The requirements of logic devices and SRAM on a chip are very different today. SRAM circuits typically require just 4 metal levels compared to 12 for logic circuits. SRAM transistors have different channel length, oxide thickness and threshold voltage compared to logic transistors too. In this scenario, it makes sense to stack SRAM and logic in 3D. The SRAM layer can be optimized for 4 metal levels and SRAM-type transistors, thereby saving cost.

- nMOS and pMOS stacking: Today's nMOS and pMOS transistors have different gate stacks, strain layers, implants and wells. Separate lithography steps are required for all of these. To save cost, one could stack the nMOS and pMOS atop each other. This reduces standard cell area too. **Analysis from IBM** shows that 30-40% reduction in standard cell area is possible for inverters, NAND and NOR cells by stacking nMOS and pMOS layers atop one another. Smaller standard cells result in shorter wires, improving power and performance.

Limitations of today's TSV technology

Like many engineers, I believe understanding a problem is important for figuring out a solution. So, let's analyze why today's TSVs are so fat. Fig. 2 shows a typical process for high-density 3D-ICs.

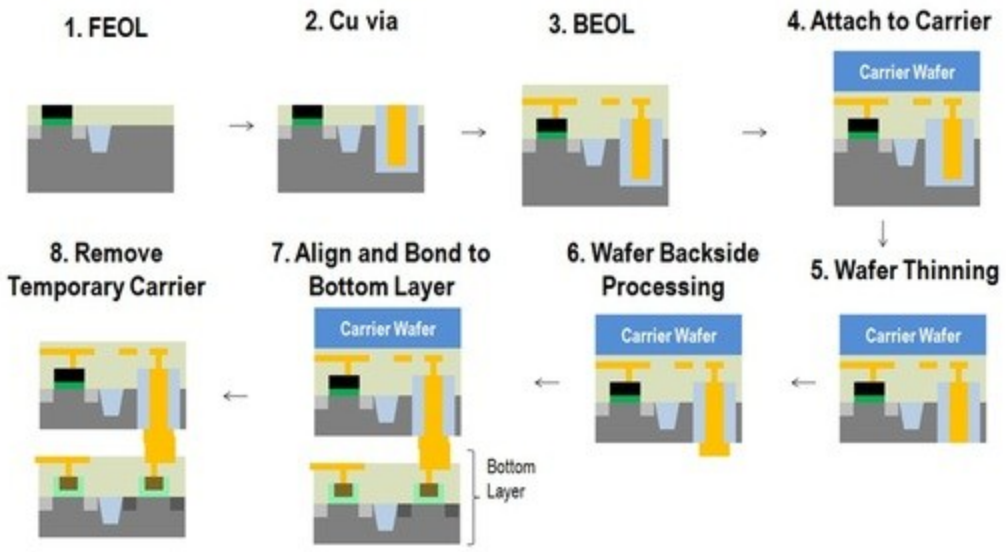


Figure 2: Process flow for a bumpless bonded 3D TSV technology.

The limiting steps for TSV size in these face-to-back bonded technologies are:

Step 5: Wafer thinning - Aspect ratio limitations of TSV manufacturing processes nowadays are around 10:1. To get 1um diameter TSVs, one needs to have a 10um thick silicon layer. For this scenario, during the thinning step, a 775um thick wafer needs to be thinned down to 10um +/- 1um (10% tolerance). This 1um tolerance is very hard to achieve at high throughput. Many manufacturers take the easy way out and thin the silicon wafer from 775um to 50um +/- 5um (10% tolerance). For an aspect ratio of 10:1, a 50um silicon thickness will lead to 5um diameter TSVs.

Step 7: Wafer alignment - In this step, the top and bottom layers are aligned with each other and bonded. Misalignment occurs due to **several reasons**:

- 3D align and bond tools on the market often do not have the stable alignment stages and image capture/storage required for sub-500nm pitch TSVs.
- Co-efficient of thermal expansion (CTE) mismatch between the top and bottom layers, wafer bow, thermal and stress induced flow of temporary bonding adhesives, localized bonding imperfections and other issues can cause um-scale misalignment.

Evolutionary Improvement of Today's TSV Technologies

In this section, I will summarize evolutionary ways to improve today's TSV technologies. IBM and MIT Lincoln Labs are the pioneers in this area, as are image sensor makers such as Sony and Omnivision.

Wafer thinning techniques - Fig. 3 shows approaches to reduce wafer thickness from 775um to less than 1um. The method in Fig. 3(a) works for SOI wafers. Buried oxide layers of SOI wafers are used as etch stops to get low silicon thickness with sufficient precision. An alternative approach for bulk silicon wafers is shown in Fig. 3(b). Silicon etch solutions such as EDP have orders of magnitude lower etch rates for p++ silicon compared to p silicon. One could therefore use a p++ layer in a silicon wafer as an etch stop. Both these techniques are starting to be used in manufacture of back-side illuminated image sensors.

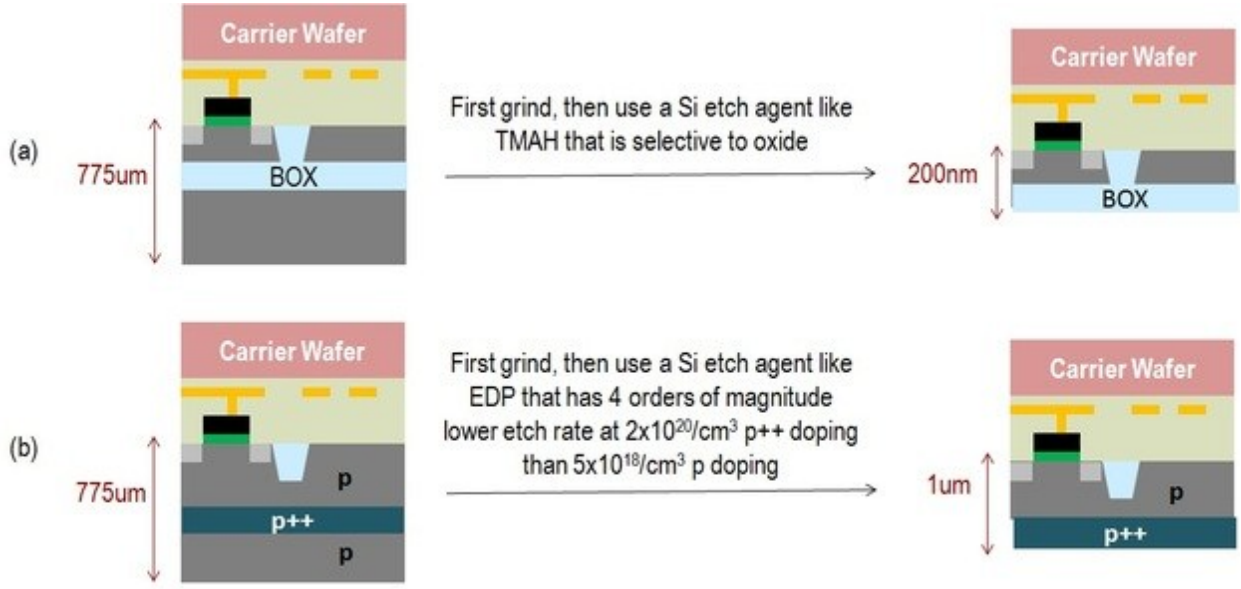


Figure 3: Next generation wafer thinning technologies that use etch stop layers.

Techniques to improve alignment accuracy - For high density TSVs, companies prefer to use glass carrier wafers at present. The transparency of glass, combined with low silicon thickness of transferred films, allows one to look through the top wafer and align. Limitations of 3D alignment tools can be overcome with this technique. In addition, if glass carrier wafers are used, adhesives for attaching silicon to a carrier wafer can be optically debondable. Optically debondable adhesives are more stable at the high temperatures needed for bumpless bonding.

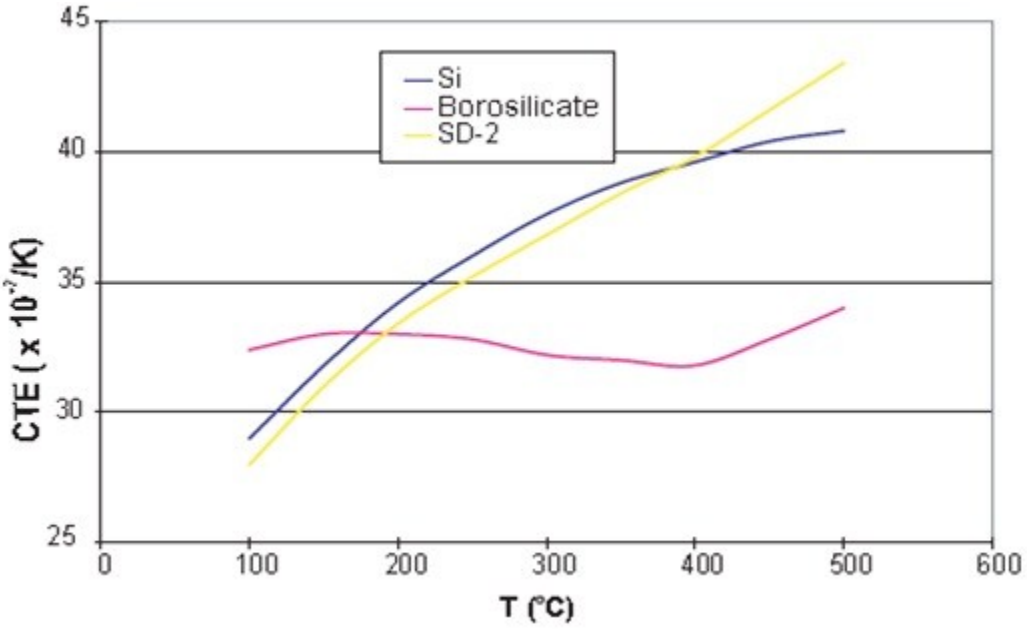
Besides using glass carriers, one could do a few more things:

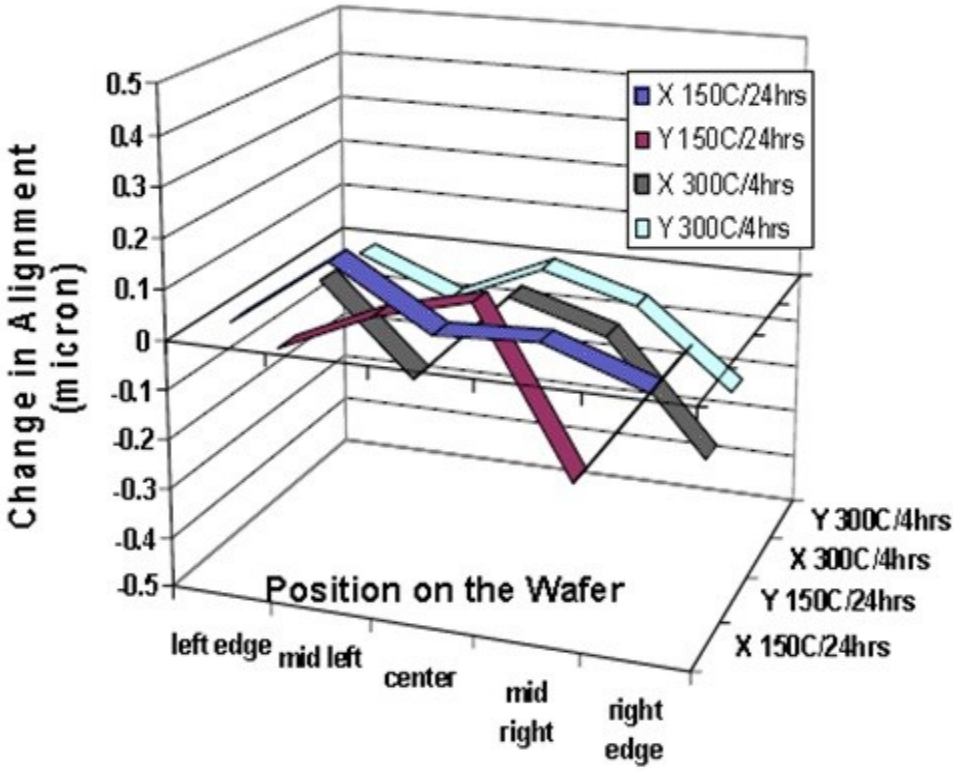
- Use CTE matched carrier wafers - Even if you use borosilicate glass with an excellent CTE match with Si, a small CTE mismatch is introduced at bond temperatures. For example, at 300C, silicon wafer diameter can increase by 314um while borosilicate glass diameter can increase by 264um. This difference in diameter can introduce alignment error. If you want to get sub-500nm pitch,

costlier glasses that have CTE-match with silicon at various temperatures are required (Fig. 4(a)).

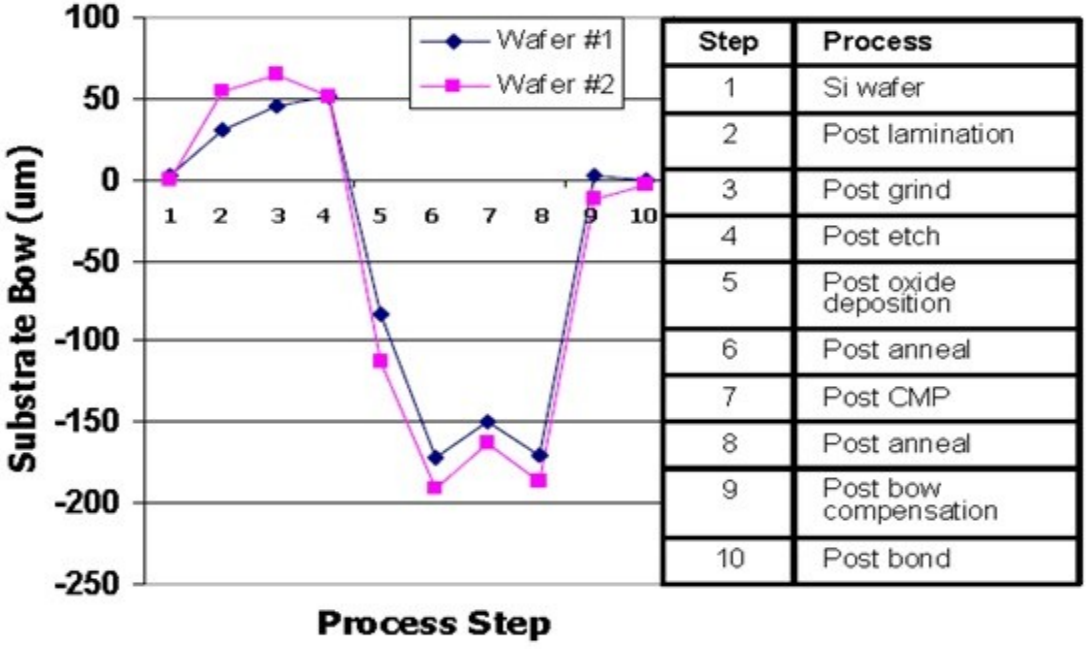
- Use oxide-to-oxide bonding - For fine-grain 3D, oxide-to-oxide bonding is the technique of choice due to the low temperatures involved vs. Cu-Cu bonding. Lower temperatures reduce CTE mismatch errors. In an oxide-to-oxide bonding process, a weak bond is formed at room temperature. Following this, a post-bond anneal (~300C) is done to get a stronger bond. The alignment got at room temperature is largely maintained. Less than 400nm misalignment is introduced by the post-bond anneal (Fig. 4(b)).

- Use wafer bow compensation - Wafers can frequently have bow of 50-100um, making sub-micron alignment accuracy difficult while bonding. IBM and MIT have developed wafer bow compensation schemes to reduce this. For example, one could deposit thin films on back sides of wafers to compensate partially for the wafer bow. See Fig. 4(c).





Ref.: [A. Topol, et al., IEDM 2005]



Step	Process
1	Si wafer
2	Post lamination
3	Post grind
4	Post etch
5	Post oxide deposition
6	Post anneal
7	Post CMP
8	Post anneal
9	Post bow compensation
10	Post bond

Ref.: [A. Topol, et al., IEDM 2005]

Figure 4(a)-(c) from left to right: (a) CTE match of various glasses with silicon. (b) Change of alignment after post-bond anneal. (c) Wafer bow compensation schemes.

IBM built prototypes utilizing many of these techniques. SOI wafers and buried oxide etch stop layers enabled transfer of thin silicon. CTE-matched borofloat glass carriers, oxide-to-oxide bonding and wafer bow compensation schemes were used. ***IBM's best prototypes had a TSV pitch of 6.7um, and they said 2um pitch would be possible when bonders with sub-0.5um alignment accuracy are available*** *(which is the case today). Essentially, we can reduce TSV pitches from the 20um we get in the marketplace today to around 2um. I believe it may be possible to lower TSV pitches to less than 500nm by improving processes further. Please see* ***slides of my talk*** *for details.*

The Monolithic 3D Path

With monolithic 3D technology, additional transistor layers are constructed monolithically atop Cu/low k layers. *This could lead to TSV size close to minimum feature size, which is needed for many of the fine-grain 3D applications described above.* Fig. 5 indicates the main barrier to creating high-quality transistors at Cu/low k compatible temperatures (sub-400C) is dopant activation.

	Sub-400°C possible?	Method
Single Crystal Silicon	Yes	Ion-Cut
Shallow Trench Isolation	Yes	Radical Oxidation, HDP
High k/Metal Gate	Yes	ALD/CVD/PVD
S-D Dopant activation	No	>750°C anneal
Contacts	Yes	Nickel Silicide

Figure 5: Steps required for constructing a silicon transistor.

Fig. 6 describes one approach to overcome this problem, which utilizes recessed channel transistors. These have been used in DRAM manufacturing since the 90nm node, and **are known to be competitive with standard planar transistors**. As can be seen in Fig. 6, high temperature dopant activation steps are conducted before transferring bilayer n+/p silicon layers atop Cu/low k using **ion-cut**. For ion-cut, hydrogen is implanted into a wafer at a certain depth creating a defect plane. Following this, the wafer is bonded to the bottom device layer using oxide-to-oxide bonding. The bonded structure can now be cleaved at the hydrogen plane using a 400C anneal or a sideways mechanical force. CMP is done to planarize the transferred surface. Transferred layers are unpatterned, therefore no misalignment issues occur while bonding. Following bonding, sub-400C etch and deposition steps are used to define the recessed channel transistor. This is enabled by the unique structure of the device. These transistor definition steps can use alignment marks of the bottom Cu/low k stack since transferred silicon films are thin (usually sub-100nm) and transparent. Minimum

feature size through-silicon connections can be produced due to the excellent alignment.

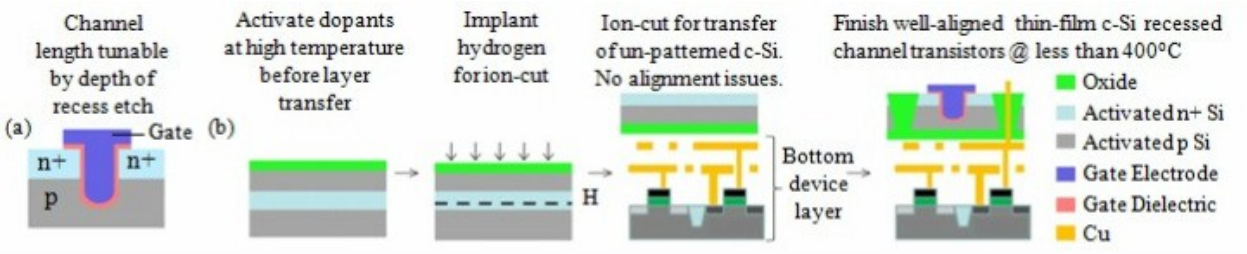


Figure 6: (a) A recessed channel transistor (b) Process flow for monolithic 3D logic. Bottom device layer with Cu/low k does not see more than 400C. Through-silicon connections can be close to minimum feature size due to the thin-film process.

A few points about Fig. 6: (i) All materials, process steps and device structures are well-known and are used in high-volume manufacturing (ii) The original donor wafer with n+ and p layers can be reused after layer transfer. This is an advantage over today's TSV processes, where one spends time and cost etching away a 300mm wafer that costs $120. (iii) Though-silicon via connections are minimum feature size, enabling large improvements (As described previously, benefits can be 2x lower power, 2x lower silicon area by doubling the number of device layers. nMOS and pMOS stacking is possible.) The main risk is the use of DRAM-type recessed channel transistors in logic technologies. My somewhat biased view is that recessed channel transistors have been used in DRAM manufacturing since the 80nm node, so they may not be difficult for logic manufacturers to bring up and make competitive (especially for low-power applications).

Click here to view slides of my presentation at the IEEE CPMT Society

Part 3: 3D-FPGA: Monolithic 3D Programmable Logic

Chapter 14 – Three Dimensional FPGAs

by Ze'ev Wurman, the Chief Software Architect of MonolithIC 3D Inc.

Rapid escalation of access price to cutting edge technology decimated the number of ASIC designs since 2000, and directed much attention to FPGAs as ASICs replacement. Yet, despite more than a decade of predictions that FPGAs will take over the semiconductor world, this has not happened. In 2000 the FPGA market stood at less than $4B, or about 2% of the total semiconductor market, and last year it was less than $5B or only about 1.5% of the total marker. Clearly, FPGAs did not conquer the (semiconductor) world.The reasons are pretty obvious. Despite their impressive advances in speed and density, FPGAs are still **20 or more times less dense than corresponding ASICs** *(PDF),* 10 times more power hungry, and at least twice slower. Sure, there are exceptions to these for some applications, but overall the picture is not very encouraging. Since much of their disadvantages stem from the area penalty of the programmable interconnect that blows up the die size—and the corresponding power dissipation and delay, efforts have been made to take advantage of three-dimensional stacking to reduce those distance-related penalties.

Tier Logic placed the configuration memory as a second-tier TFT layer. Lin and El Gamal from Stanford **explored three-dimensional architectural FPGA variants** *(PDF)* such as in figure 1 and found potential area reduction of up to a factor of 3.2, with concomitant reduction of power and delay by up to 1.7. Le, Reda and Behar from Brown University **suggested 3D architectural partitioning across block types** *(PDF)*, such as relocating large user memories or DSP blocks to other tiers, and finding smaller potential improvements. Yet the big issue with all these ideas is the fact that nobody knows how to manufacture them: even with state of the art TSVs the vertical connectivity demands are overwhelming, while Tier Logic found that it could not resolve the reliability problems associated with TFT devices.

Recently Xilinx came up with a hybrid solution that placed multiple FPGA dies on a passive silicon interposer that connects among the logic of the FPGA dies (fig. 2). Xilinx claims large power savings by avoiding the need for full-sized off-chip drivers for the short signals on the interposer, yet, at best, this is a half-way measure rather than anything close to a true 3D IC architecture.

Finally, there is a matter of development cost. Coming out every couple of years with a new device family, with the cost to design and tapeout a dozen or more family members, while porting all the IO, PLLs and SerDes to a new technology is not a cheap operation and must cost in the hundreds of millions. Xilinx and Altera have invested years in this process and even they barely have the resources to execute it over and over again with ever-increasing technology costs. Something will have to give.

An effective three-dimensional FPGA solution will address two key issues: it will provide a solution that will take advantage of the third dimension to significantly reduce the average distances between circuits that will result in a large decrease in power dissipation and increase in performance; and it will provide a manufacturing solution so that building these device will be less expensive, and allow the reuse of older-generation analog and quasi-analog elements, which do not need to track the inexorable march of logic technology, such as IOs.

We believe we have a good solution to this problem, which we will present in the near future.

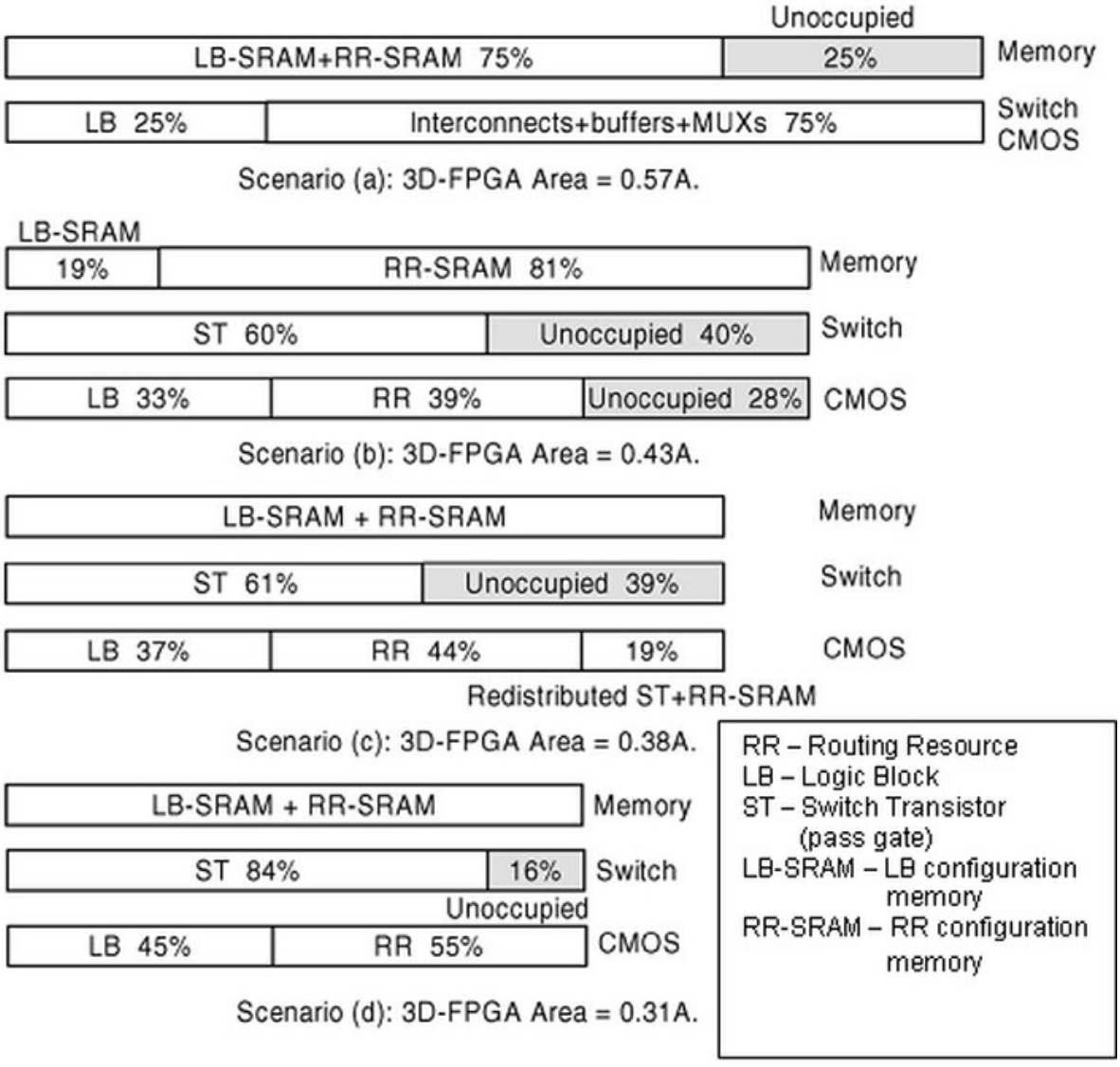
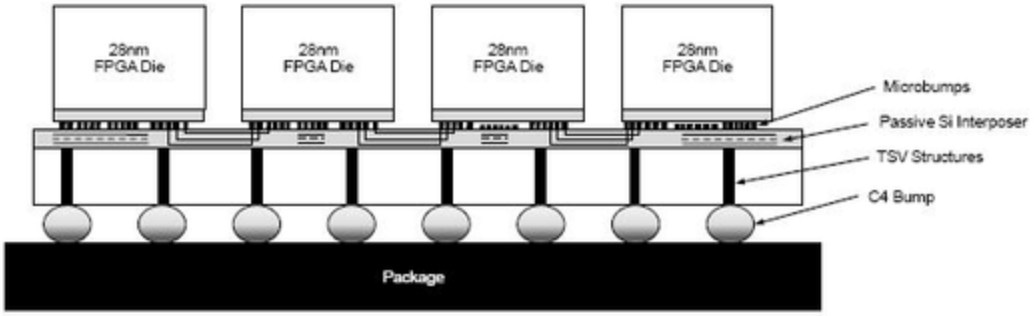


Figure 1: 3D FPGA, Lin & El Gamal, 2007.

The approach that Xilinx has taken to address the challenges of ramping up high-capacity FPGAs in 28nm technology is shown in the following figure.

FIGURE 2

Xilinx Stacked Silicon Interconnect Technology



Chapter 15 – Three Dimensional FPGAs – Part II

by Ze'ev Wurman, the Chief Software Architect of MonolithIC 3D Inc.

Last time we discussed **three-dimensional FPGAs**, it became clear that there are two major areas that block wider acceptance of current 2D FPGAs: their relative inefficiency in area (i.e., cost) power and performance as compared to ASICs, and the limited number of sizes that are offered by vendors due to the high cost associated with each family member.

The advantage of going to 3D was also discussed. Architectures such as suggested by Lin and El Gamal from Stanford, and by Le, Reda and Behar from Brown promise to reduce the footprint of a three dimensional stack by a factor of 3 or more, resulting in long wire distance reduction of more than 40%. Experiments show that also the average wire length is reduced in such cases. As we know, the majority of dynamic power dissipation in deep submicron designs resides in the interconnect power, so a shorter average wire length will directly reduce the power of such three-dimensional FPGAs. Simultaneously, the shorter long wires will also increase the FPGA performance.

But such architectures require dense vertical connectivity between device layers that TSVs cannot provide. Only now, with the true monolithic 3D technology we bring to the market, this dream may be realized.

And the story just gets better. Antifuse-based FPGAs have been on the market for many years, but their efficiency was always hampered by the large, high-voltage, programming transistors that needed to share the terrain with the logic block. Three dimensional FPGAs allow designing highly effective antifuse-based FPGAs, where the high voltage programming transistors reside in layers above and below the FPGA fabric itself. Antifuses can be as small as regular vias and allow for a much better programmable connectivity as compared to SRAM-based FPGAs. This arrangement is shown in Figure 1, with antifuses marked in red. An added advantage of the two layers of programming transistors above and below the FPGA is that the one below can program the CLB, while the one above can program the interconnect. Thus each programming path does not have to unnecessarily cross multiple metal layers and increase routing congestion.

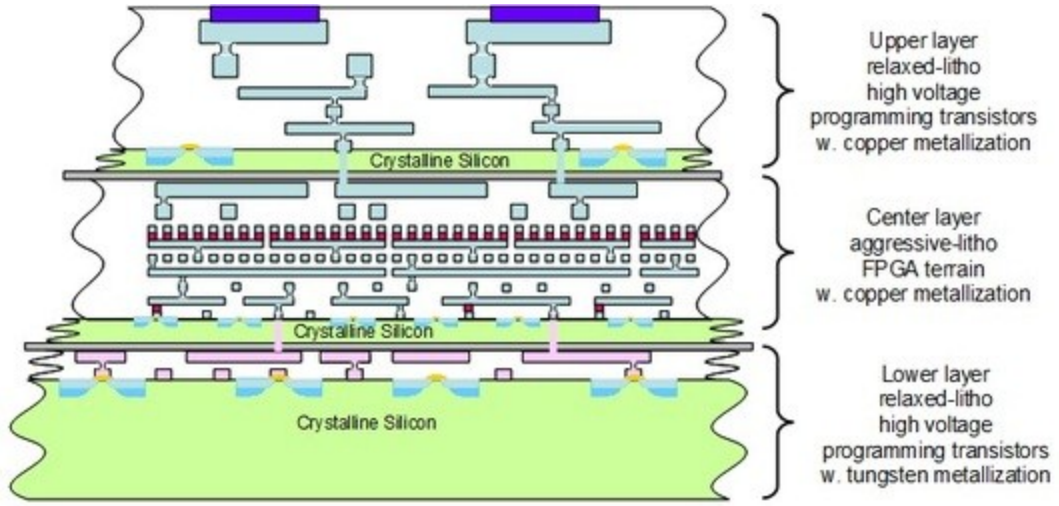


Figure 1: 3D Antifuse-based FPGA

Flexible manufacturing

As mentioned before, the cost of designing each member of an FPGA family is very high, and the cost of the whole family is prohibitive. Even Xilinx and Altera struggle with the huge R&D expenditure every other year to follow the technology curve. Yet, despite this huge investment by the vendors, most customers eventually have to pay for chips that are typically 20% bigger than actually necessary for their designs. Not only that, but each such chips will carry many additional elements that are not fully utilized – be it multipliers, SerDes circuits, memory blocks, or I/O pins.

Imagine instead if a whole wafer of an FPGA was dedicated to its logic fabric, but without any I/O. Every so often this terrain would be interrupted by a gap of perhaps 100 microns, with only long tracks crossing perpendicularly across that gap. This is a concept known as Continuous Array, and is illustrated in Figure 2.

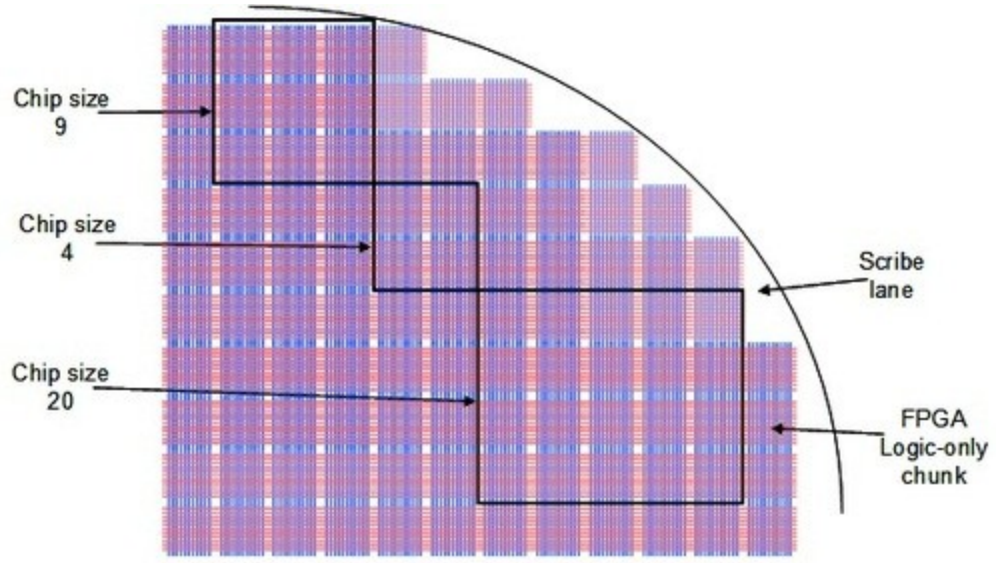


Figure 2: Continuous Array

The gaps between the “logic chunks” of FPGA terrain serve as potential scribe lanes, and based on customer demand the wafer can be diced in a variety of sizes. The top metal layer of the Continuous Array has a TSV (or microbump) prepared in a regular pattern, connected to the programmable interconnect. Now imagine that we design “chiplets” of I/O, SerDes, block memory and similar, each chiplet being of the exact physical size of the FPGA terrain logic chunk, with corresponding TSV (or microbump) pattern below, and with flip-chip bumping on its top. A customer can then specify the size of the logic needed for his design, and the type of chiplets needed to complete the design – how many I/Os, how much block memory, and how many SerDes macros. As can be seen in Figure 3, an almost infinite variety of configurations is possible with just a handful of mask sets.

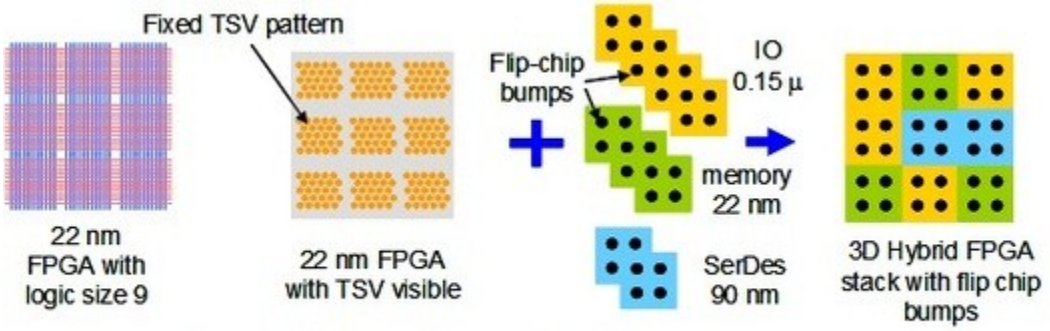


Figure 3: Continuous Array and Chiplet Assembly

Yet another advantage of this configuration is that one can reuse chiplets from an older technology over multiple generations of FPGA products. This makes it much easier to come to the market with pre-qualified IO from previous products without the tedious and difficult process of re-certification.

As we see, there are many options and savings that open up with monolithic 3D integration. For example, one can imagine a stack of single or multiple monolithic block memory layers on top of FPGA logic, topped with variety of IO chiplets, to offer a wider range of logic to memory ratios than currently available today.

Part 4: 3D-Gate Array: Monolithic 3D Gate Array

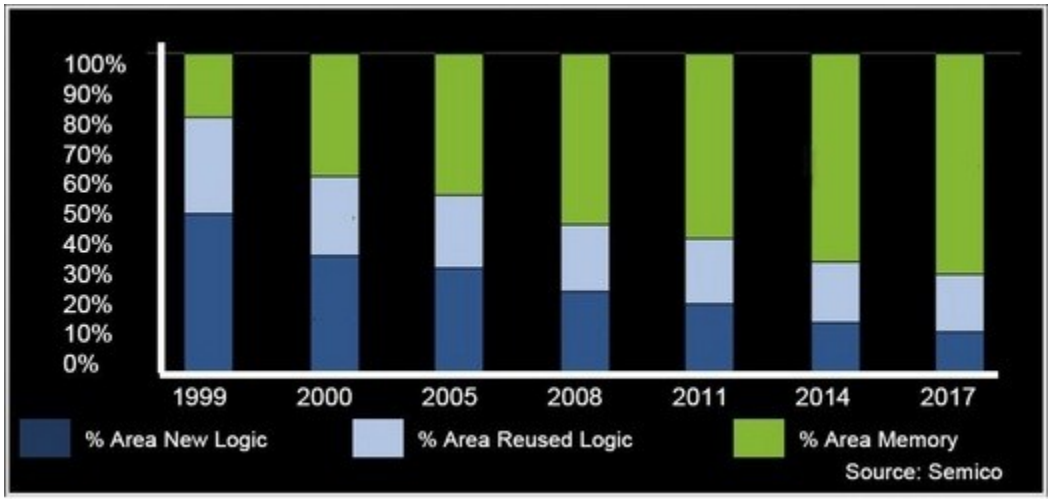
Chapter 16 – Embedded Memory and MonolithIC 3D

by Zvi Or-Bach, the President and CEO of MonolithIC 3D Inc.

Introduction

SoCs represent a significant part of the semiconductor industry ~40%. The logic market, which has SoCs and microprocessors, forms 60% of the industry.

The logic market is highly diversified and comprises hundreds of designs, yet within these devices the embedded memory portion is becoming the dominating element – see chart below:



In most cases, the embedded memory within the SoC is predominantly SRAM. In many designs, the internal memory (or as many refer to it, the eMemory) comprises hundreds of different structures, including a few large chunks of single port memories and hundreds of smaller chunks of memories, many of which are multiple port memories.

As SoC devices represent a great variety of products and market segments, there are requirements for various types of memory, including high speed, high density and non volatile. Yet due to the need for a simple manufacturing flow, the dominating memory type in most SoCs is the conventional 6 transistor SRAM.

For better illustration of the embedded memory in SoCs, lets look at embedded memory offered by Altera in their programmable devices:

Summary of Memory Features in Stratix V Devices

Feature	MLABs	M20K
Maximum performance	600 MHz	600 MHz
Total RAM bits (including parity bits)	640	20,480
Configurations (depth x width)	64 x 8 64 x 9 64 x 10 32 x 16 32 x 18 32 x 20	16K x 1 8K x 2 4K x 4 4K x 5 2K x 8 2K x 10 1K x 16 1K x 20 512 x 32 512 x 40
Parity bits	✓	✓
Byte enable	✓	✓
Packed mode	—	✓
Address clock enable	✓	✓
Single-port memory	✓	✓

Memory Capacity and Distribution in Stratix V Devices

Family	Device	MLABs	M20K Blocks	Total Dedicated RAM Bits (M20K Blocks Only) (Mb)	Total RAM Bits (Including LABs) (Mb)
Stratix V GX	5SGXA3	3,776	800	15.6	17.9
	5SGXA4	5,880	1,344	26.3	29.8
	5SGXA5	8,020	2,304	45.0	49.9
	5SGXA7	11,736	2,560	50.0	57.2
	5SGXA9	15,850	2,640	51.6	61.2
	5SGXAB	17,960	2,640	51.6	62.5
	5SGXB5	9,250	2,100	41.0	46.7
	5SGXB6	11,270	2,660	52.0	58.8
Stratix V GT	5SGTC5	8,020	2,304	45.0	49.9
	5SGTC7	11,736	2,560	50.0	57.2
Stratix V GS	5SGSD2	2,450	450	8.8	10.3
	5SGSD3	4,536	686	13.4	16.2
	5SGSD4	6,264	1,062	20.7	24.6
	5SGSD5	8,630	2,014	39.3	44.6
	5SGSD6	11,000	2,320	45.3	52.0
	5SGSD8	13,280	2,624	51.3	59.4

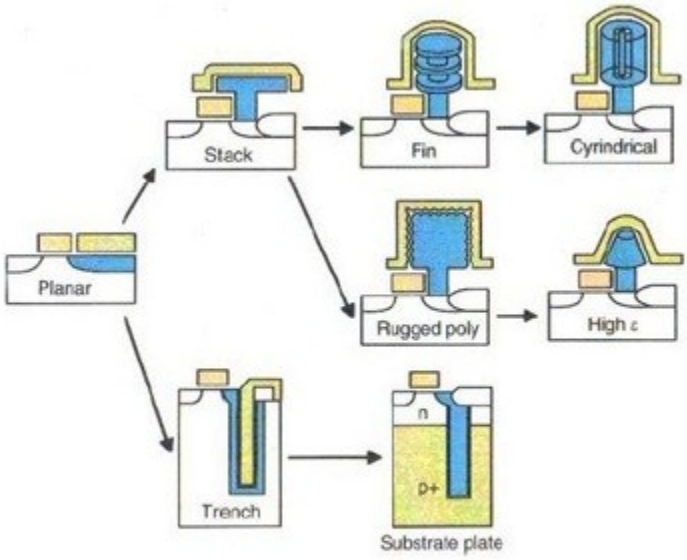
The Monolithic 3D Advantage

The expected and well studied effect of monolithic 3D is the reduction of average wire length. At MonolithIC 3D™ Inc., we have develop a simulation tool –IntSim v2.0 which provides top level simulations for 2D and 3D implementation options. In most cases, for every device folding the average wire length and total silicon area will be reduced by about 50%.

An additional advantage of monolithic 3D could be achieved by placing the embedded memory in dedicated strata. Memory-only strata could be processed in a

flow optimized for memory such as a DRAM flow, to allow the much better density offered by DRAM.

SoCs built with monolithic 3D could be constructed with trench capacitor eDRAM as the first stratum or stack capacitor eDRAM on the upper most stratum. Additional variation that could leverage monolithic 3D would be dual-port eDRAM. This could be done using two strata of transistors so each port may use its own transistors providing two transistors for each capacitor. This could enable user accessibility that is not impacted by refresh accessibility.



DRAM memory cell evolution

Alternatively a more advanced form of eDRAM – Floating Body DRAM (FB-DRAM) could be used. FB-DRAMs use the transistor own body as the charge holder instead of the dedicated capacitor. This form of DRAM had been suggested to save area and simplify the fabrication process. It is very appealing for monolithic 3D as multiple layers of DRAM could be stacked vertically without the bulky capacitors. Yet the FB-DRAM has yet to become an acceptable option, due to the small charge stored and the requirement for rapid refresh. The concept of dual-port could be applied to support rapid refresh with no interference with the user's use of the memory.

Additional advantage of a multi-stratum monolithic 3D SoC is the ability to have a mix of technologies while being efficient in device processing. So for applications that require a decent amount of non volatile memory, a device stratum could be dedicated to Flash memory which utilizes a fabrication flow quite different from a logic flow.

The Monolithic 3D options

MonolithIC 3D Inc. offers two flows for monolithic 3D.

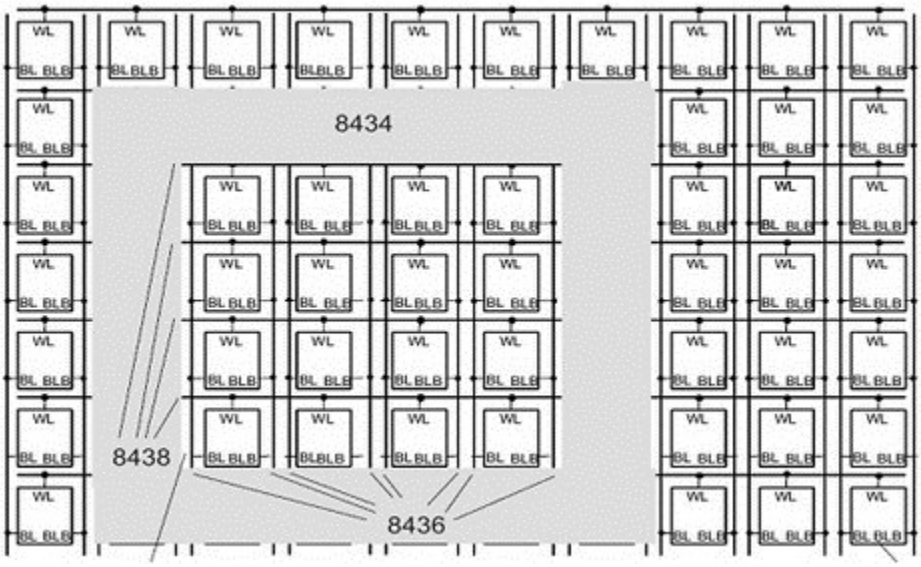
- **Path 1 to Monolithic 3D**: Construct recessed channel transistors in single crystal silicon, common in DRAM manufacturing, above copper interconnects at <400C.
- **Path 2 to Monolithic 3D**: Employ any state-of-the-art replacement gate transistor, along with repeating layouts and a novel alignment scheme, to obtain a high density of vertical connections. The advantage of this technique is its use of state-of-the-art transistor technologies.

Memories being a repetitive structure would work well with Path 2, while DRAMs being the proponent of RCAT transistors would also work well with Path 1.

In short the embedded memory of a 3D SoC could effectively utilize both flows for monolithic 3D fabrication.

The Continuous Array

An additional advantage of monolithic 3D SoCs, and quite a non-obvious one, is the concept we call 'Continuous Array'. The following drawing illustrates the idea:



The drawing illustrates a stratum dedicated to a specific continuous memory array of bit cells. The idea is to process a reticle size continuous terrain of bit cells from which specific memories will be constructed. The memory peripherals could be constructed on the upper or lower strata. The continuous terrain would be customized to the specific SoC need by etching 'borders' around the desired memory structure as required and than connecting to the peripheral logic as required.

This concept provides significant reduction of the NRE and mask cost with benefits for low to medium production volumes.

Part 5: 3D-Repair: Yield recovery for high-density chips

Chapter 17 – Can Yield Increase with 3D Stacking?

by Ze'ev Wurman, the Chief Software Architect of MonolithIC 3D Inc.

When the subject of vertical stacking of active layers is discussed, the question of yield comes up frequently. We all know that chips have defects – after all, that's the main reason why Xilinx chose to offer their large 28nm FPGAs as **stacked dies on an interposer** instead of simply making a larger chip. But when one stacks one aggressive litho die on top another – or worse yet, four or six on top of each other – surely the aggregate yield of this expensive stack must plummet, right?

Turns out that such simplistic approach does not have to be right. In fact, we will see that a clever use of monolithic stacking allows us to **increase** the yield, and reliably manufacture much bigger devices than previously possible.

The basic idea behind yield improvement in monolithic 3D is the concept of repair. We are familiar with this concept from big memory arrays, where we create spare rows or columns, and switch them in as needed using some form of programming to replace faulty memory elements. This works for memory arrays because they are designed to have uniform access time across the whole array, and replacing one column by another that is physically located elsewhere makes no functional difference. In logic terrain, however, this is effectively impossible. Many logic paths are finely tuned and have little slack. Replacing a faulty element in such path with another, which may be far away from the location of the original element, is bound to fail because of the additional delay that is introduced.

This picture changes with monolithic 3D design. We can design our logic on N layers, and we can then place an additional (N+1) layer on top of the stack, dedicated to the repair of the layers below. One example of such architecture is schematically depicted below.

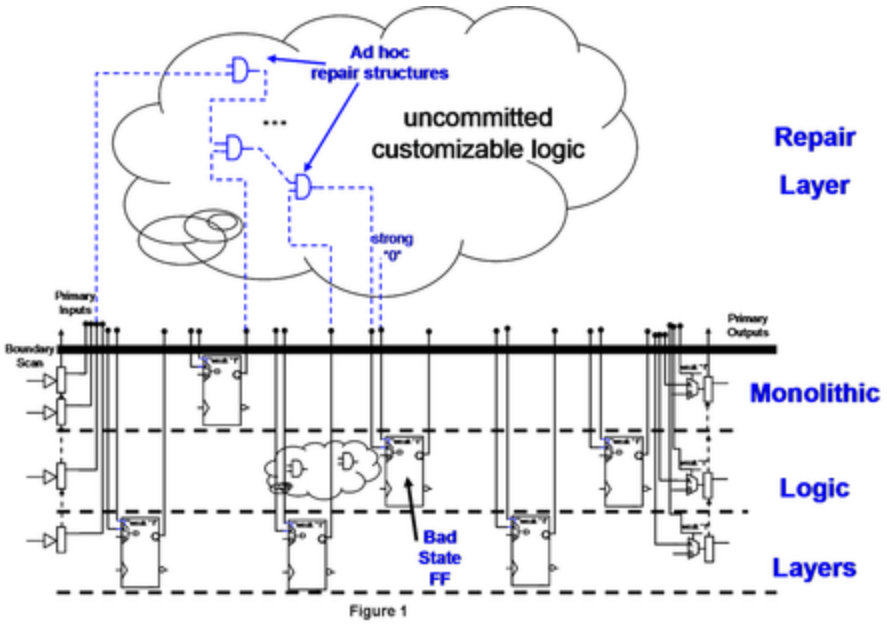


Figure 1

With the dense vertical connectivity that monolithic stacking offers, we can bring the output of every flop to the top repair layer, and we can multiplex an additional input to every flop from that layer. The repair layer itself consists of uncommitted logic that can be programmed late in the manufacturing process through, for example, direct-write e-beam machine. Using this technique we can create large number of ad-hoc repair structures as needed, based on the diagnosed faults in the lower N layers. The beauty of this architecture is that one can create the repair structure right above the fault, and with each monolithic layer being perhaps only 1-2 microns thick, the replacement delay will be similar to the delay of the original logic. One can even make the repair layer of ultra fast (and power hungry) logic to provide additional timing margin, as only a tiny fraction of that repair layer is ever used. A true “drop in” replacement!

We have described here one repair architecture, but others are possible. The key point to remember is that with multi-layer stacking we can afford to have silicon dedicated to repair right above where any potential logic fault can occur.

Before I finish this post, let me touch on another intriguing possibility. Thirty years ago Gene Amdahl gave up on his dream of wafer-scale integration, when he realized that the yields needed for a wafer-scale device will not be attainable for perhaps another 100 years. Yet monolithic 3D stacking with a repair layer brings Amdahl’s dream within our reach. After all, with a repair capability on a logic cone-by-cone basis, nothing stops us from achieving close to 100% yield even at the level of a full wafer.

Chapter 18 – Monolithic 3D IC Could Increase Circuit Integration by 1,000x

by Zvi Or-Bach, the President and CEO of MonolithIC 3D Inc.

Since the invention of the Integrated Circuit by Jack Kilby and Bob Noyce, we been pursuing Moore's Law by doubling device integration every two years. Higher integration has been the key ingredient to end product cost reduction and performance improvement. It has been well documented and demonstrated in the literature that integrating functions that were spread on a PC board onto a single chip could provide order of magnitude reduction of operating power and similar benefits to cost and performance.

The following information was presented recently by Chris Malachowsky, nVidia's Founder and senior VP of research: [Reference]

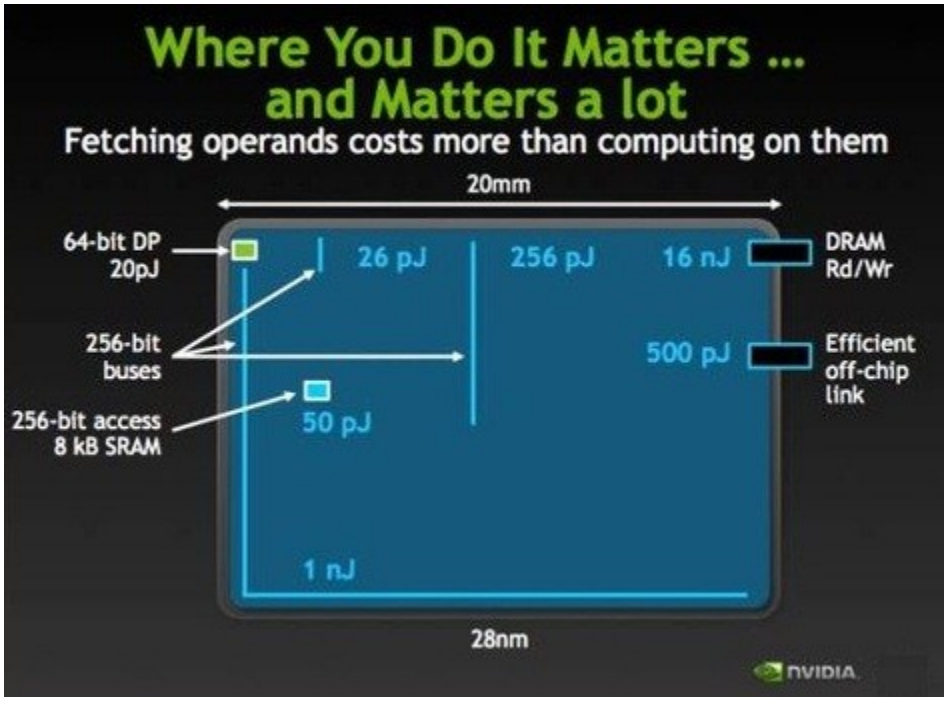


Figure 1: Energy estimates for different operations in nVIDIA's 28nm chips.

Simply stated: **"loading the data from off chips takes >> 100x the energy".** And clearly energy is today the limiting factor of future electronic systems and computing.

Figure 2: More estimates from nVIDIA.

So why are we not integrating more?

The main limit to integration is yield. A secondary limitation is reticle size (~20x30 sq. mm). The semiconductor industry has an amazing skill to continuously improve device yield with scaling. At every new process node, yield gets improved so the new node with double the complexity gets yield similar to the previous node for about the same die size.

It is expected that a die of 10x10 sq. mm will have better than 50% yield. But as yield get reduced exponentially with die size, only in extreme cases, we see designs that are full reticle size and those tend to have very low yield.

MonolithIC 3D Inc. has innovated practical technologies to process multiple tiers of circuits with vertical connectivity comparable with horizontal connectivity. The technology utilizes very thin layers (<100nm) of mono-crystalline silicon, so each tier with its interconnect layers would add about 1 micron to the chip, allowing super high integration if the yield limit could be overcome.

Overcoming yield of non-repeating circuits (such as memory) is considered a hard problem. Trilogy System had attempted to do so with systematic application of *"Triple Modular Redundancy". **Every logic gate and every flip-flop were triplicated with binary two-out-of-three voting at each flip-flop.*** Trilogy systems was known as one of the largest financial failures in Silicon Valley before the burst of internet/dotcom bubble in 2001. Aparently Trilogy's failure had a lasting effect and it seems that for over two decades no other attempts towards Wafer Scale Integration were made.

We believe that a new approach and new technology, alligned with many times larger market and far higher value for integration merits the development of super scale integration. The follwing provides an illustration of MonolithIC 3D Inc.'s 3D super scale integration scheme:

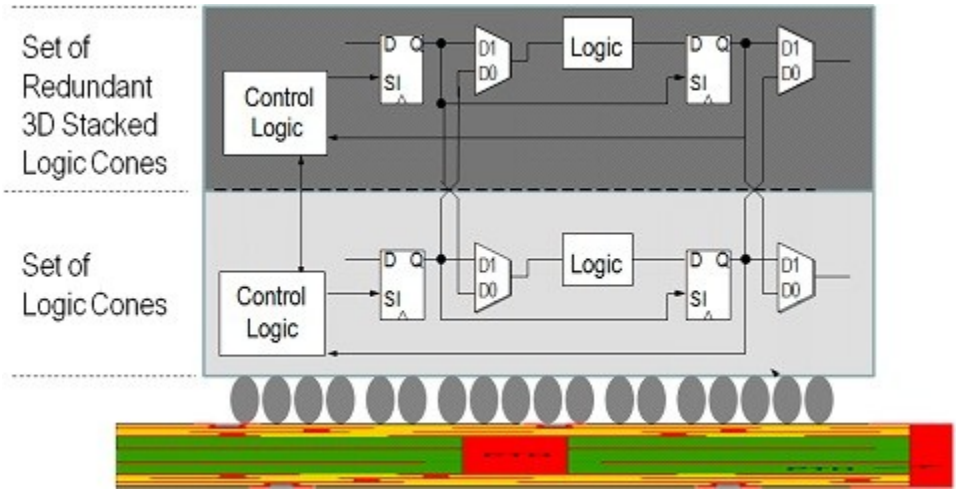


Figure 3: MonolithIC 3D Inc.'s super-scale integration scheme.

There are three primary ideas:

- Swap at logic cone granularity.
- Redundant logic cone/block directly above, so no performance penalty.
- Negligible design effort, since redundant layer is exact copy.

The new concept leverages two important technology breakthroughs.

The first is the Scan Chain technology that enables circuit test where faults are identified at the logic cone level. The second is the 3D IC which enable replacement of defective logic cone by the same logic cone ~1 micron above.

Accordingly, by just building the same circuit twice one on top of the other with minimal overhead, every fault could be repaired by the replacement logic cone above. Such repair should have negligible power penalty and minimal cost penalty whenever the base circuit yield is about 50%. There should be almost no extra design cost and many additional benefits can be obtained (which we will discuss later)

So the immediate question would be how far can we go with such an approach ?

A simple back-of-the-envelope calculation should start with the number of Flip-Flops in a modern design. In today's designs we would expect more than 1 million F/F (logic cones). So, if we expect one defect, then the device with redundancy layer would

work unless the same cone is faulty on both layers which probability wise would be one in a million!

Clearly we have removed yield as a constraint to super-scale integration. We could even integrate 1,000 such devices!!!

Chapter 19 – Repair in 3D Stack: The Path to 100% Yield with No Chip Size Limits

by Ze'ev Wurman, the Chief Software Architect of MonolithIC 3D Inc.

Last month we **described** how monolithic 3D layers enable super large scale integration using redundancy layers. In that approach each logic layer is duplicated and adding a regular symmetric vertical connectivity between these layers allows swapping in a replacement logic cone for each faulty logic cone on the main layer. This permits close to 100% yield for arbitrary-sized chips, up to wafer-scale size, at the cost of dedicating a repair layer for each logic layer.

Today I will describe an alternative method that addresses the repair of multiple stacked layers of logic by a single stacked repair layer. Like before, this method offers close to 100% yield and enables super large scale integration devices up to a wafer size.

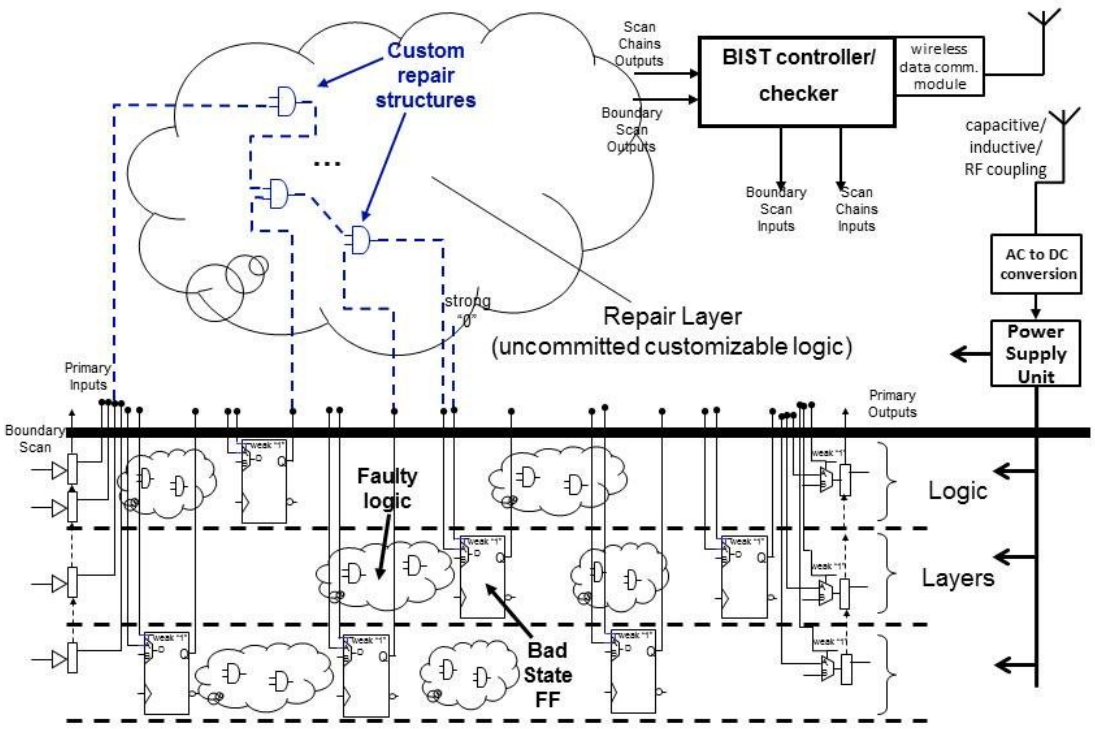


Figure 1

The principle behind this method is quite simple and relies on the dense vertical connectivity offered by monolithic 3D technology, as well as the inexpensive availability of direct-write e-beam lithography. (The word “inexpensive” in the previous sentence is not a typo.) Figure 1 provides an overall view of this approach. It consists of N stacked layers of logic, with an N+1 repair layer on top. The logic is conventionally scan-based, but uses a special flip flop that has an additional multiplexer in front of the FF input, as described in figure 2. By default, this mux is steering the regular logic input into the flop through a weak pullup at its control. The additional mux input and its control are vertically routed to the repair layer, which also has the output of the flop available. We should observe that having three vertical connections for every flop, and at multiple logic layers, can be easily achieved with monolithic 3D but is not feasible for most designs with TSVs – they are simply too big.

The repair flow is pretty straightforward. The wafer is completed through its N layers of logic and half-way into the top repair layer, up to its metal 3 or 4. At that point the BIST controller and the contactless data communication and power harvesting modules should be completed (more on them later). Scan testing is performed using this contactless powering and probing and the on-board BIST controller, and any failing logic cones are identified. External CAD software then synthesizes the failing logic cones in the repair layer using the flop outputs available there, and places them in a close proximity to the original x,y location of the logic – except that on the repair layer -- to maintain timing similar to the original one. Output of the synthesized replacement logic is fed to the appropriate flop mux input, and the mux control is tied to logic 0 to steer this replacement logic to the flop. This is depicted by the blue repair structures in figure 1.

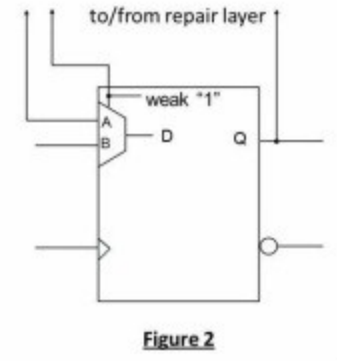


Figure 2

The repair layer can be made of a gate-array-like terrain, or of some other metal-programmable type of terrain. An important element is that this terrain needs to be routable (and programmable, if need be) using a small number of metal segments on a single metal layer or, ideally, only metal vias on a single via layer. Similar segmented routing fabrics are routinely used by FPGA companies and by structured ASIC manufacturers such as ChipX and eASIC. With such segmented metal fabric, the e-beam machine needs to spend minimal time – a matter of minutes per wafer – to implement the repair structures on the repair layer. After that step, the fabrication of the wafer continues to completion, except that now each chip-site/die has a customized repair structure in place.

A few points are worth noting. First, since the vertical distance added by each layer is on the order of a micron, the distances (and timing) are essentially preserved

when using the repair layer. Further, the transistors on the repair layer can be made somewhat faster (and more power hungry) than the logic layer transistors, as only a handful of the repair transistors will ever be actually used; hence their impact on the overall power dissipation is miniscule. Second, in a typical manufacturing flow one expects faults on the order of one per square centimeter or less and, even with multiple stacked layers, a single repair layer contains plenty of transistors available to effect repairs of a few logic cones at this fault density. Third, it should be noted that this particular approach does not address the case when the fault is in the flop itself. Flops typically occupy only a fraction of the silicon area and the impact of this restriction on yield is minimal.

One may wonder how realistic the contactless approach to testing wafers is. Just last month ST Microelectronics **announced** first commercial wafer-level contactless testing.

At the 2011 ISSCC, Keio University (Yokohama, Japan) researchers **announced** inductive harvesting of 6 watts of energy with a 5x5 mm square chip. A year before that they **demonstrated** a 6 Gb/s wireless transfer rate per pin with a 300x300 micron antenna size, and in 2009 a group from the same university **demonstrated** contactless probing that can perform DC measurement. (The links require IEEE subscription). Clearly, contactless testing is coming just in time to assist with the testing of large 3D chips.

Part 6: 3D – DRAM: Monolithic 3D DRAM

Chapter 20 – Introducing our Monolithic 3D DRAM technology

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

A few months back, we received an invitation to speak at the AVS 3D workshop in San Jose. We felt it would be a good opportunity to discuss our monolithic 3D DRAM technology, so we accepted the organizers' kind invitation. The workshop happened last week.Overall, it was a fun event to speak at. I was impressed with the questions asked by people in the audience, and also their enthusiasm for the subject. The organizers had planned the event very well and the room was packed to it's capacity (with ~150 people). You can find details of this workshop **here**. Other speakers at the workshop were Sesh Ramaswami from Applied Materials, Valeriy Sukharev from Mentor Graphics and Robert Rhoades from Entrepix.

Let's now talk about the technology itself. As many of you know, the industry has been aggressively pursuing monolithic 3D approaches for NAND flash memory wherein litho steps are shared among multiple memory layers (see my old blog-post titled "**Looking beyond lithography**"). Toshiba has their version called Bit Cost Scalable (BiCS) Technology, while Samsung, Hynix and Intel/Micron have their own approaches. Fig. 1 summarizes these schemes. The common thing with all these approaches is the use of polysilicon for making NAND flash transistors

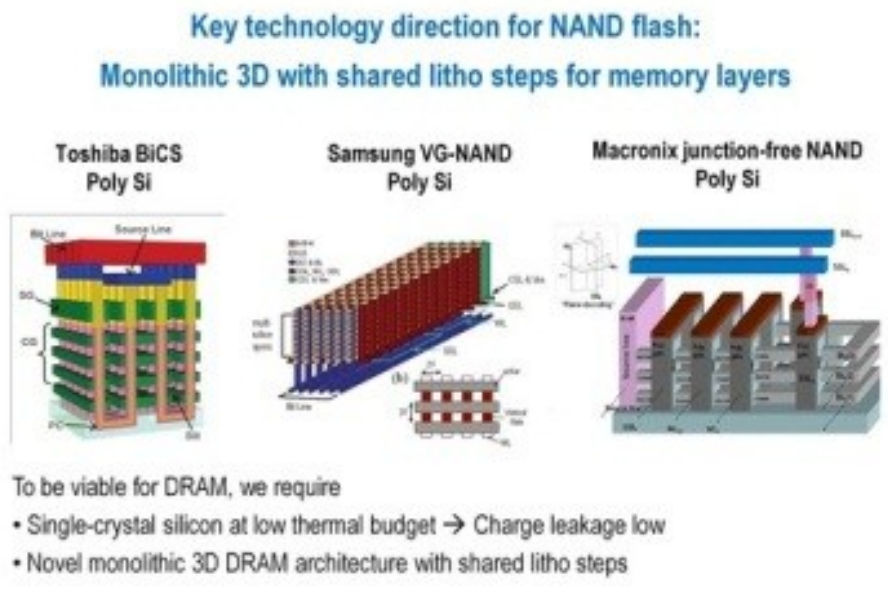


Fig. 1

While the NAND flash memory industry has gone after this technology direction in a big way, the DRAM industry has not explored it at all. One of the key reasons for this is the fact that NAND flash can live with polysilicon transistors but DRAM cannot. Charge stored in the DRAM would just leak out if polysilicon is used. The transistor's performance would not be high enough either :-(DRAM doesn't use the large amount of ECC and redundancy NAND flash does, and this makes the use of polysilicon even more difficult. In our company, we looked at this as an opportunity. If we could invent a way to apply single crystal silicon to these 3D memories, we could potentially come up some disruptive 3D DRAM technologies! The key innovations we needed were:

- Stacked single crystal silicon layers produced with low thermal budget
- A novel monolithic 3D DRAM architecture with shared litho steps

It turns out both these problems can be solved. Ion-cut, the technology used for manufacturing all SOI wafers nowadays, can provide stacked single-crystal silicon at low thermal budgets. It's shown in Fig. 2. Ion-cut involves bonding a hydrogen implanted top layer wafer onto a bottom layer wafer, cleaving the bonded stack at it's hydrogen implant plane and later polishing the surface. This process was invented in the early 1990s at CEA -LETI and has been in production since the late 1990s. As Fig. 3 and Fig. 4 show, our novel 3D DRAM architecture uses double-gated floating body RAM, a technology that has been developed by several manufacturers for 2D DRAM including Hynix and Intel. Essentially, the DRAM is capacitorless, with charge stored in the body of a transistor. Capacitorless DRAM is quite helpful for stacking multiple memory layers with shared litho steps, since it avoids the bulky stacked capacitor (we do have approaches to do monolithic 3D DRAM with shared litho steps even with capacitors, but the amount of capacitance is not that high). As Fig. 4 indicates, ***our novel 3D DRAM architecture innovatively combines three well-studied and mature technologies: monolithic 3D with shared litho steps, stacked single crystal silicon with ion-cut and double-gated floating body RAM.***

Ion-Cut:
Stacked single crystal Si at low thermal budget

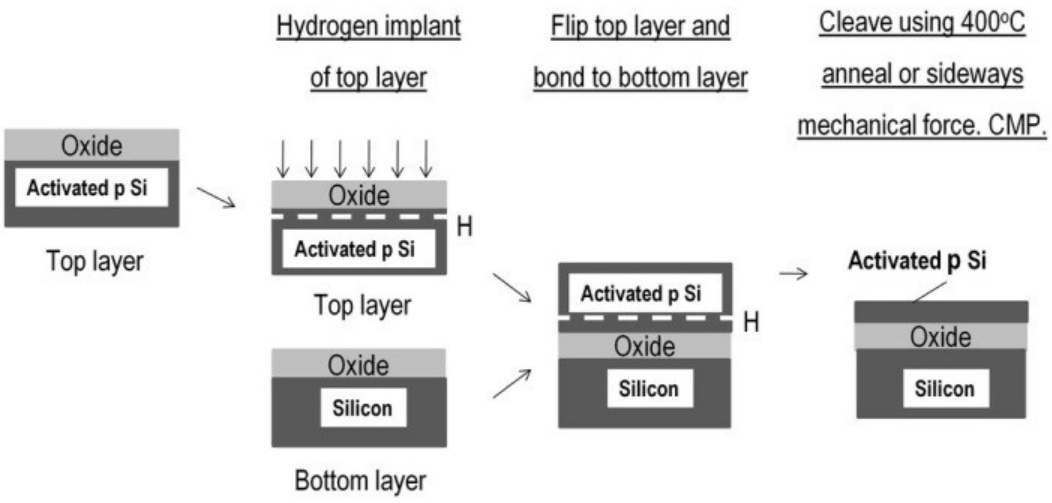
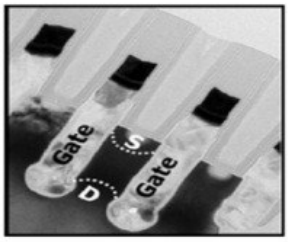


Fig. 2

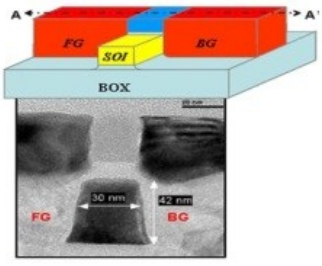
Double-gated floating body RAM:
Well-studied in silicon for 2D-DRAM

Hynix + Innovative Silicon
VLSI 2010



- 0.5V, 55nm channel length
- 900ms retention
- Bipolar mode

Intel
IEDM 2006



- 2V, 85nm channel length
- 10ms retention
- MOSFET mode

Fig. 3

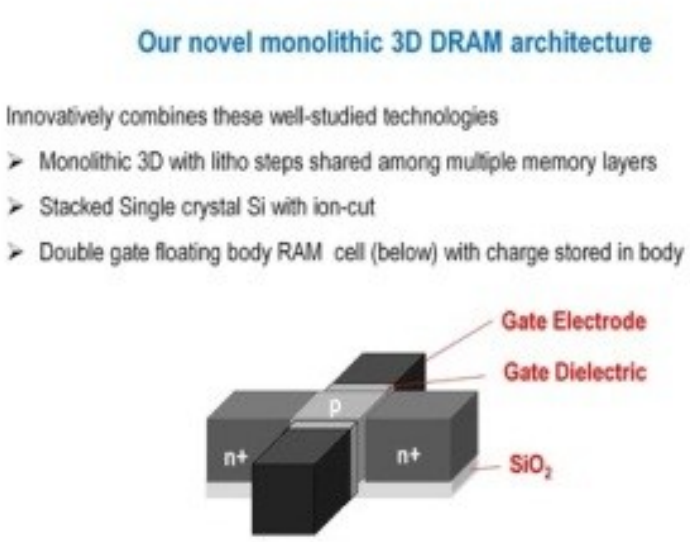


Fig. 4

The steps required for this monolithic 3D DRAM are summarized in Fig. 5-Fig. 10.

- Step 1: Ion-cut is used to transfer a p-type single crystal silicon layer atop the peripheral circuits of the DRAM as depicted in Fig. 5. Notice how the peripheral circuits are placed under the memory array... this improves the array efficiency and allows smaller-size blocks that offer high performance.
- Step 2: Using litho and implant, n+ doped regions are formed as shown in Fig. 6.
- Step 3: Using steps similar to Step 1 and Step 2, a silicon-silicon dioxide multilayer sandwich is formed as described in Fig. 7. A high temperature anneal is conducted to activate dopands in multiple layers of memory at the same time.
- Step 4: Using the same litho and etch step, multiple layers of memory are defined as shown in Fig. 8.
- Step 5: Gates are formed for multiple levels of memory at the same time as described in Fig. 9. Since the source and drain regions are defined in Step 2 and Step 3 and gates are formed separately in Step 5, the process is not self-aligned, which will produce a density penalty of around 20%.
- Step 6: Using another shared litho step, bit-line contacts are formed to multiple levels of memory. Bit-lines are then made. Contacts to multiple levels of memory are defined with shared litho steps using a process described in [Tanaka, et al., Symposium on VLSI Technology, 2007]. Fig. 10 reveals the structure after this

step. Using carefully chosen biases to bit-lines (BLs), word-lines (WLs) and source-lines (SLs), each bit in the memory array can be uniquely addressed.

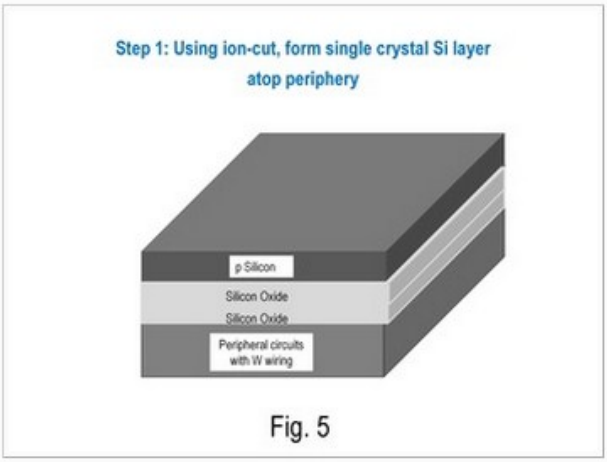


Fig. 5

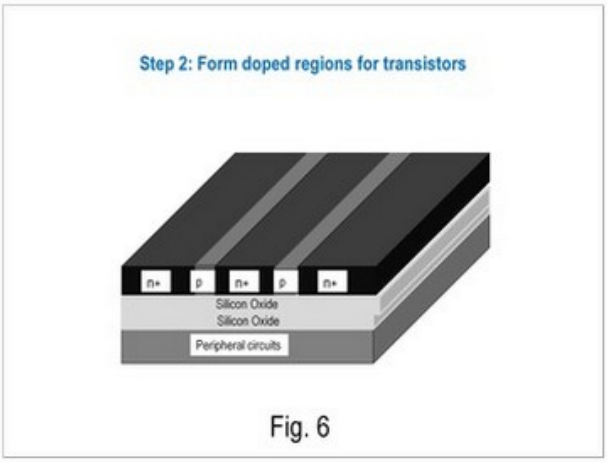


Fig. 6

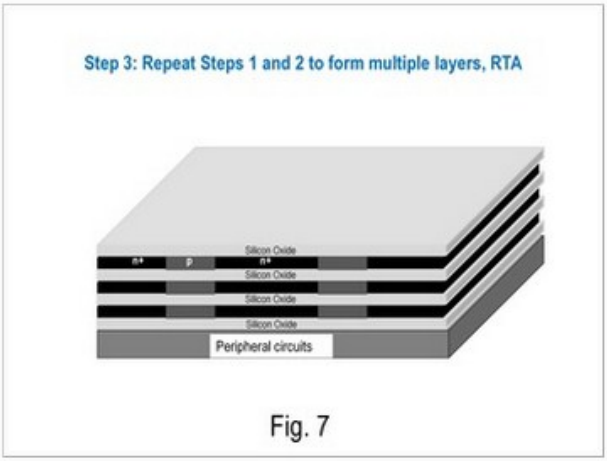


Fig. 7

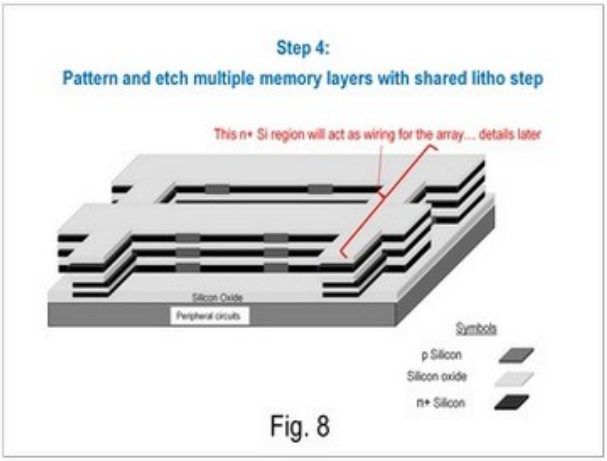


Fig. 8

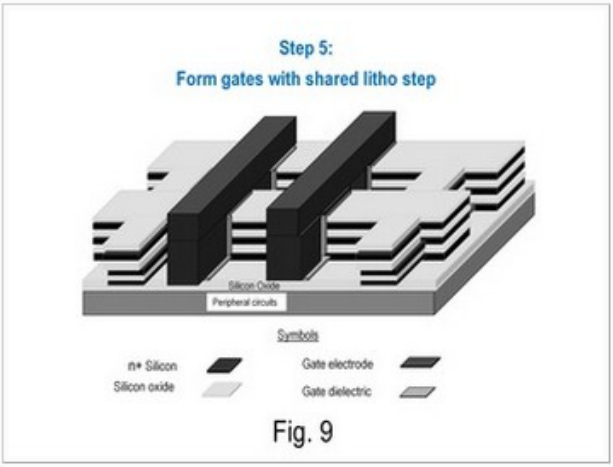


Fig. 9

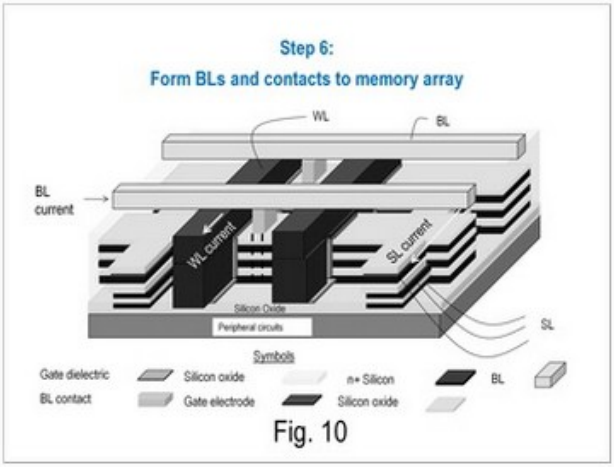


Fig. 10

Fig. 11 shows approximate density estimations for this technology. You'll notice the monolithic 3D DRAM offers more than 3x the density of standard capacitor-based

DRAM without an increase in the number of critical litho steps :-) For a commodity industry such as DRAM, that's a huge gain!

Density estimation

	Conventional stacked capacitor DRAM	Monolithic 3D DRAM with 4 memory layers
Cell size	$6F^2$	Since non self-aligned, $7.2F^2$
Density	x	3.3x
Number of litho steps	26 (with 3 stacked cap. masks)	~26 (3 extra masks for memory layers, but no stacked cap. masks)

3.3x improvement in density vs. standard DRAM, but similar number of critical litho steps!!!

Fig. 11

The other key implications of this technology are shown in Fig. 12 and Fig. 13. You'll see that one can get multiple generations of cost per bit improvement without necessarily upgrading the litho tool. For example, a company can do 22nm 2D, then go to 22nm 3D with 2 device layers after two years and then go to 22nm 3D with 4 device layers after another couple of years. So, you'll be able to use the same litho technology for 6+ years and still get cost per bit improvement! Since tools depreciate in value quite significantly every two years, it is a key win. With conventional 2D scaling, one would need to move to a new and costly litho technology every two years. New litho tools such as EUV ones are projected to cost around $100M... one can delay this :-)

Fig. 13 shows companies can avoid some of the difficulties of standard DRAM scaling with this monolithic 3D approach. Please see my previous post titled "**The most cut-throat portion of the semiconductor industry**" to learn more about the difficulties with standard DRAM scaling.

- One of the biggest challenges to DRAM today is the need for continuous upgrades to litho tools every few years. The next big thing in litho, EUV, has been delayed by many-many years. It was supposed to be in production in 2007, but people now say it's too late for 2015! (see Fig. 13) In the absence of EUV,

companies have moved to costly double patterning technologies, and in fact, are within a year of going to quad patterning (for NAND flash). The risk of next-generation litho can be avoided by using monolithic 3D and sticking with the same litho tools for more years.

- DRAM stacked capacitors require aspect ratios of >150:1 and dielectric constants of around 70 in a few years. You'll see projections from the Intl. Technology Roadmap for Semiconductors (ITRS) in Fig. 13. To put these numbers in perspective, 20:1 aspect ratios are considered challenging in most parts of the industry and dielectric constants of around 70 require exotic new high-k materials; well-known high-k dielectrics such as hafnium oxide, aluminum oxide and zirconium oxide will not suffice. The ITRS puts a big portion of the stacked capacitor roadmap in red, which means "*no-known-solution*". If a company moves to monolithic 3D DRAM, it can potentially avoid these challenges.

- The DRAM industry's roadmap requires a major overhaul of it's cell transistors every generation or two. This challenging problem can potentially be avoided by moving to monolithic 3D DRAMs as well :-) If you stick with the same feature size and just add additional device layers every generation, you may not need to upgrade the transistors for that.

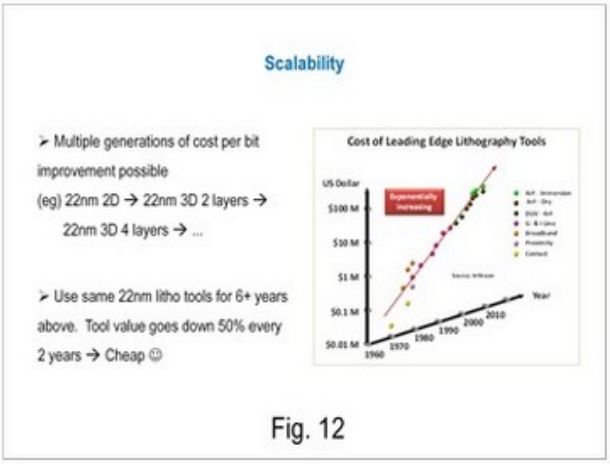


Fig. 12

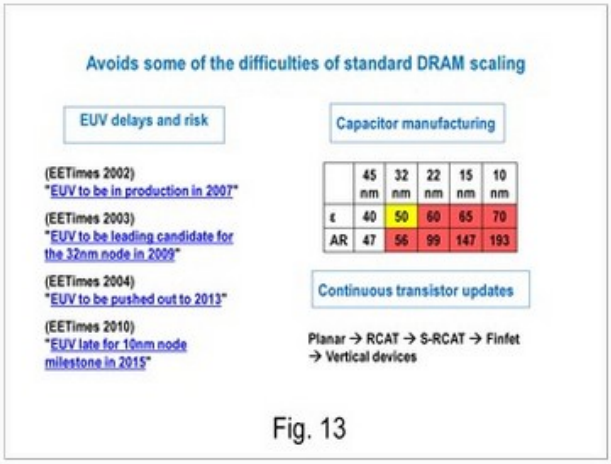


Fig. 13

Like any other technology, this technology has risks as well. One risk is the floating body RAM technology. It hasn't moved to production for 2D-DRAMs yet and is known to have issues with refresh times, reliability and scalability to smaller feature sizes. These challenges will require engineering work to overcome... Furthermore, for the monolithic 3D DRAMs to scale for many generations (for more than 4 device layers), the ion-cut cost needs to reduce significantly to <$50. This is possible since it is an implant, bond and cleave process. In fact, several companies in the cost-sensitive solar

industry, such as SiGen and Twin Creeks Technologies, are using ion-cut nowadays and have figured out creative ways of getting the cost down.

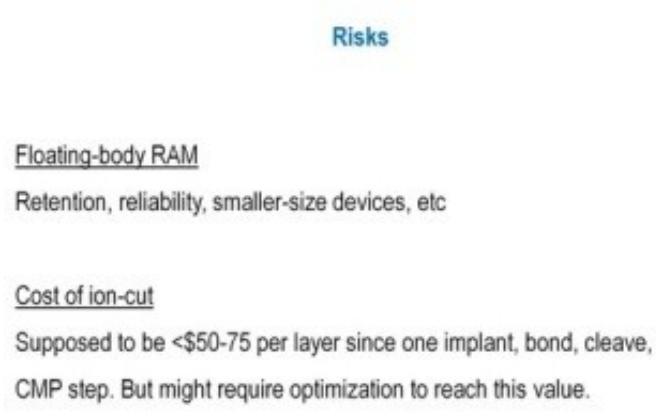


Fig. 14

To summarize:

I just showed you an approach to increase DRAM density by 3x or more without increasing the number of critical litho steps. Thismonolithic 3D DRAM technology can provide several years of continuous cost per bit reduction and reduces the burden we put on next-generation lithography. It also tackles challenges with the stacked capacitor and cell transistor that are inherent to 2D-DRAM scaling. There has been almost no prior work on monolithic 3D with shared litho steps for DRAM, and we've got some pretty fundamental patents allowed by the patent office for this technology. Exciting, isn't it?

To get more details of this technology, please see my presentation at the AVS workshop at the following **link**.

Part 7: 3D – RRAM: Monolithic 3D RRAM

Chapter 21 – Introducing our Monolithic 3D Resistive Memory Architecture

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

Over the past decade, we've seen a slew of rewritable (RW) memory devices: Phase change memory, resistive RAM and MRAM, just to name a few. What is sorely needed is an architecture that allows these RW devices to be built into chips that compete with NAND flash memory. We'll describe MonolithIC 3D Inc.'s solution to this problem here.

Its amazing how many rewritable memory startups Silicon Valley has today... check this list out!

- Ovonyx: Phase-change memory (PCM) startup. Has licensed to Samsung, Micron, Intel, others
- Unity Semiconductor: Resistive RAM (RRAM) startup. Has a partnership with Micron
- Adesto technologies: Resistive RAM startup. Funded by Applied Materials, among others.
- Crossbar: Resistive RAM startup. Funded by Kleiner Perkins.
- 4DS: Resistive RAM startup. Working with Sematech.
- Qs Semiconductor: Resistive RAM startup. Working on SiC memory.
- Nantero: Nanotube RAM startup.
- Grandis: MRAM startup.
- Crocus: MRAM startup.

You'll notice these startups seem to be mainly developing three types of rewritable memory elements: resistive RAM, phase change memory or MRAM. Resistive RAM involves the use of ionic conduction to produce a change in resistance for the memory element, while phase change memory involves the change in phase of a material from amorphous to crystalline and vice versa. This is depicted in Fig. 1. MRAM is another interesting technology, but is not considered a NAND flash replacement, so we'll not discuss it here.

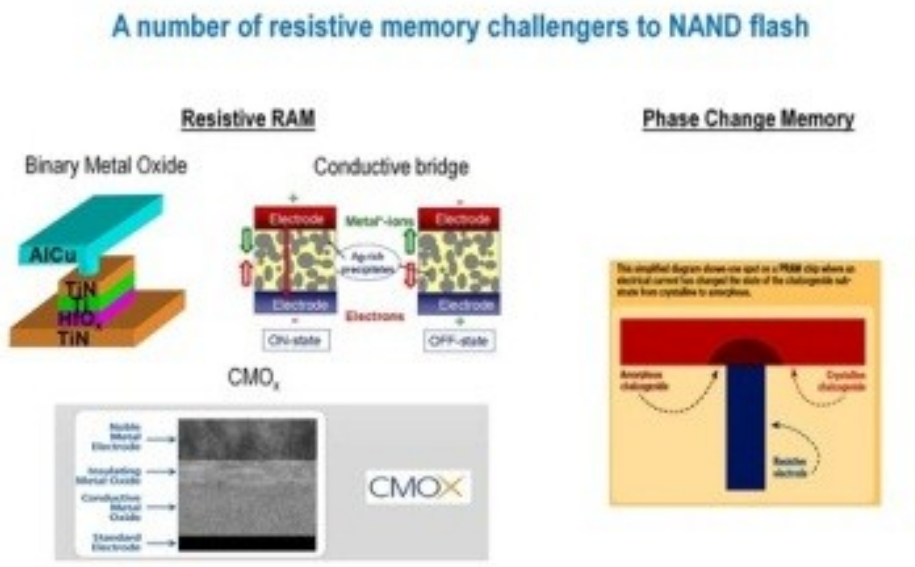


Fig. 1

Fig. 2 shows a plot from Dr. Eli Harari, the recently retired SanDisk CEO, that reveals 3D stacking is necessary for all these rewritable memory technologies to compete with NAND flash. This is because the main driver for NAND flash is cost, and without a 3D architecture, its hard to reach costs of NAND flash which is in volume production today, has 3 bits per memory cell and requires just 4 critical lithography steps.

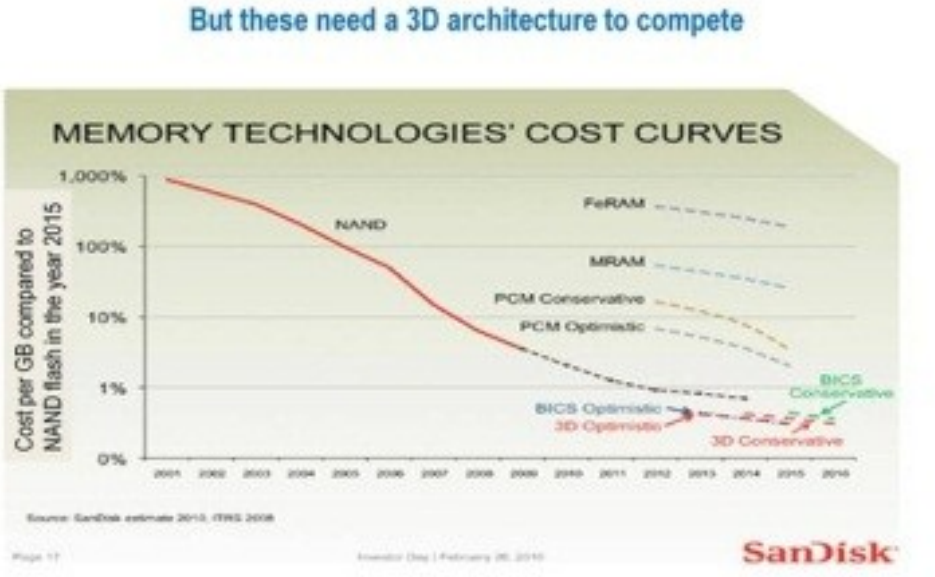


Fig. 2

Let's first summarize the 3D architectures that are being explored today by the above startups, and also by bigger companies such as SanDisk, Samsung, Toshiba, Hynix, IBM and others.

- The most commonly explored architecture is the polysilicon diode selected 3D memory (Fig. 3). This was taken to production by Matrix Semiconductor with antifuse memory in 2003... please see **my blog-post on Matrix** for more details. While the Matrix architecture is mature and uses well-understood polysilicon diode technology, it has some challenges. One key challenge is cost. The Matrix group revealed in their ISSCC 2003 presentation that they needed 2 litho steps per layer of memory, so for 10 memory layers, they need 20 critical litho steps!! With litho costs sky high nowadays, it is hard to compete with NAND flash which has just 4 critical litho steps :-(Furthermore, while the poly diode selector works well with the antifuses Matrix Semiconductor took to production, it is harder for it to work with rewritable memory elements. PCM compatibility is difficult since the poly diode requires more than 700C process temperatures while PCM melts at 620C. RRAM compatibility is impacted by the fact that the pn junction diode conducts current unidirectionally, while many viable RRAM devices require bidirectional current. To tackle these challenges, companies in the industry took different paths.

- Intel demonstrated a test-chip in IEDM 2009 that had multiple layers of PCM in series with Ovonic Threshold Switch (OTS) selectors. Please see Fig. 4 for an illustration. The OTS selector could be constructed at less than 400C and it could conduct either unidirectionally or bidirectionally. While this approach tackles the compatibility issue with RW materials such as PCM and RRAM, it still has challenges with litho cost :-(Furthermore, the OTS selector uses complex materials and is harder to process and optimize than a polysilicon diode.

- Samsung showed their approach to 3D resistive memory at the 2009 VLSI Symposium (see Fig. 5). They used shared litho steps to pattern multiple levels of memory at the same time, thereby tackling the litho cost problem. Their use of a transistor selector also allowed compatibility with common RW materials such as PCM and bipolar RRAM. The key challenge with the Samsung architecture is the sharing of a transistor selector among many RW devices. This caused sneak leakage paths which limited block sizes and degraded array efficiency and cost per bit.

It is clear that the industry would benefit significantly from an improved 3D RW memory architecture. That's precisely what I'm going to describe to you today (see Fig. 6). Our architecture uses shared litho steps to pattern multiple memory levels thereby

keeping cost low. It has single-crystal transistor selectors and being a 1T-1R architecture, doesn't have issues with sneak leakage paths. It is compatible with most common RW materials too.

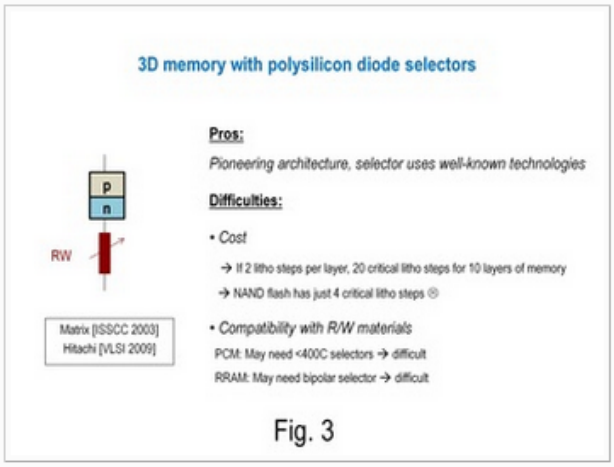


Fig. 3

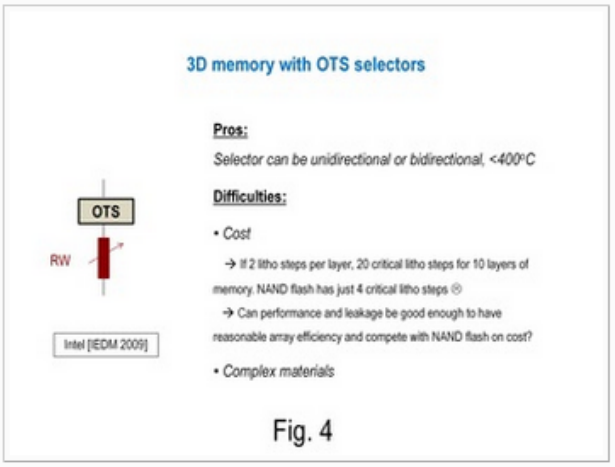


Fig. 4

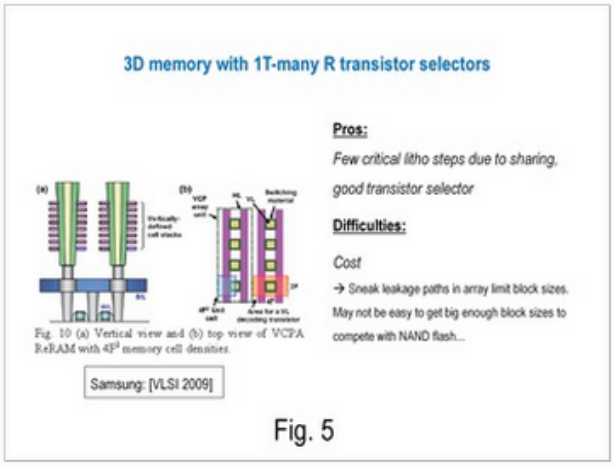


Fig. 5

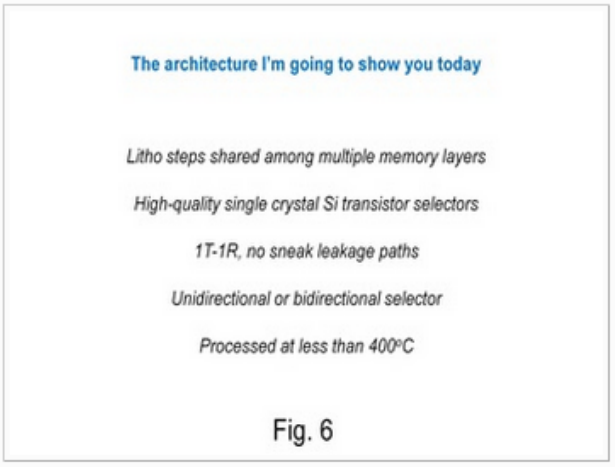


Fig. 6

Ion-cut, the technology used for manufacturing all SOI wafers nowadays, is one of the key ingredients of this new architecture of ours. It can provide stacked single-crystal silicon at low thermal budgets and is shown in Fig. 7. Ion-cut involves bonding a hydrogen implanted top layer wafer onto a bottom layer wafer, cleaving the bonded stack at it's hydrogen implant plane and later polishing the surface. This process was invented in the early 1990s at CEA -LETI and has been in production since the late 1990s for applications such as SOI wafers.

The other key ingredient of our architecture is the use of **junctionless transistors** as resistive memory selectors. These transistors rely on using thin silicon channels that are depleted of charge carriers at voltages close to 0V. Macronix has demonstrated NAND flash memory structures made out of junctionless transistors - please see **this article** for more details. The steps involved in constructing our 3D resistive memory are as follows:

- Step 1: Ion-cut is used to transfer a n+ single crystal silicon layer atop the peripheral circuits of the resistive memory as depicted in Fig. 8. Notice how the peripheral circuits are placed under the memory array... this improves the array efficiency and allows smaller-size blocks that offer high performance. Also, the n+ dopants are pre-activated before layer transfer.
- Step 2: Using steps similar to Step 1, a silicon-silicon dioxide multilayer sandwich is formed as described in Fig. 9.
- Step 3: Using the same litho and etch step, multiple layers of memory are defined as shown in Fig. 10.
- Step 4: Gates are formed for multiple levels of memory at the same time as described in Fig. 11.
- Step 5: Using another shared litho step, a via hole is made to multiple levels of memory. A resistive memory element (such as titanium oxide) is deposited following which an electrode is deposited and CMPed (Fig. 12). WL, SL and BL are acronyms for Word Line, Source Line and Bit Line respectively.
- Step 6: Bit-lines are then made. Contacts to multiple levels of memory are defined with shared litho steps using a process described in [Tanaka, et al., Symposium on VLSI Technology, 2007]. Fig. 13 and Fig. 14 reveal the structure after this step. Notice how each memory cell consists of a junctionless transistor in series with a RW memory device. Using carefully chosen biases to bit-lines (BLs), word-lines (WLs) and source-lines (SLs), each bit in the memory array can be uniquely addressed.

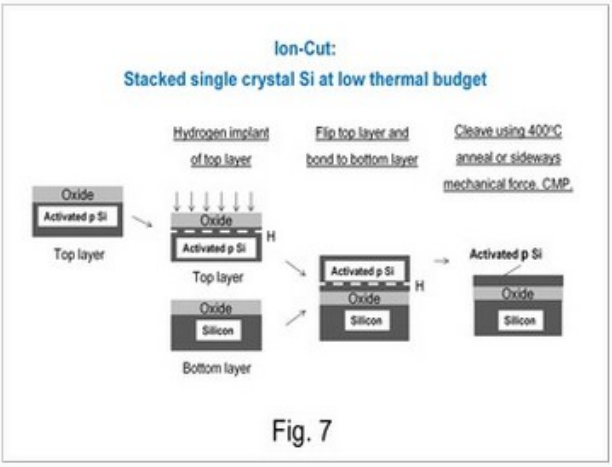


Fig. 7

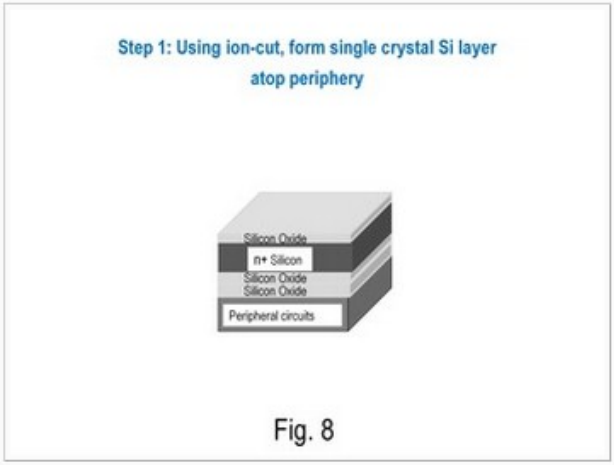


Fig. 8

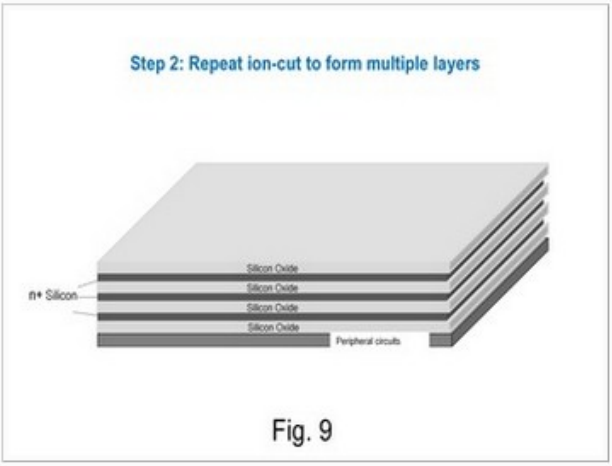


Fig. 9

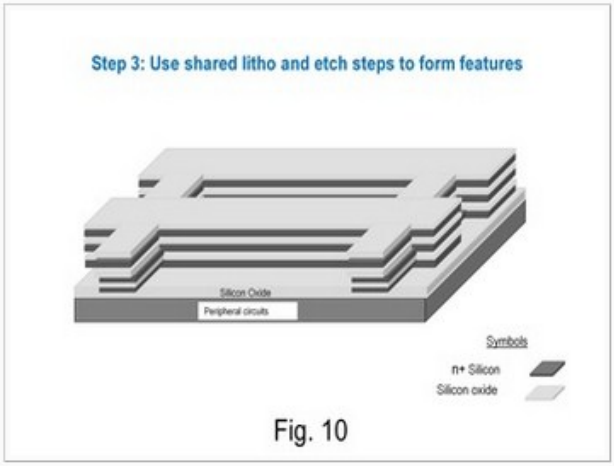


Fig. 10

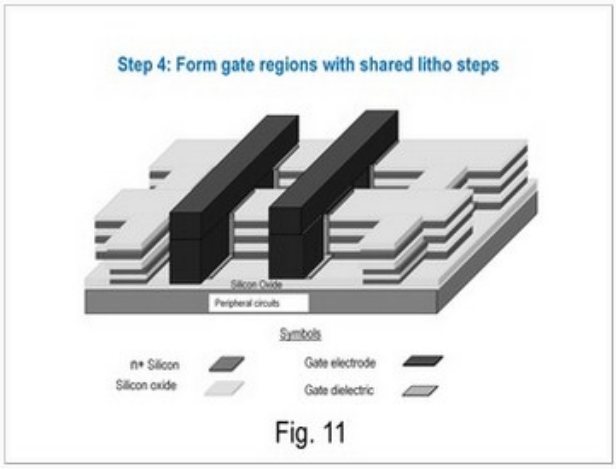


Fig. 11

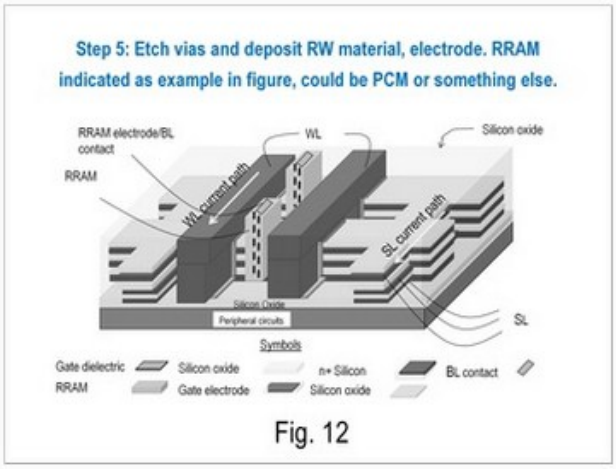


Fig. 12

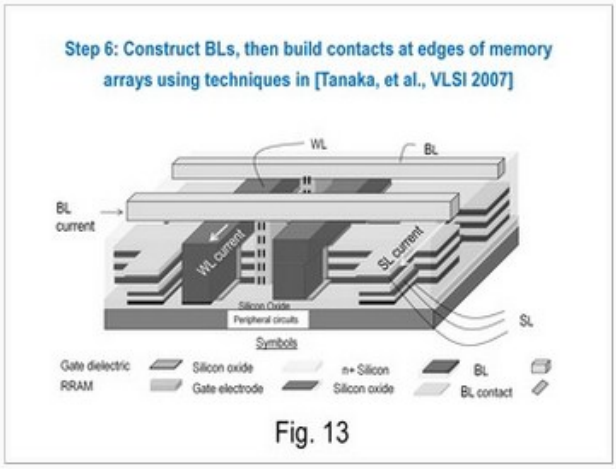


Fig. 13

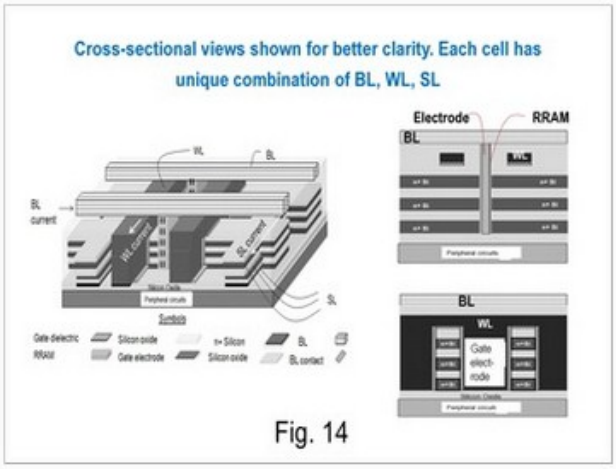


Fig. 14

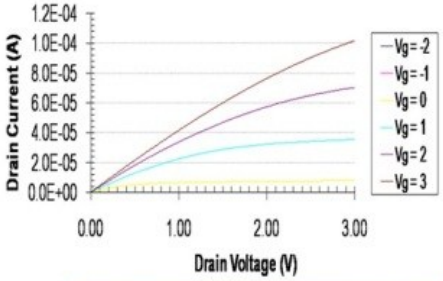
Fig. 15 shows simulation results from **Paul Lim** for these junctionless devices. It reveals that at the 15nm node, these junctionless transistor selectors can have very small leakage currents (<0.1pA) and can still drive power hungry RW materials. Fig. 16 shows array bias schemes with these selectors... you'll notice leakage currents for unselected cells are negligible, indicating large array sizes and excellent performance are possible. These are key advantages of single crystal silicon transistor selectors.

I-V curves of select transistors at the 15nm node

Simulation results

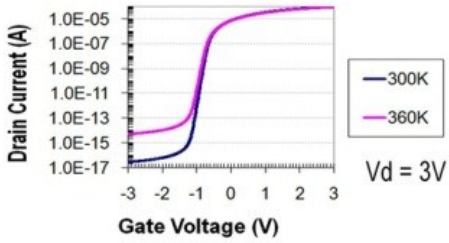
T_{si} = 20nm, T_{box} = 20nm, W_{si} = 15nm, L=90nm, 10 layers of memory

Id-Vd curves



Drive current @ Vg= 3V → 100uA

Id-Vg curves



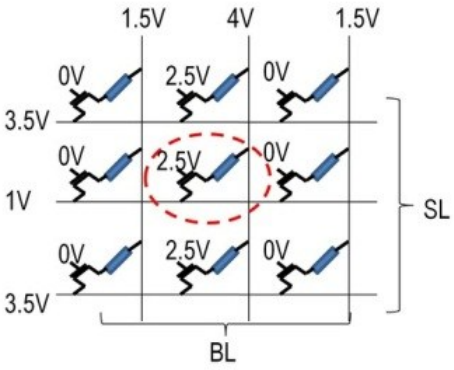
Leakage @ Vg = -1V → pA
Leakage @ Vg = -1.5V → negligible (fA)

I-V curves show value of single-crystal silicon transistors → Selector can drive even power-hungry RW elements like PCM, and still have negligible leakage

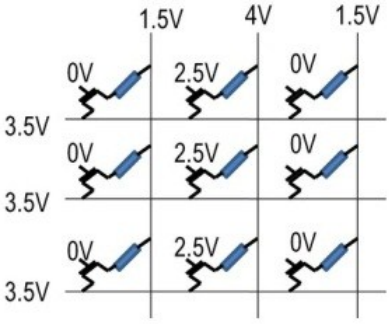
Fig. 15

Array bias schemes

Top layer of 3D memory



Bottom layer of 3D memory



Selected cell: Drive current > 40uA as long as voltage drop across select transistor > 1.3V

Un-selected and half-selected cells: Leakage negligible. Huge array sizes possible

Fig. 16

Fig. 17 shows approximate density estimation for our architecture. The architecture has 0.9F(square) cells and ~5 critical litho steps, while a poly diode selected 3D memory requires 16 critical litho steps for 0.5F(square) cells. To put this in perspective, NAND flash has 2F(square) cells and ~4 critical litho steps. As mentioned previously, the large number of litho steps needed for poly diode selected 3D memory are a key challenge both for cost per bit and fab cap-ex. Our new architecture tackles this issue. Fig. 18 reveals other advantages of our architecture over the poly diode selected 3D memory, such as use of a three terminal selector, high performance due to low leakage and high forward current drive of the selector, possibility for bipolar operation and sub-400C construction of the selector.

Approximate density estimation

	NAND	Poly Diode Selected 3D memory	This architecture
Cell size	$4F^2$	$4F^2$	$18F^2$
Bits per cell	2	1	2
Number of memory levels	1	8	10 for 26:1 aspect ratio
Critical Litho steps per level of memory	4	~2 per level	~5 for 10 levels
Effective density @ 15nm node (memory only)	$2F^2$ and 4 critical litho steps	$0.5F^2$ and 16 critical litho steps	$0.9F^2$ and **5** critical litho steps

Fig. 17

Comparison with poly diode selected 3D memory

	Poly Diode Selected 3D memory	This architecture
Effective density	$0.5F^2$ and 16 litho steps	$0.9F^2$ and 5 litho steps
Selector	Two-terminal poly device	Three-terminal single crystal device
Leakage in array	High	Negligible
Bipolar operation possible?	No, pin diode is unidirectional	Yes, transistor selector
Forward current drive	Low	High

Fig. 18

To summarize,

We've described a novel 3D resistive memory architecture here (see Fig. 19). It can be useful for many types of rewritable memory materials such as phase change memory and resistive RAM due to its sub-400C process temperatures and use of a three-terminal selector. It offers significant density advantages over NAND flash without incurring an increase in litho cost... this is a key differentiator from other types of 3D RW memories and is enabled by the use of shared litho steps. This architecture could produce an effective storage class memory due to the possibility of getting high endurance and high performance at NAND flash-like densities.

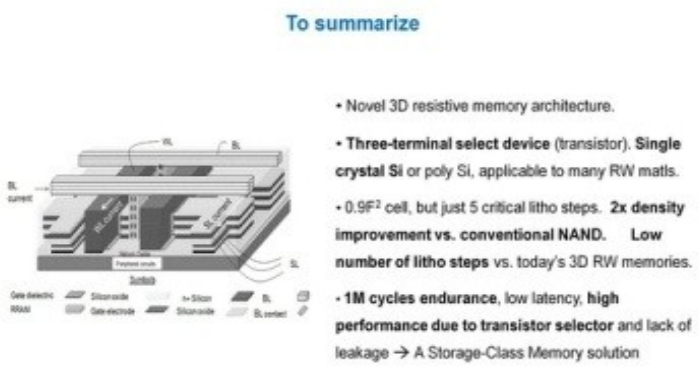


Fig. 19

Part 8: 3D – Flash: Monolithic 3D Flash Memory

Chapter 22 – The Flash Industry's Direction and MonolithIC 3D Inc.'s Solution…

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

Toshiba, Samsung, Hynix and Micron are developing polysilicon-based monolithic 3D flash memories. Today, I'll talk about these and also introduce our company's monocrystalline silicon solution...

You can argue about when NAND flash scaling will end. Some people say two years, others say five. However, there is little argument that a monolithic 3D solution is required when conventional NAND flash scaling ends. Figure 1 shows Monolithic 3D NAND flash memory approaches pursued by Toshiba, Samsung, Hynix and Macronix.

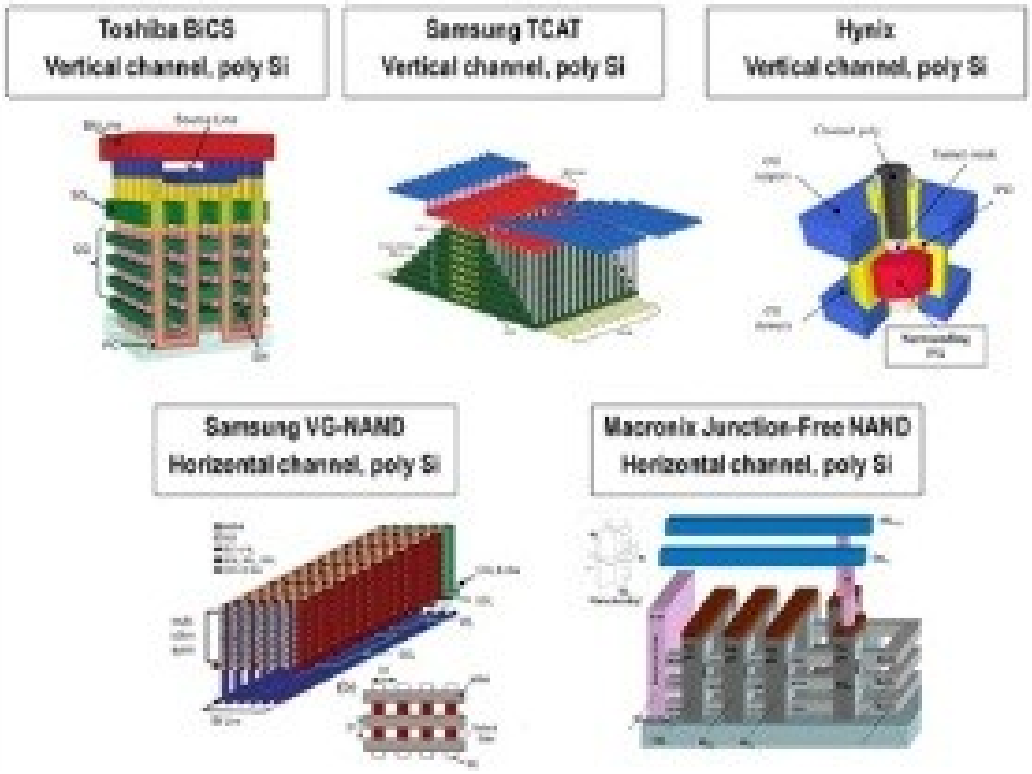


Figure 1: Today's polysilicon-based Monolithic 3D NAND Flash Memories.

The key points to note are:

- Lithography steps for patterning multiple memory layers are shared, which lowers cost.

- Polysilicon is used as the channel material for transistors.
- To be cost-competitive with scaled 2D NAND flash memory, aspect ratios to be etched and filled are often 50:1 or higher. For future generations, aspect ratios need to be increased further!

For more details, please read my old blog post: **Looking beyond lithography**. As you can imagine, polysilicon transistors and high aspect ratios pose significant challenges. Polysilicon has 6x lower mobility, higher sub-threshold slope and significantly larger variability than single crystal silicon, which makes 2 bits/cell and 3 bits/cell difficult. High aspect ratios are problematic to manufacture and yield too.

The questions to ask are therefore: *Can we build 3D NAND flash memories with single crystal silicon instead of polysilicon? In addition, can we use low aspect ratios and still have cost-competitive products?* I will now describe MonolithIC 3D Inc.'s technology, where both these important problems are solved. We were awarded fundamental patent coverage on this technology just a few months back.

Ion-Cut: The Building Block

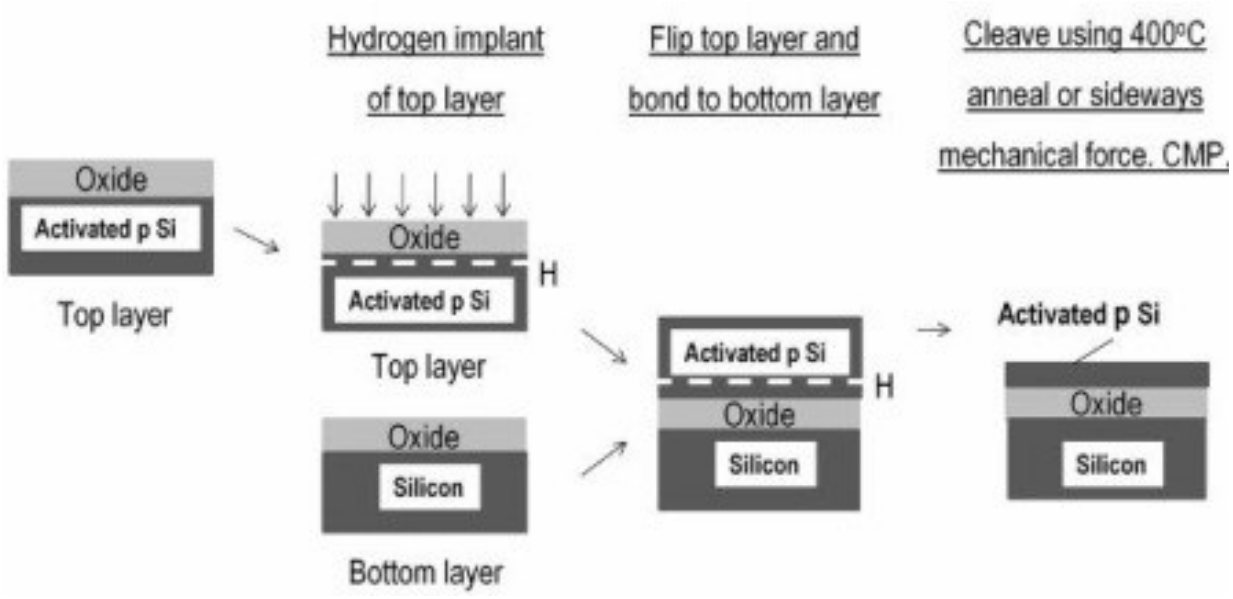


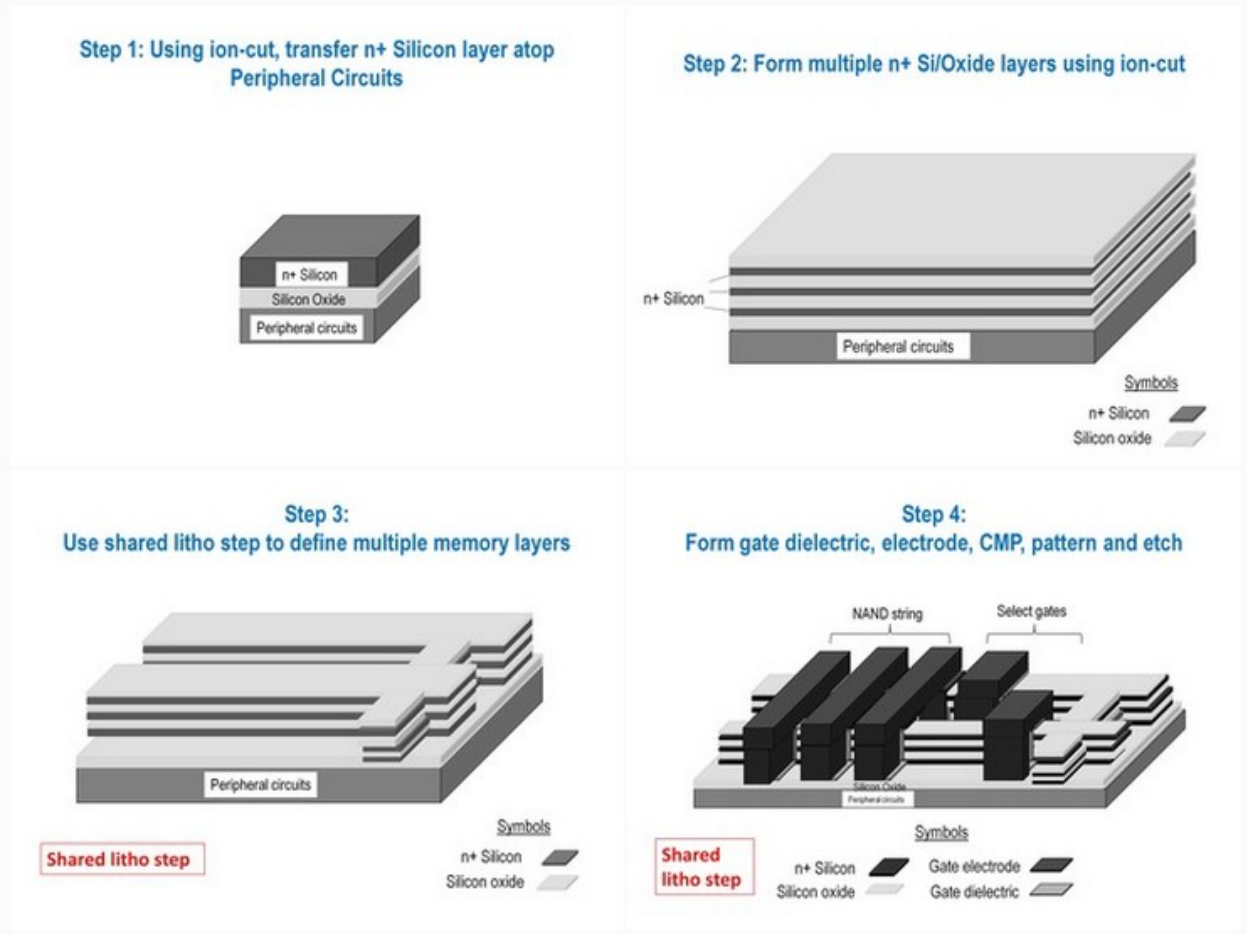
Figure 2: The Ion-Cut process can provide stacked single crystal silicon at low thermal budget.

Ion-cut, the technology used for manufacturing all SOI wafers nowadays, can provide stacked single-crystal silicon at low thermal budgets. Its shown in Figure 2. Ion-cut involves bonding a hydrogen implanted top layer wafer onto a bottom layer wafer, cleaving the bonded stack at its hydrogen implant plane and later polishing the surface.

This process was invented in the early 1990s at CEA -LETI and has been in production since the late 1990s.The process costs around $60 per layer of memory, which is affordable. Ion-cut will become a public-domain technology in 2012, when its basic patent expires. For more cost information on ion-cut, please see my old blog post: **How much does ion-cut cost?**

Process Flow

Figure 3 describes the process flow for constructing our company's monolithic 3D NAND flash memory technology. The key point to note is how lithography steps for patterning multiple memory layers are shared, keeping cost per bit down. The memory cell is a double gate depletion mode single crystal silicon transistor that utilizes charge-trapping as the storage mechanism.



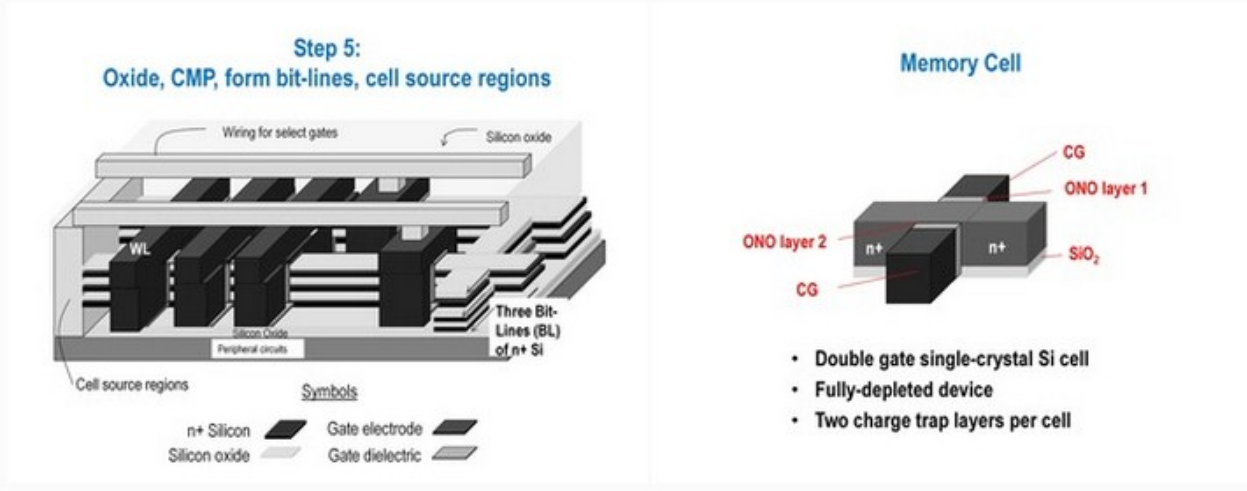


Figure 3: Process flow for constructing our company's Monolithic 3D NAND Flash Memories.

The steps involved in this process are:

- Step 1: Ion-cut is used to transfer a n+ single crystal silicon layer atop the peripheral circuits as depicted in Figure 3. Notice how the peripheral circuits are placed under the memory array... this improves array efficiency. Tungsten may be used for the wiring of the periphery.
- Step 2: Using steps similar to Step 1, a silicon-silicon dioxide multilayer sandwich is formed as described in Figure 3. A high temperature anneal may be conducted (if desired) to reduce defect levels in the layer transferred silicon.
- Step 3: Using the same litho and etch step, multiple layers of memory are defined.
- Step 4: Gate dielectrics and electrodes are formed for multiple levels of memory at the same time.
- Step 5: Cell source regions are formed. Contacts to multiple levels of memory are defined with shared litho steps using a process described in [Tanaka, et al., Symposium on VLSI Technology, 2007]. Figure 3 reveals the structure after this step. Using carefully chosen biases to bit-lines (BLs), word-lines (WLs) and the cell source, bits in the memory array can be accessed.

Implications

140 sq. mm die	Conventional NAND 22nm node	BiCS 32 layers @ 45nm node (around the limit for CD)	MonolithIC 3D Flash 8 layers @ 22nm node
Density	64Gbit (3 bits/cell)	128Gbit (SLC)	256Gbit (2 bits/cell)
Aspect ratio		60:1 → hard to manufacture	16:1

Figure 4: Estimates for density based on data presented at the 2010 VLSI Symposium Short Course.

Figure 4 gives estimates for density and aspect ratio based on data presented at the 2010 VLSI Symposium Short Course. MonolithIC 3D Inc.'s single crystal silicon solution can provide 4x higher density than conventional NAND flash memory at the 22nm node. Aspect ratios are manufacturable, unlike today's poly-based solutions.

Our technology is, of course, applicable to any monolithic 3D NAND flash memory architecture where the transistor's channel is horizontal. For more details, please check out our issued US patent #8,026,521 or contact me by e-mailing deepak@monolithic3d.com.

Part 9: IntSim v2.5

Chapter 23 – IntSim v2.0: An Open-Source Simulator for Monolithic 2D and 3D-ICs

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

Some background on IntSim

I first began work on IntSim during my PhD studies at Georgia Tech almost 5 years back. We folks in Prof. James Meindl's research group had derived compact models for various device and interconnect phenomena. There was opportunity to combine together all these models to get a chip simulator. IntSim v1.0 was the result. **It could simulate 2D-ICs and we described it in ICCAD 2007** in San Jose (I had just got married... my wife was based in San Jose while I was based in Atlanta finishing up my PhD, so I liked attending conferences in the San Jose area!). Over the past 5 years, a number of university researchers and professors have used IntSim v1.0 :-) Some used it to evaluate chip-level performance/power benefits of novel transistor technologies, some used it as an architecture simulator, and others used it to set homework assignments for classes they taught.

Well, I joined NuPGA/MonolithIC 3D Inc., and we were coming up with some great new ways to do Monolithic 3D-ICs. The question we began asking ourselves was: how does going to monolithic 3D impact chip performance, power and die size? We couldn't design an actual monolithic 3D chip since CAD tools for this were still under development. So, I suggested to Zvi Or-Bach, our CEO, that hey, there was this CAD tool I built for 2D-ICs at Georgia Tech, I can extend it to monolithic 3D using 3D wire length distribution models in the literature. Zvi liked the idea, and suggested I go ahead. He also said, "Let's offer it on our website for people to use, let them play with it and have fun simulating monolithic 3D chips too". Thus began the efforts for IntSim v2.0.

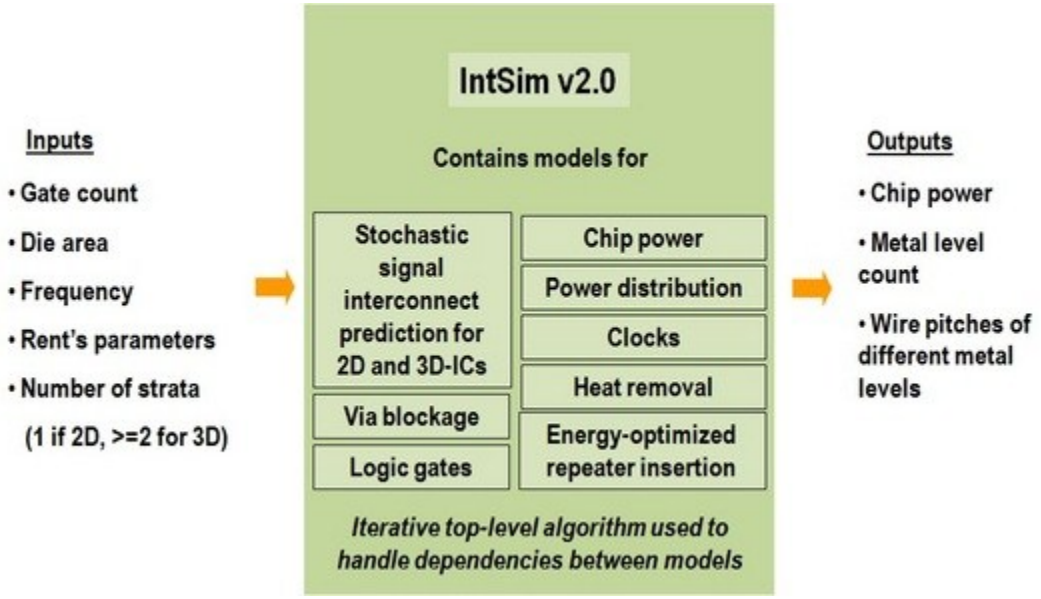
Structure of IntSim v2.0

You can see a diagrammatic representation of IntSim v2.0 below. For a detailed account of models used in IntSim v2.0, please visit **the "IntSim's models" page**. There are a few key improvements compared to IntSim v1.0:

- Support for monolithic 3D-ICs: Signal wire length distributions for monolithic 3D are obtained using Arif Rahman's models (**link to Arif's PhD thesis in MIT**). I extended **Kaveh Shakeri** and **Reza Sarvari**'s models for power distribution to

3D, and developed my own models for 3D heat removal... I'll talk more about the 3D heat removal models once they're published.

- Java and Open-Source: IntSim v1.0 was written in MATLAB, and required a (somewhat costly) MATLAB license :-(IntSim v2.0 is in JAVA, so you can run it as an app on any OS: Windows, Mac OS X or Unix. The tool is very well-documentated, and is Open Source. So, if you feel you'd like to contribute and improve IntSim v2.0, please let us know. We'll give you the source code of IntSim v2.0, you can add in your features... we can then release your features in IntSim v3.0 and list you as a contributor to the tool.



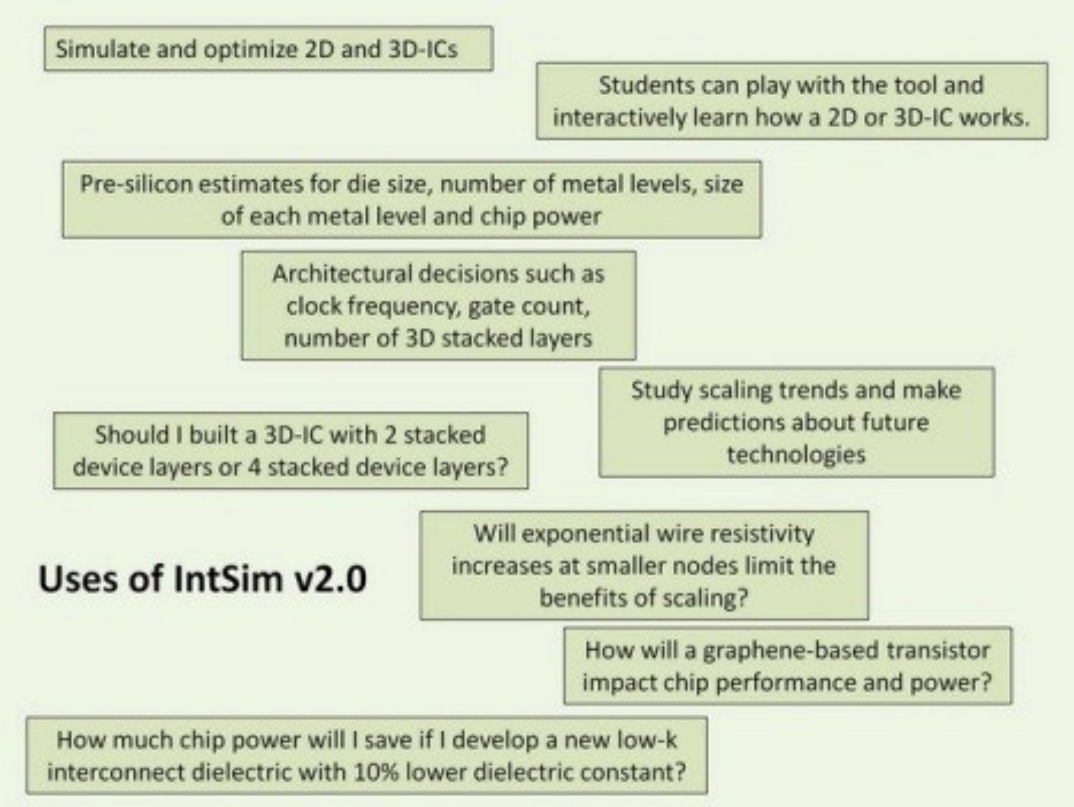
Want to run IntSim v2.0 and check it out?

It's simple! Double-click on the icon below to run a beta version of IntSim v2.0.

Utility

You can use IntSim v2.0 for a number of things: simulate 2D and 3D-ICs, determine scaling trends and get estimates for quantities such as die size, pitches of metal levels in a multilevel interconnect network, chip power and clock frequency prior to design. What excites me the most is that some professors are using IntSim as a fun way for students to learn how a chip works. For example, they set homework

assignments asking students to use IntSim and find out how chip power and the interconnect stack changes as a function of clock frequency - this helps students appreciate and understand why clock frequency increases in the future are not that attractive. The picture below shows some common uses of IntSim.



Comparison with Actual Data from a Commercial Microprocessor and case studies showing use of IntSim v2.0

Please check out our **"comparison with actual data" page** and our **"case studies" page**.

What's next?

We are building a small group in our company to develop open-source CAD tools for 3D-ICs. We'll add more features to IntSim moving forward, and we also plan to develop other 3D open-source CAD tools... It is clear to us that having good CAD tools and simulators will accelerate the industry's transition to Monolithic 3D.

Many thanks to Zvi Or-Bach, MonolithIC 3D Inc.'s CEO, for supporting development of IntSim v2.0 and for various useful inputs. I'd also like to thank Prof. James Meindl of Georgia Tech under whose guidance most of the models in IntSim were developed. Jeff Davis, Ragu Venkatesan, Arif Rahman, Keith Bowman, Kaveh

Shakeri, Reza Sarvari, Azad Naeemi, Ajay Joshi and several others helped with useful discussions while developing IntSim... I'd like to thank them for their help.

Chapter 23 – Introducing IntSIm v2.5

by Deepak Sekar, former Chief Scientist of MonolithIC 3D Inc.

Today, let's check out IntSim v2.5, the latest version of our open-source chip simulator. IntSim v2.5 has a powerful and simple-to-use GUI and helps optimize 2D and 3D chips.

As many of you know, IntSim is an open-source 2D/3D chip simulator that's been developed at Georgia Tech and MonolithIC 3D Inc. Using IntSim, one can optimize various parameters of a 2D/3D chip such as power, die size, number of metal levels, size of metal levels, gate count and clock frequency. The simulator helps study scaling trends and is a fun tool for students to intuitively learn how a chip works.

MonolithIC 3D Inc. is pleased to introduce the next version of the simulator, IntSim v2.5, to you today. We've added a great new GUI in this version of the simulator.

Double-click on the icon below to download and run IntSim v2.5

Here is a summary of features added in this version:

- Significantly improved Input and Output GUIs,
- Store and load technology files, and
- Sweep and optimize parameter values.

Please see the slideshow below for pictures of our GUI. For more details of IntSim, please refer to our **Simulators page**. Here is an **EETimes story** where IntSim was used to study how a tri-gate transistor impacted chip power.

You can watch a video of IntSim v2.5 here.

We had an intern, Parthiv Mohan, over for the summer and he implemented these features. Parthiv is a student at Saratoga High School, and that's him grinning at you from the picture alongside :-) We were quite impressed with Parthiv's work, and if you try using the GUI, you'll know what I mean. We told Parthiv the features we wanted, and he implemented all of them independently in JAVA without requiring too much guidance. Pretty good for a high-school student!

We hope you will have fun using IntSim v2.5. And if you have any questions, please don't hesitate to e mail us at intsim@monolithic3d.com.

