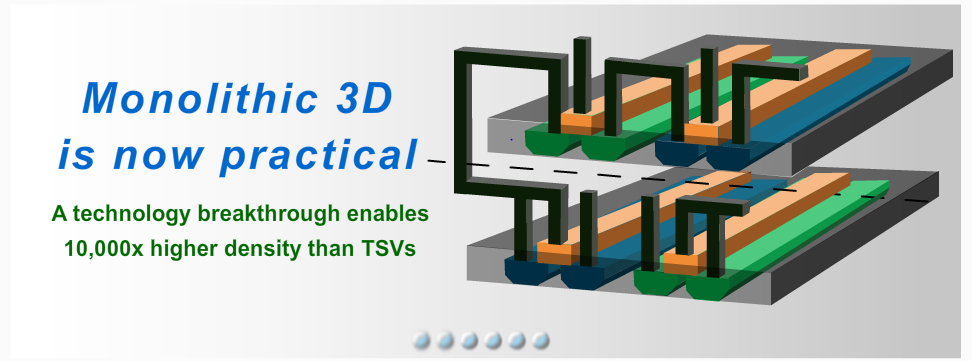
The Monolithic 3D Advantage

Monolithic 3D is far more than just an alternative to 0.7x scaling!!!



**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](http://www.monolithic3d.com/2/post/2012/12/iedm-moores-law-seen-hitting-big-bump-at-14-nm.html)

## [Is the Cost Reduction Associated with Scaling Over?](http://www.monolithic3d.com/2/post/2012/06/is-the-cost-reduction-associated-with-scaling-over.html)

## [Entanglement Squared](http://www.monolithic3d.com/2/post/2011/03/guest-contribution-entanglement-squared-by-zvi-or-bach.html)

## [IEDM 2012 - The Pivotal Point for Monolithic 3D IC](http://www.monolithic3d.com/2/post/2013/01/iedm-2012-the-pivotal-point-for-monolithic-3d-ic.html)

## 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:

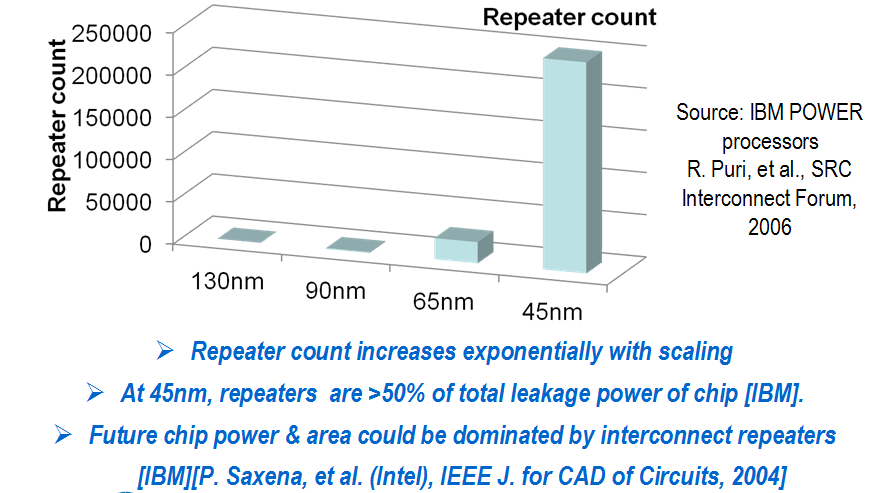
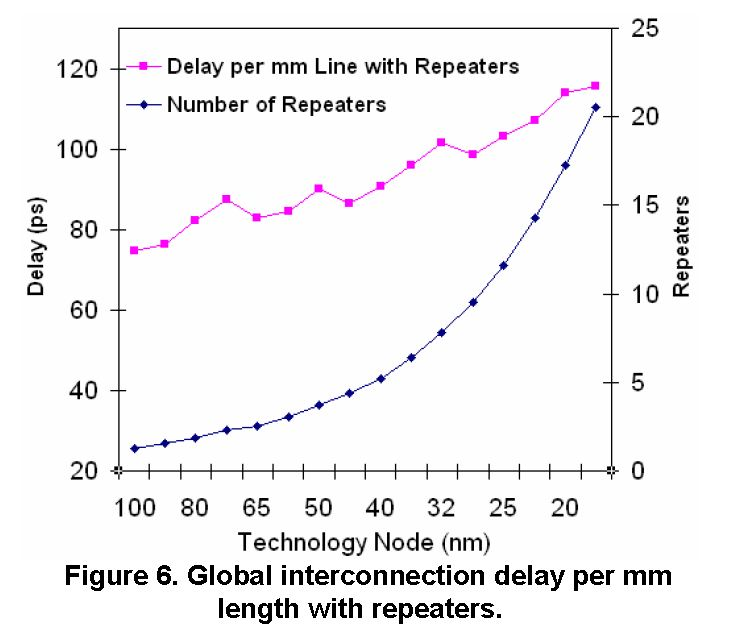
* + Continuing reductions in die size and power
  + Significant advantages for reusing the same fab line and design tools
  + Heterogeneous Integration
  + Processing multiple layers simultaneously, offering multiples of cost improvement
  + Logic redundancy, allowing 100x integration at good yields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

**2. Reduction in die size and power**

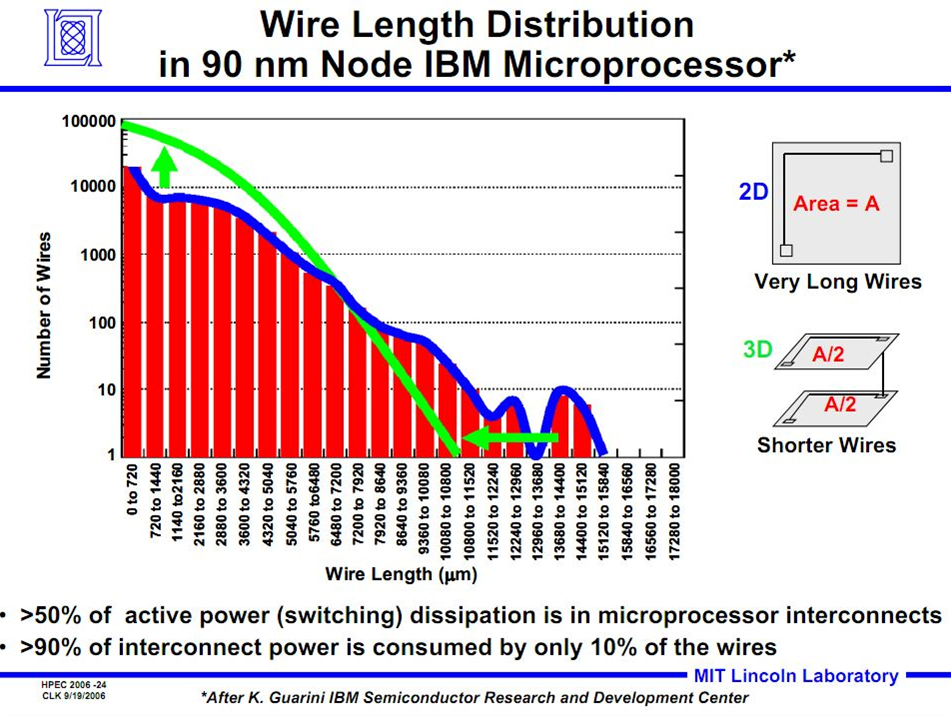
**A. Reduction in die size**

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.

Source: ISQED07 Alam

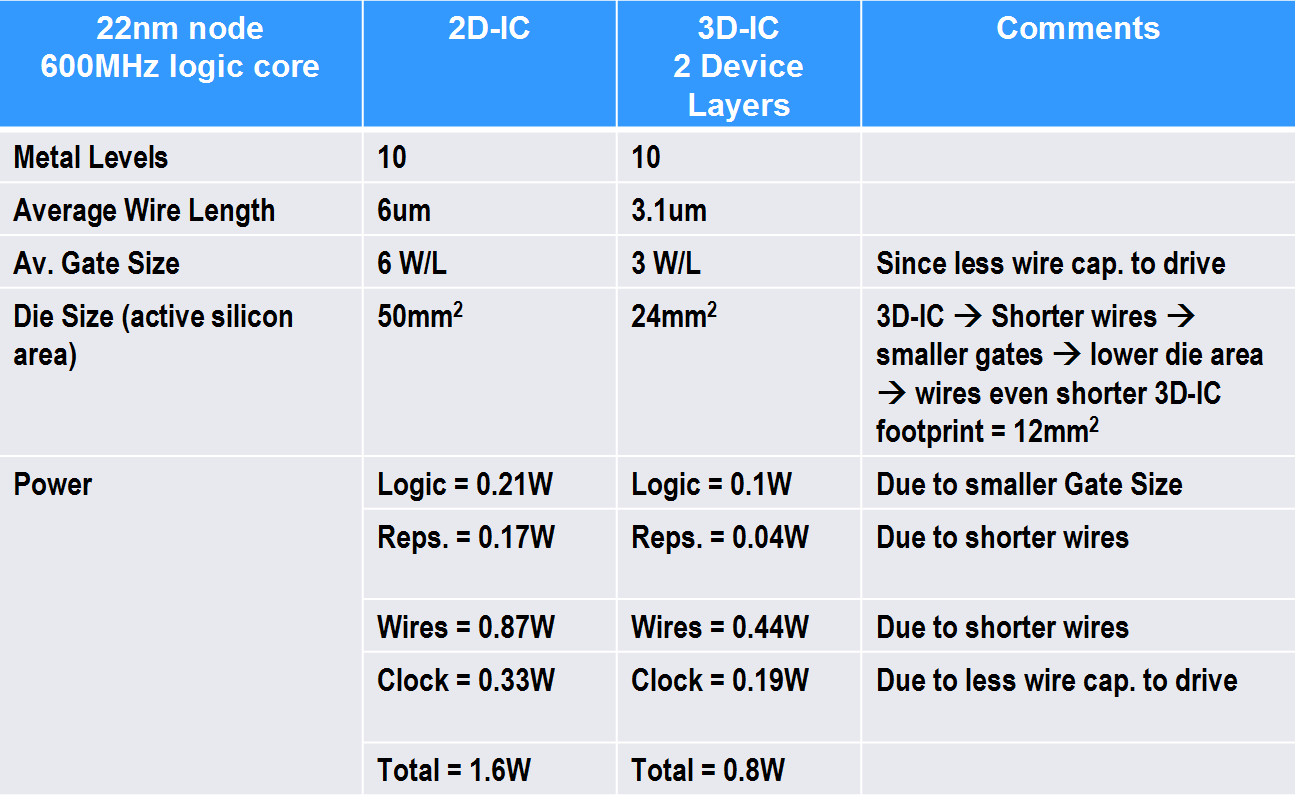


Monolithic 3D enables the folding of a circuit, with the each stratum only about 1µ above or below its neighbor, combined with a very rich vertical connectivity between the strata. The following IBM/MIT slide illustrates the effectiveness of such a folding.



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](http://www.monolithic3d.com/simulators.html) 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).

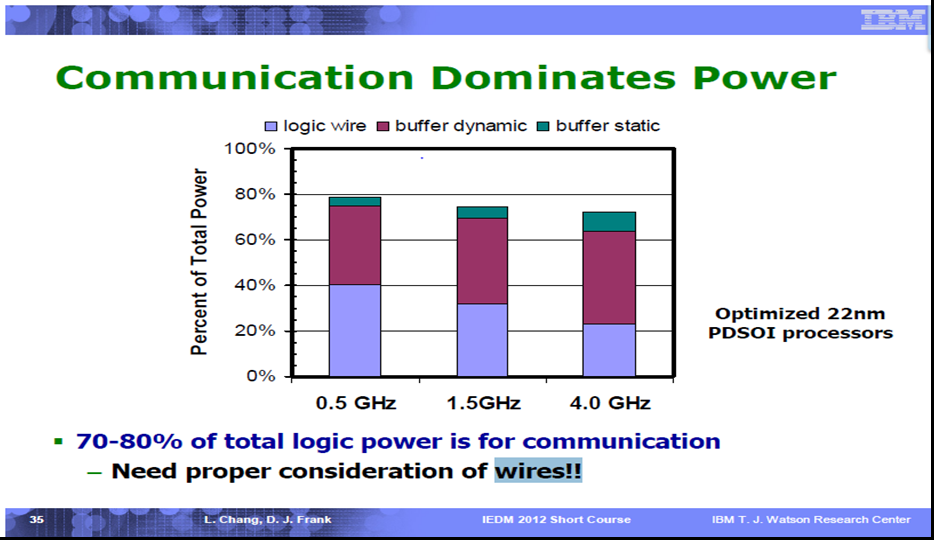


These results are in-line with many other monolithic 3D research results.

**=> 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.



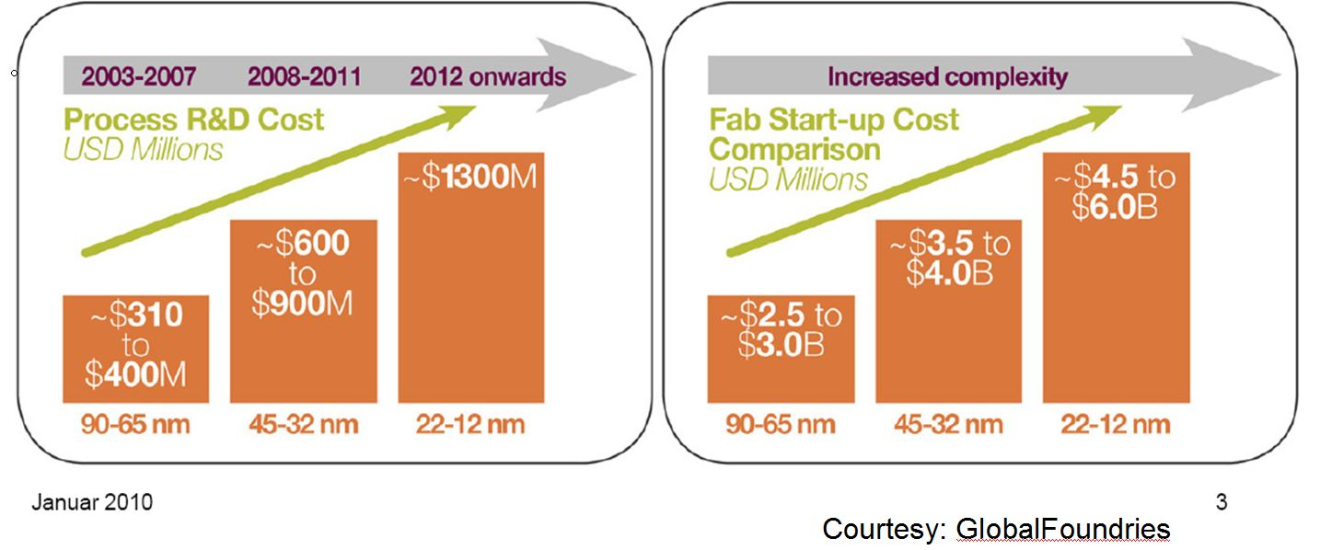
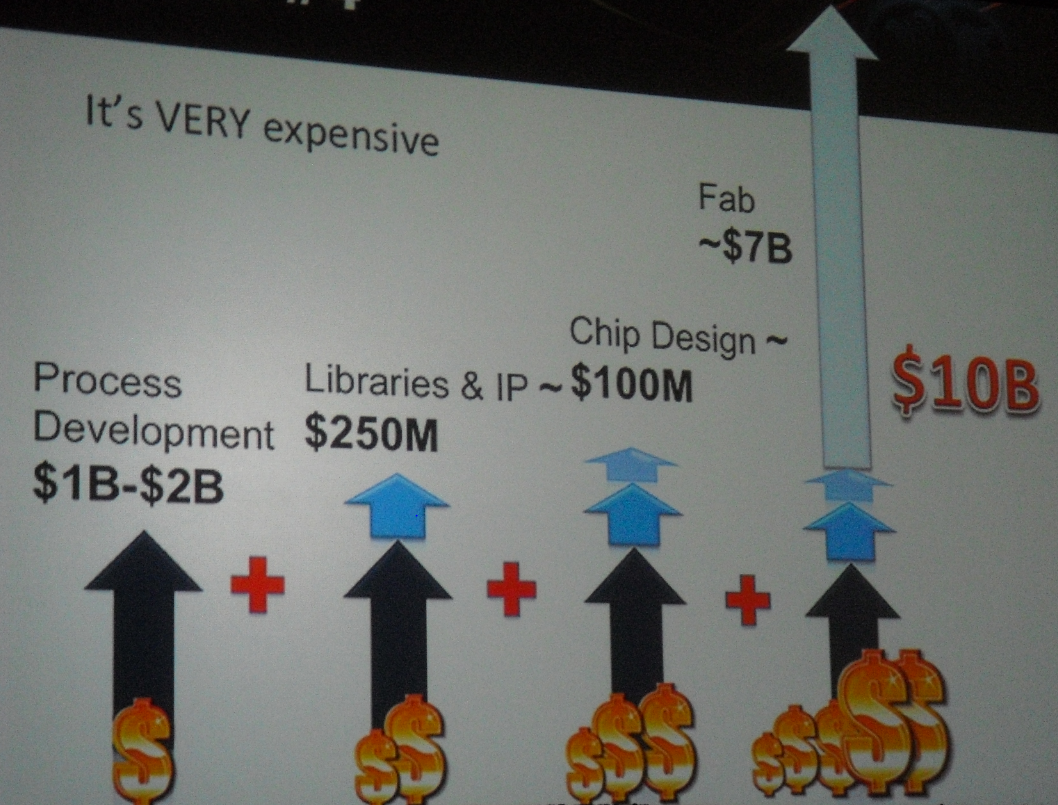
**=>As every 'folding' effectively reduces the average wire length by about 50% it results in reducing the average power by 50%.**

(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**

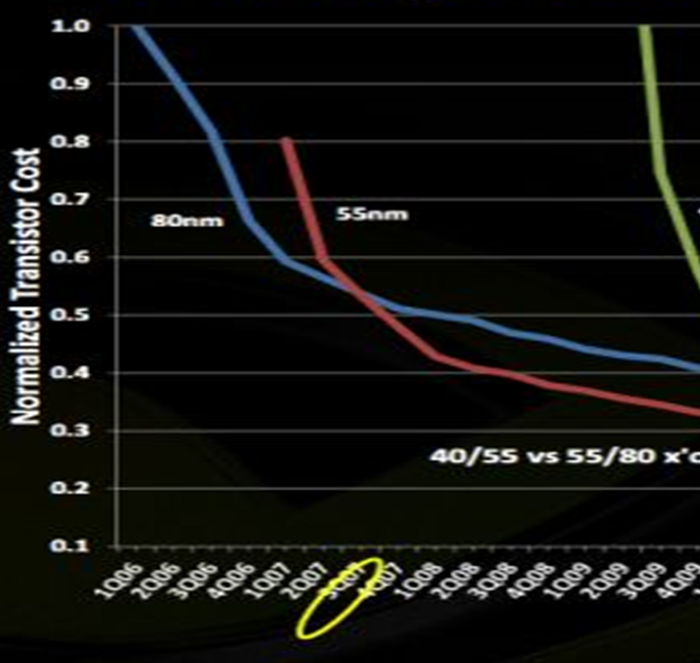
With dimensional scaling every technology/process node requires a significant capital investment for new processing equipment, significant R&D spending for new transistor process and device development, and the building of an ever more complex and costly library and EDA flow. The following charts illustrate this escalating cost trend:



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

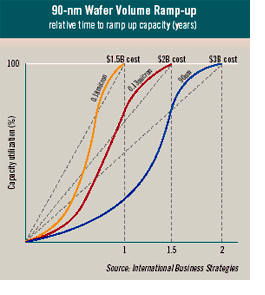


**B. Learning Curve - Yield**

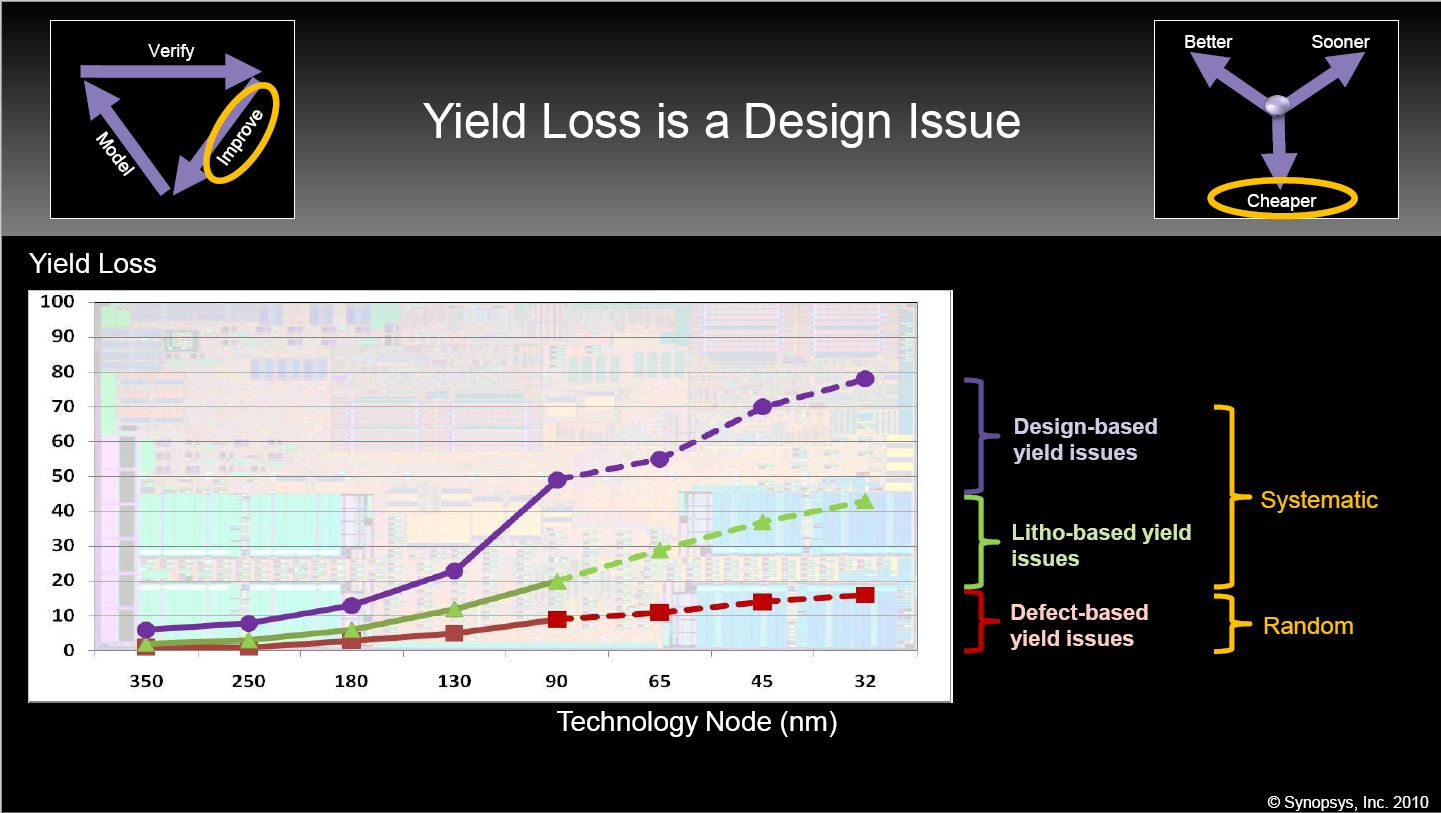
Using the same transistor tools and EDA has an additional important benefit. Learning curve equals yield improvement. With dimensional scaling we face the predicament that by the time we know how to manufacture a process node well, that learning quickly becomes obsolete as we are quickly moving on to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



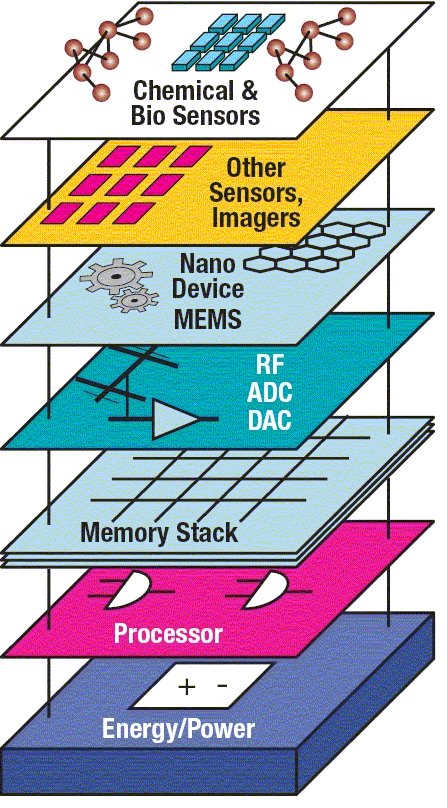
The design and litho based yield loss is growing quickly as the technology node gets dimensionally smaller.

**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.

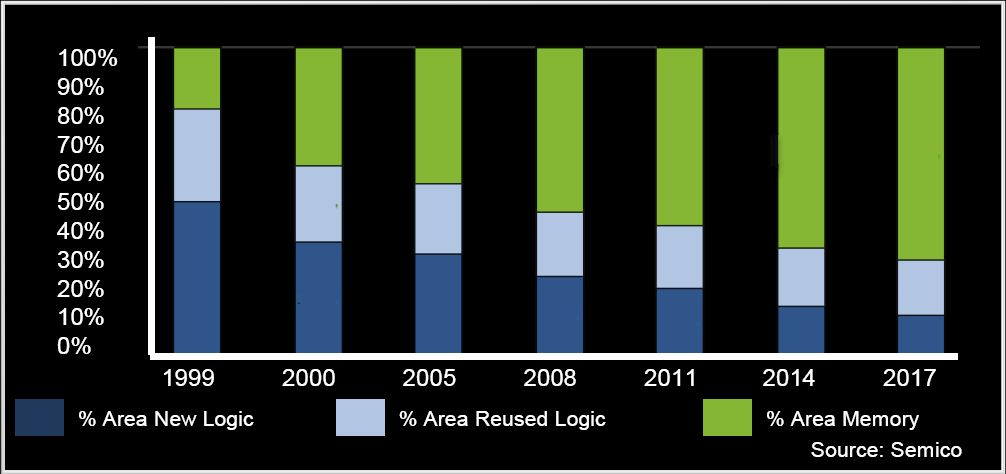


**A. Logic, Memory, IO**

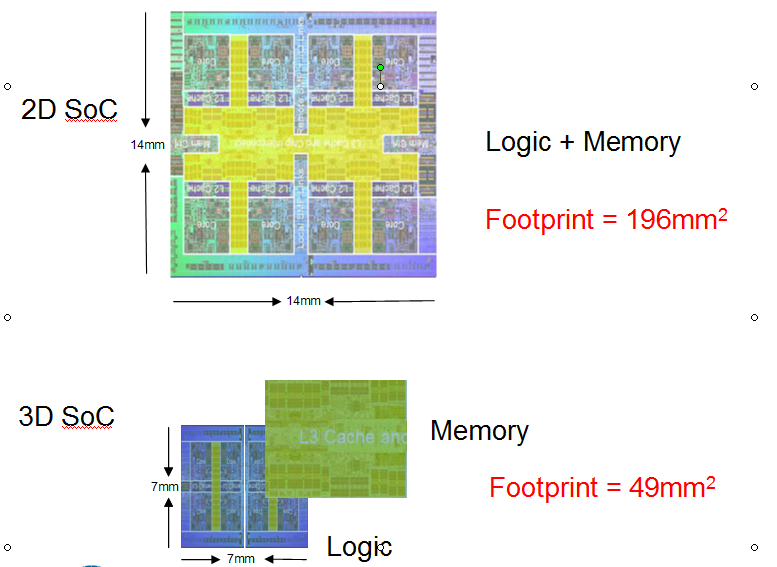
Let’s start with quoting Mark Bohr, in charge of Intel’s process development:

"[**Bohr**:](http://semimd.com/blog/2011/07/14/one-on-one-with-mark-bohr/%3e) 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](http://www.monolithic3d.com/3d-embedded-ram.html)



**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.

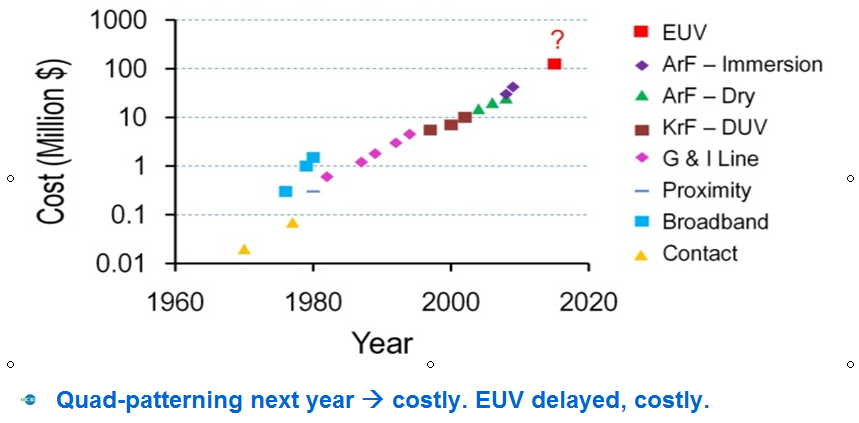
**C. Strata of different substrate crystals and fabrication processes.**

3D enabledheterogeneous 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 Together**

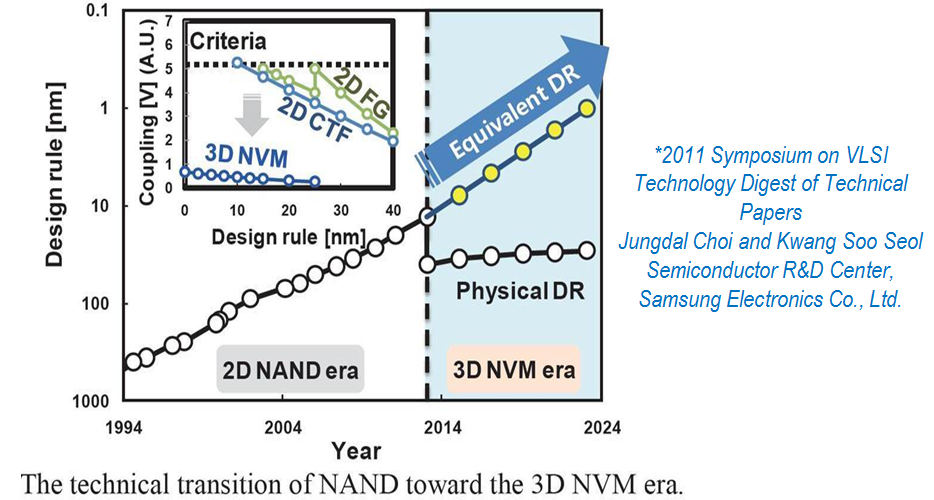
An extremely powerful unique advantage of monolithic 3D is the option to process multiple layers in parallel following one lithography step. This option is most natural for regular circuits such as memory, but it is also available for logic circuits.

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.



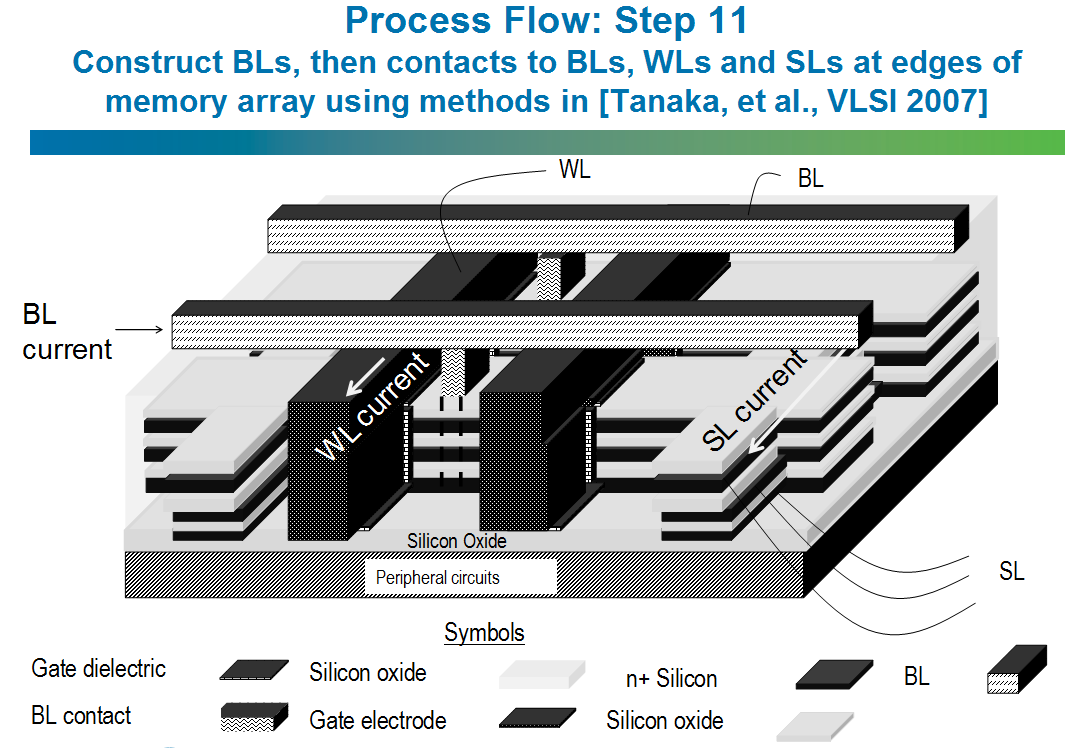
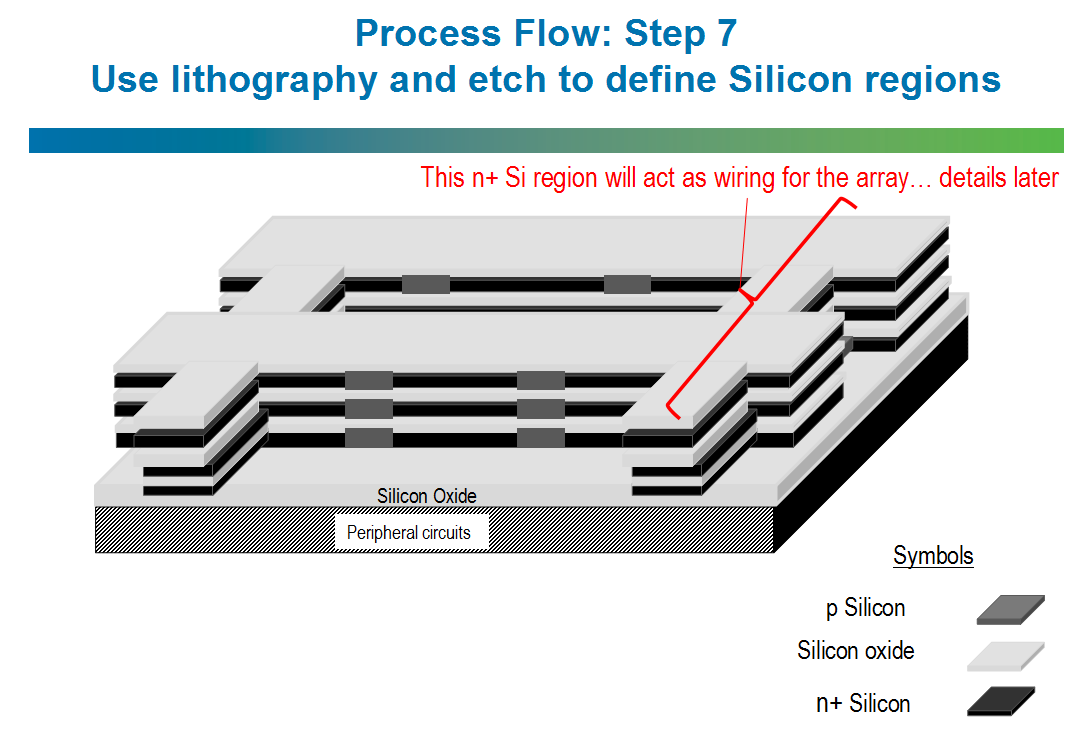
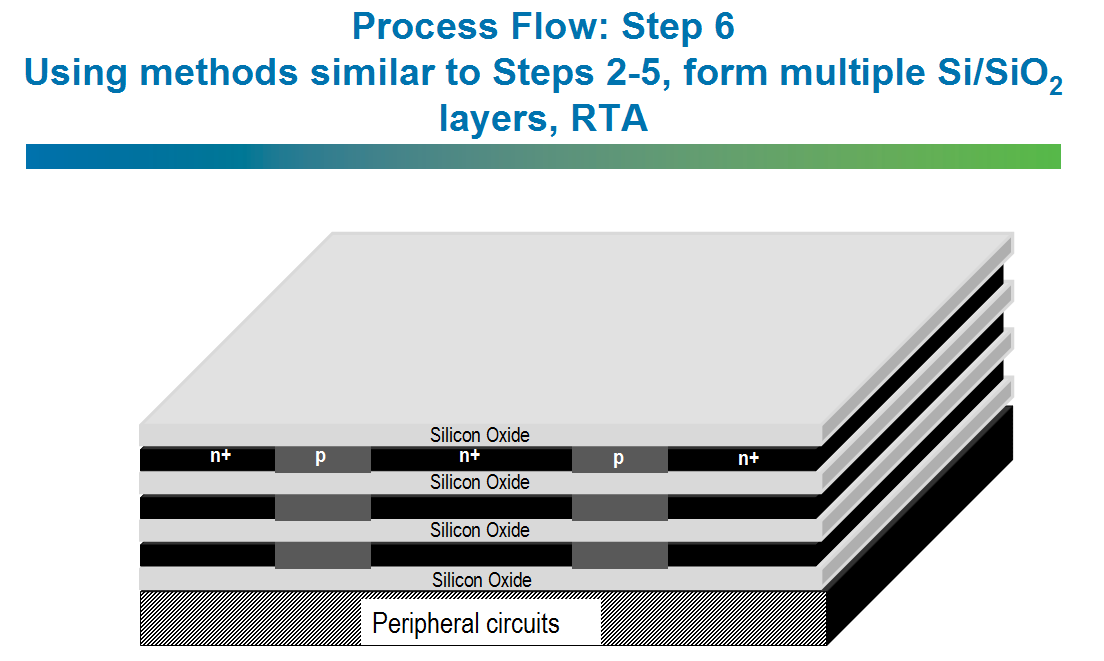
Currently the critical lithography steps dominate the end device production costs. Accordingly, if the critical lithography step could be used once for multiple layers rather than multiple times for each single layer, then the end device cost would roughly be reduced in proportion to the number of layers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



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:



The MonolithIC 3D Inc. website presents more details for such a [DRAM](http://www.monolithic3d.com/monolithic-3d-dram.html) flow, and also related flows for [RRAM](http://www.monolithic3d.com/monolithic-3d-resistive-memories.html) and [NAND Flash](http://www.monolithic3d.com/monolithic-3d-flash.html) memories.

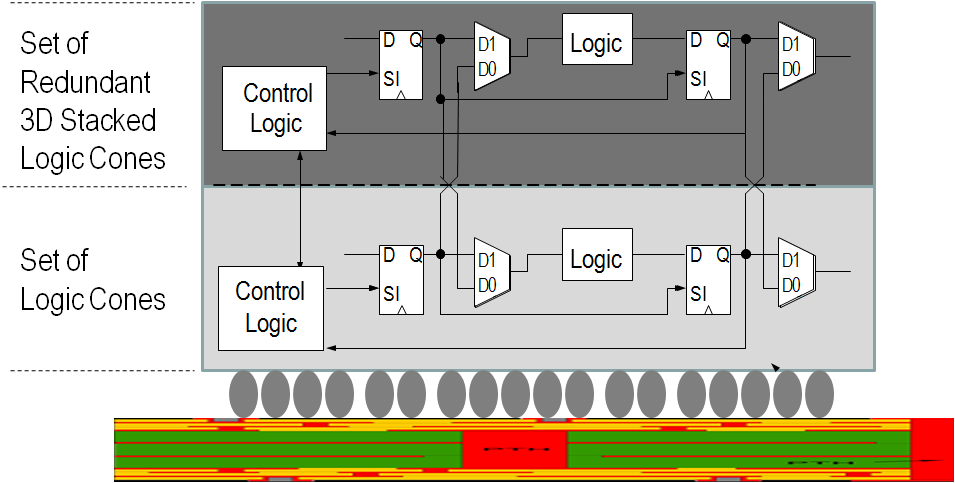
**6. Logic redundancy allowing 100x integration with good yield**

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!!!***

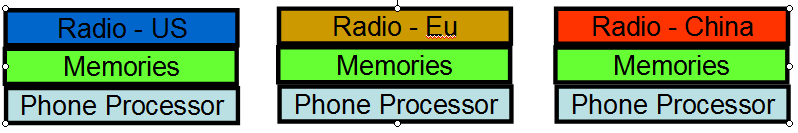
The ultra-integration value could be as much as:

* ~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.



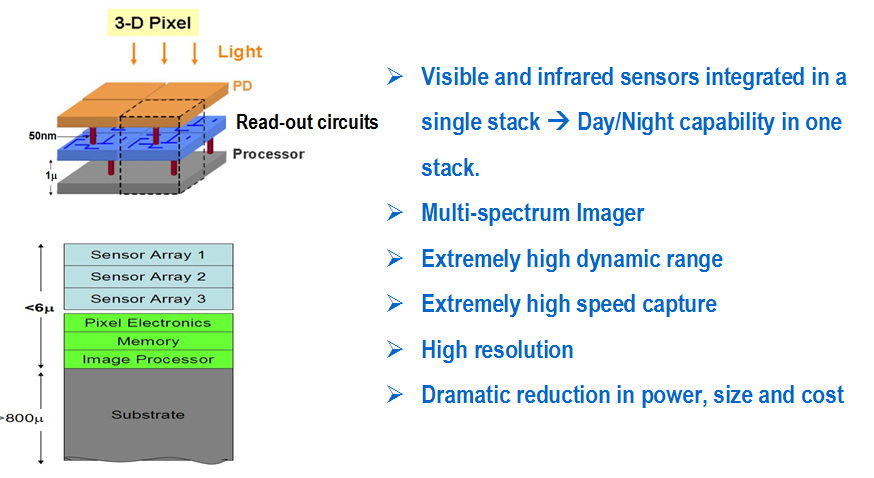
Such a system architecture could be inexpensively used in many market segments and with multiple variations. An interesting one could be in the FPGA sector where the same platform could come with many flavors of memories and I/O.

**8. 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**

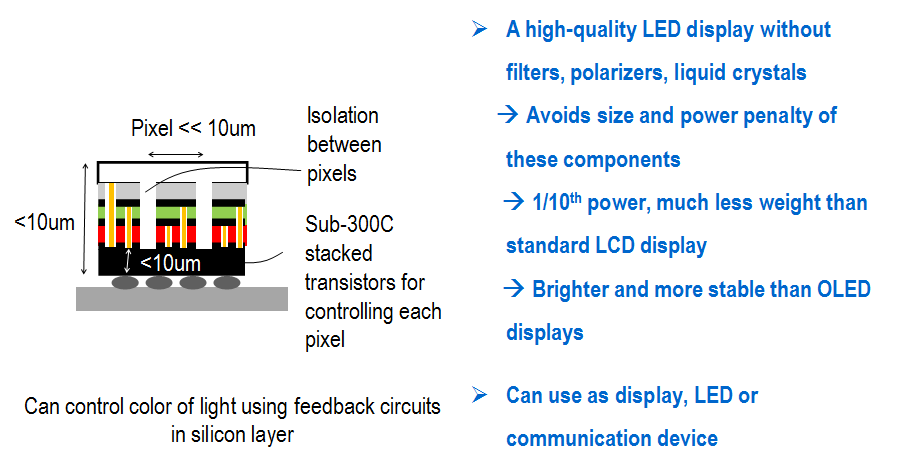
The image sensor industry has moved to back-side illumination to increase the image sensor area utilization. By adding the option for multiple layers many additional benefits could be gained as illustrated below:



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.

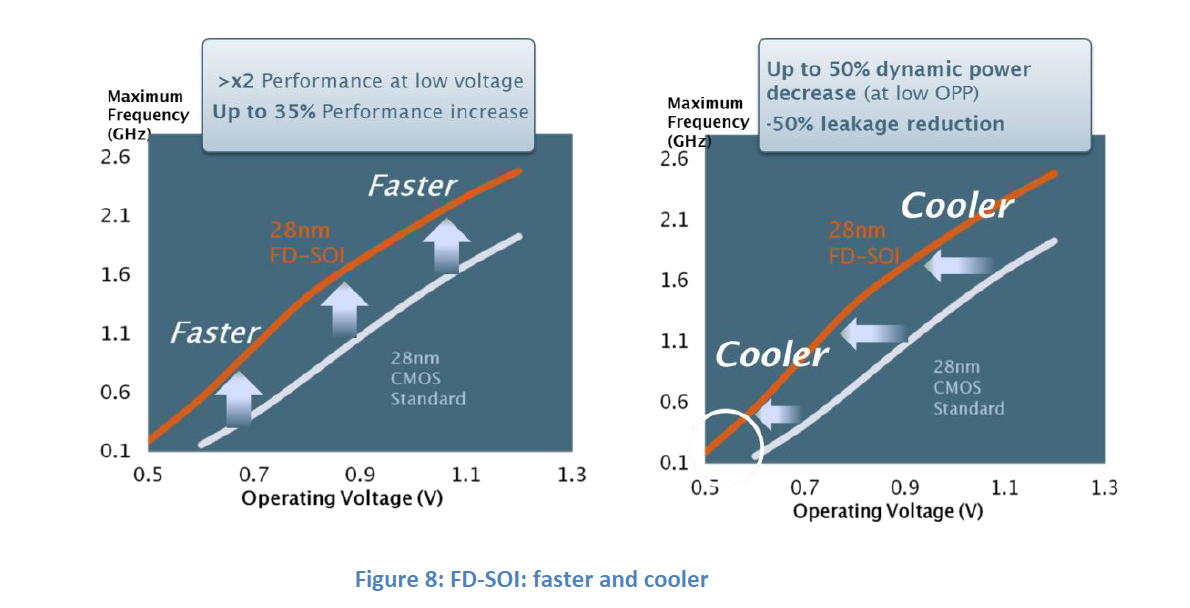
**B. Micro-display**

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.



**C. SOI (FD-SOI)**

The upper layer of monolithic 3D devices are naturally Silicon-On-Insulator (SOI). The advantage of SOI is well known and the recent progress of Fully Depleted SOI (FD-SOI) and SOI-FinFet has taken that advantage much further as illustrated below.



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

**9. 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.**