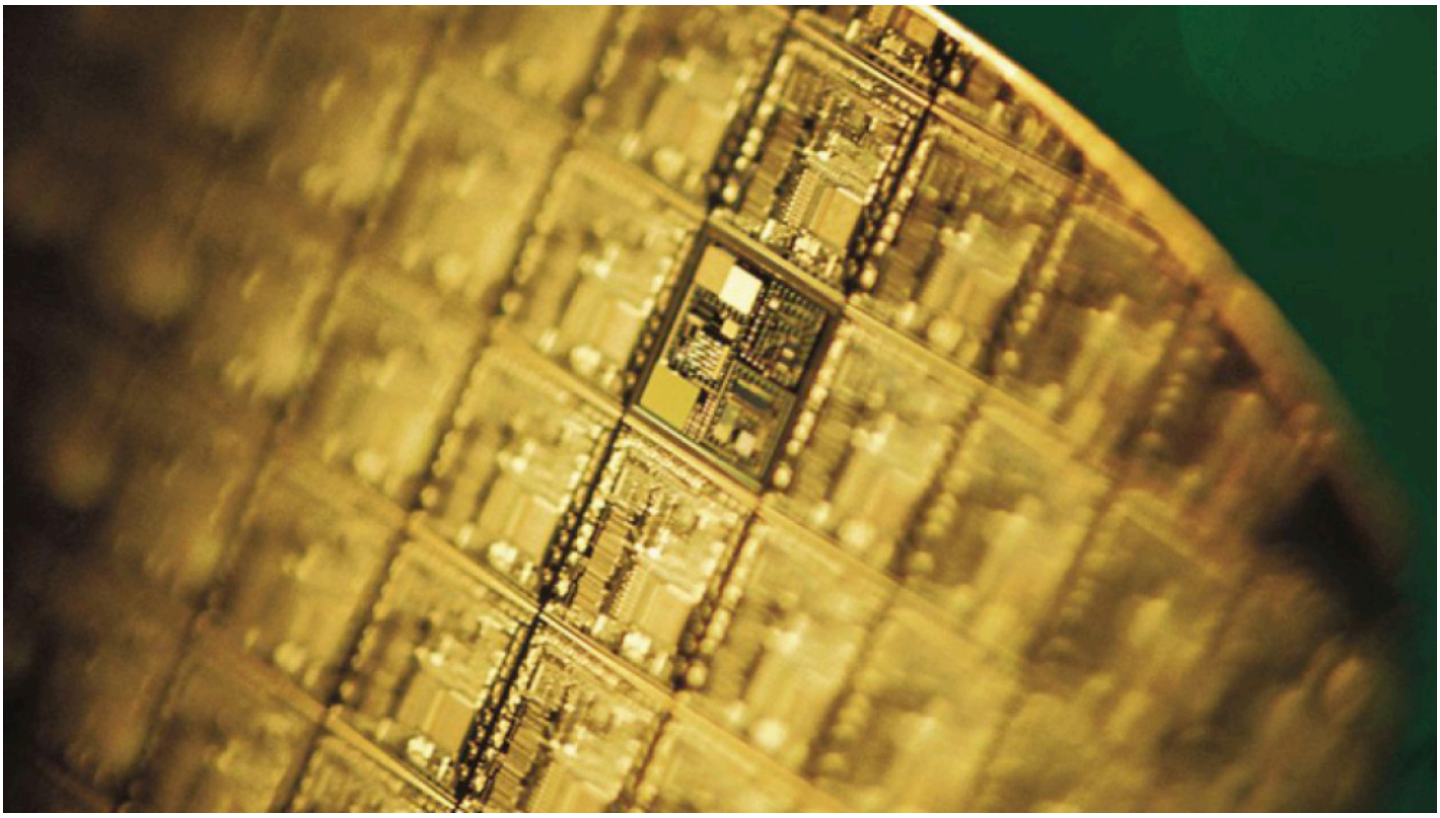
The future of CPU scaling: Exploring options on the cutting edge

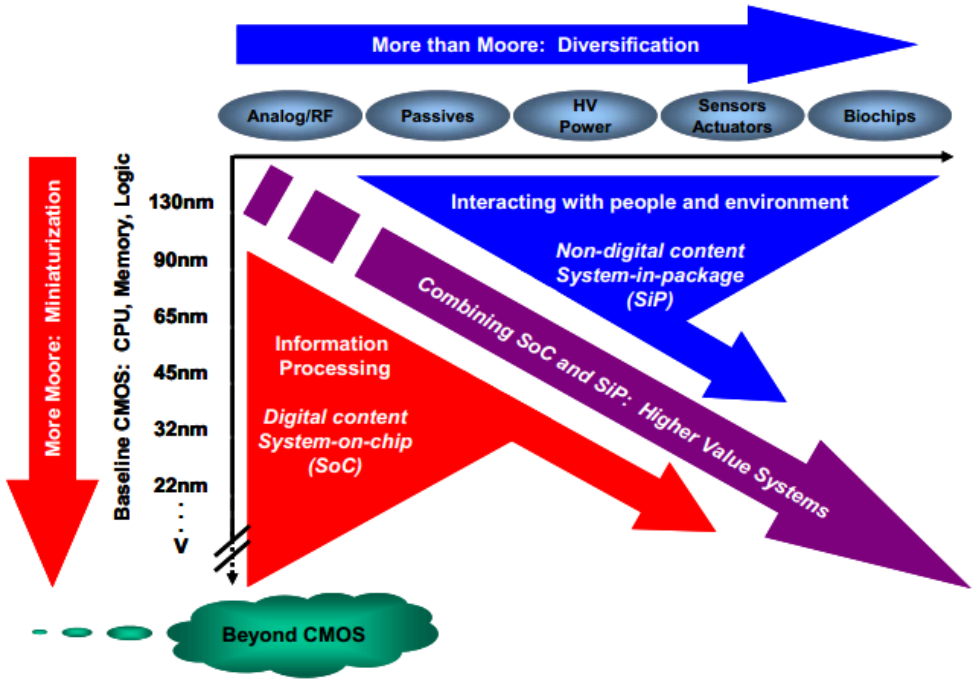
In our first piece, we considered the death of CPU scaling -- here, we examine alternative materials, the future of product integration, and what lies beyond Moore's law

By Joel Hruska February 28, 2012

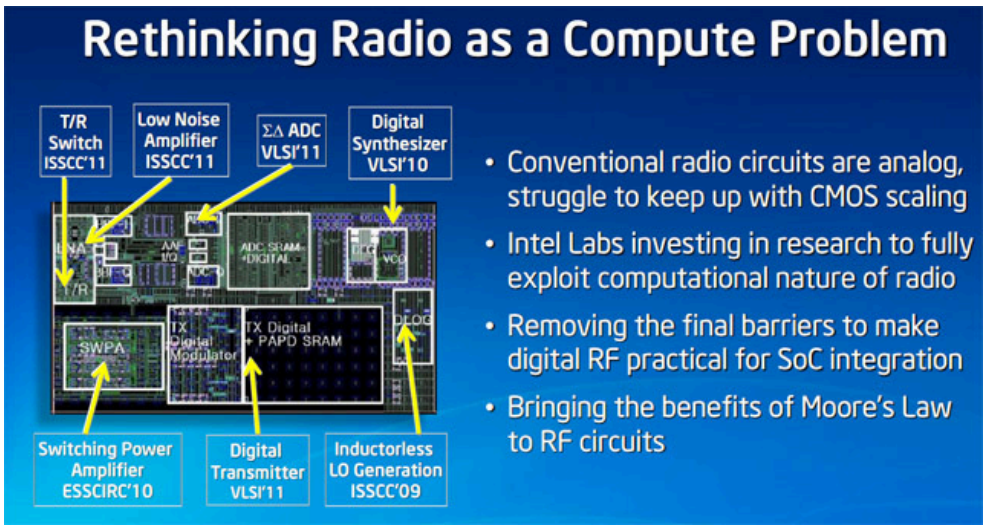


In our first article, we discussed the problems facing CPU scaling and how **neither multi-core or heterogeneous many-core designs** are a long-term solution. This follow-up addresses what the semiconductor industry is doing about it. It's a question best answered in two parts: Near-term innovations (think 3-5 years out) and longer-term research initiatives.

Near-term, the ITRS (International Technology Roadmap for Semiconductors) is focused on what's referred to as "More-than-Moore" (MtM) scaling. The goal of MtM scaling is to extend the same design principles that've driven digital device scaling for decades over to analog circuitry, and to integrate those technologies on-die within an SoC/SiP. The goal of MtM scaling is to increase system-level power efficiency and capabilities, provide a coherent, regular roadmap for the relevant technologies, and increase device complexity. The original proposal that laid out the MtM concept states "it is the heterogeneous integration of digital and non-digital functionalities into compact systems that will be the key driver for a wide variety of application fields, such as communication, automotive, environmental control, healthcare, security and entertainment."



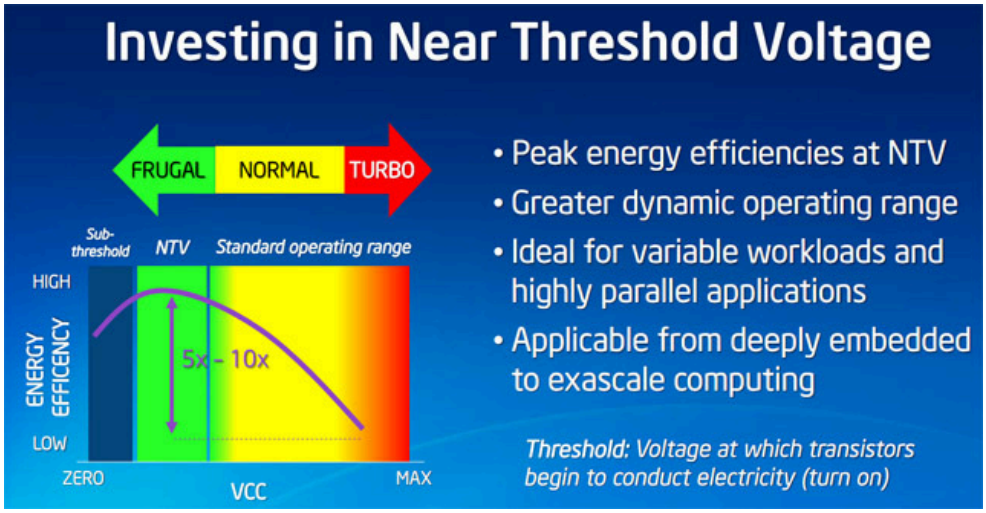
A recent article at the *New York Times* discussing the use of medical devices **capable of interfacing with a mobile phone** is a timely example of exactly the sort of integration More Than Moore is meant to address. Instead of focusing strictly on the CPU as the enabler of experiences, MtM emphasizes integration and efficiently designing every component. As the MtM paper states, "Whereas More Moore may be viewed as the brain of an intelligent compact system, 'More-than-Moore' refers to its capabilities to interact with the outside world and the users."



The original MtM proposition didn't discuss converting analog circuits to digital ones, but Intel's recent description of a fully digital radio in an ISSCC 2012 paper is another demonstration of the theory. Existing smartphone radios have digital basebands but rely on analog circuits for most other components. Analog transistors are more difficult to scale than their digital counterparts, which is part of why a true digital radio makes sense. Intel's paper claims it's possible to effectively shield the CPU and

radio from interfering with each other -- previously, this obstacle would've prevented such a combination.

One of the most striking characteristics of current semiconductor research is how completely the search for lower-power devices has subsumed the old clockspeed obsession. 0W has become the new 1GHz; success is measured in how long your chip spends in its lowest-possible power mode and how quickly it transitions. Performance, the old God of Computing is now merely an efficient means to achieve the lowest possible minimal power usage.



Intel's research into digital radio and Near Threshold Voltage (NTV), Samsung's **PenTile displays**, which use an RG-BG subpixel configuration rather than the conventional RGB-RGB stripe, **3D chip stacking**, and **increased SoC integration** are all examples of how industry focus has shifted to a device-centric viewpoint. The work in this area, however, isn't limited to mobile hardware. The K supercomputer, designed by Fujutsi and Riken, was **fine-tuned for power efficiency** by matching power supply output to the optimal voltage requirements of each physical CPU. The research team estimates that this saved ~1MW of power and reduced operating costs by ~$1M USD per year.

Futjitsu's approach may never translate directly to the PC or mobile industries, but it's indicative of how manufacturers are looking in non-typical areas to find ways to reduce power consumption and improve efficiency. It dovetails with research projects

into improving multi-core and many-core designs and ties to features as diverse as Intel's TSX extensions in Haswell and the idea of implementing programs directly in hardware as a way to leverage increased transistor counts.

These are the sorts of projects and initiatives we expect to dominate device cycles for the next 3-5 years. Past that point, the situation changes.

Next page: The long-term picture

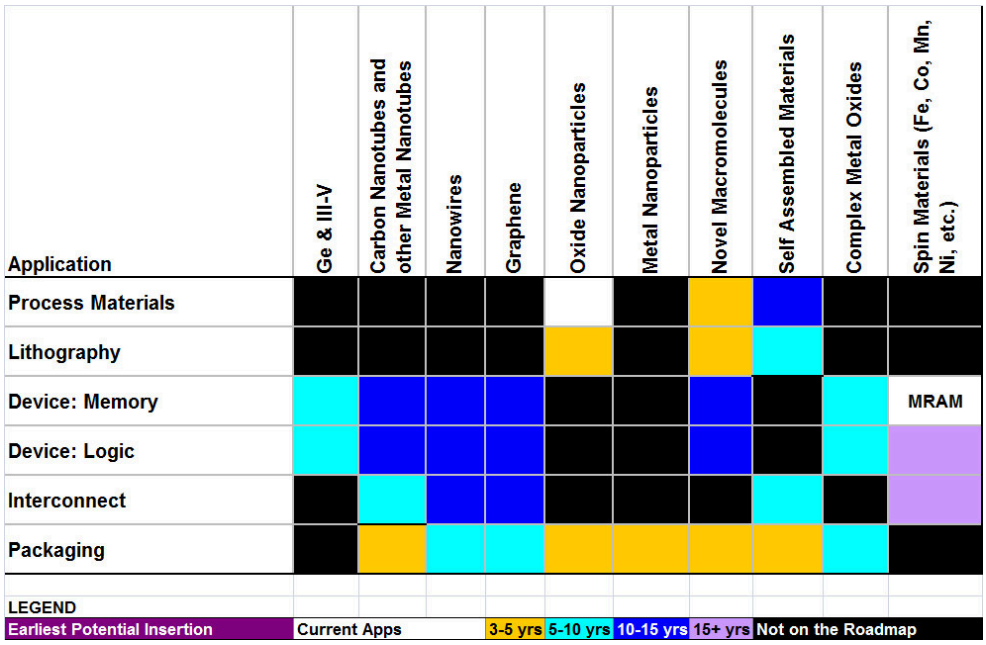
The long-term picture

The 2011 ITRS report contains an exhaustive consideration of the currently known options for driving semiconductor development. The difficulty of this process can't be overestimated -- trying to develop replacements for silicon and/or CMOS is like trying to find a replacement for water.

Dihydrogen monoxide is fantastically interesting stuff. It's a universal solvent, transparent to the visible light spectrum, and capable of flowing against gravity via capillary action. It's highly effective for cooling and requires a relatively high energy input before it transforms into a gas. It becomes less dense (and expands) when frozen and combines easily with a number of other liquids. It has a high refractive index, it's abundant, has a neutral pH, is easy to purify, and is stable at room temperature. There are plenty of fluids that exhibit superior characteristics to water in specific areas, and none that match its overall characteristics. Similarly, there are a number of technologies that could augment current manufacturing processes or replace key elements like silicon -- but nothing even half so simple as a drop-in replacement.

The ITRS report examines the current and future utility of carbon nanotubes, graphene, oxide nanoparticles, metal nanoparticles, complex metal oxides, spintronics, nanowires, and the adoption of germanium and so-called III-V semiconductors. The only recent technology not mentioned is molybdenum; memristors are discussed in a different chapter of the report.

The documents aren't exactly grim, but they highlight the seriousness and complexity of the problem. The following table breaks down various steps in the manufacturing process and when a host of technologies might be ready for integration. Device: Logic refers to CPUs and GPUs, from smartphone chips to high-end workstation hardware.



If you've wondered why so many companies are focused on device packaging (primarily 3D chip stacking) these days, here's part of the reason. A number of cutting-edge technologies should be ready for rollout in this area within 3-5 years, which makes packaging an area where companies like Intel and AMD can expect to offer innovation. Virtually all of the other areas are choked with roadblocks, challenges, and obstacles. Graphene and carbon nanotube transistors may grab headlines, but the ITRS believes both technologies are 10-15 years away from commercial deployment in memory and logic devices.

The following table breaks down how the adoption of III-V materials, Germanium, graphene, nanowires, and carbon nanotubes are expected to compare to existing technologies. A score of 2 indicates that a new technology is equivalent to existing methods, a score of 3 would be far better than current technology. Anything below a 2 is a step backwards.

Assessment Ratings						
1 = Worse than existing technologies		2 = Same as existing technologies			3 = Better than existing technologies	
Earliest Potential Insertion Date		2013-2018	2013-2018	2018-2023	2018-2023	2018-2023
Chip-Package Interconnect Metrics	Research Targets	ACM Application Opportunities and Ratings				
		p-III-V Materials	Ge [For n-channel devices]	Graphene	Nanowires [Group IV and	Carbon Nanotubes
Demonstrate co-integration on Si: Ability to achieve high performance	Mobilities: >> Scaled strained Silicon, Quasi ballistic, saturation	2.3	2.4	2.3	2.3	2.2
Gate dielectric compatibility and control	Unpinned Fermi level, 10% thickness (1s), Dit <1E11 cm-2	1.8	1.9	1.3	1.9	1.3
Low contact resistance and variability	Contact resistivity <1E9	2	1.9	1.3	1.8	1.3
Property control [Eg, Ion: Ioff, assembly/ packaging stress, etc.] in an integrated structure	10% (1s)	1.9	2.3	1.2	1.8	1.4
Ability to dope	Demonstrate and understand controlled n and/or p channel & S/D	1.9	1.8	1.4	2	1.3
Control of formation, location, and direction	CMOS compatible catalyst, as warraned, 10% of half pitch, Vertical and/or	1.9	2	1.4	1.7	1.6
Defects	Understand formation mechanism, develop low defect density strategy	1.8	1.6	1.8	1.9	1.6

All of these solutions offer higher potential performance than silicon, but they come with significant tradeoffs. In many cases, the ITRS doesn't actually know of a solution -- the 5-15 year roadmap estimates assumes that problems will be solved within this time frame.

Rocket scientists wish they had it this easy

The sheer scale of the problem is best demonstrated in a paragraph in the ITRS ERD (Emerging Research Materials) chapter. Section 4.2.1.1 discusses carbon nanotubes. It reads: "For SWCNTs (single-walled carbon nanotube) to be viable for future CMOS applications, the ability to grow them with a tight bandgap distribution must be demonstrated... Little progress has been reported in the past two years... with DNA purification purities approaching 99% have been achieved. These levels of bandgap distribution control are far short of the projected requirements *(better than parts per trillion)*." (emphasis added)

Growing CNTs at 99% purity isn't enough; in order to serve as a replacement, SWCNTs need to hit greater than 99.999999999999% purity. That's the equivalent of a twentieth of a drop of water in an Olympic swimming pool. As features become smaller, the need for perfection rises -- ironically, any new technology that tolerated errors more effectively than CMOS would likely be a prime candidate for adoption.

The number of technologies under investigation is proof that no one, not even the ITRS, is certain which approaches will prove most useful. Despite occasional claims to the contrary from those who ought to know better, there is currently no agreement on what technology is likely to replace NAND flash or DRAM.

The big-big picture

Over the mid-to-long term, transistors will continue to get smaller, the advent and eventual adoption of 450mm wafers will reduce per-chip manufacturing costs, and power efficiency will all continue to improve. The industry's focus on concepts like MtM scaling and improving power efficiency and device utilization at every level means that while the gains will be smaller than what we saw historically, they may also occur in more interesting places.

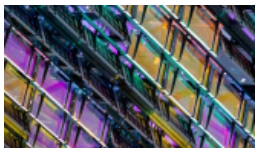
One thing that ITRS charts and diagrams make absolutely clear is that while CMOS may continue scaling well past 2020, the practical *benefits* of that scaling decline from generation to generation. The only way to reinvigorate the roadmap is to find the new driver of semiconductor utility that we mentioned back in part one -- assuming one exists.

The good thing about all this uncertainty is that it gives everyone from Intel and TSMC to Google and Nvidia reason to re-evaluate the fundamental nature of how data is pushed, pulled, shoved, and modified. The failure of Dennard scaling after 30 years of continual improvements may have slammed the door on the easiest and most common type of performance scaling, but there are other avenues to pursue. This necessary exploration of other ways to improve semiconductors could pave the way for a better, more useful class of product than we've had seen if clock speeds and voltages had continued to scale at historic levels.

Tagged In

NTV | ITRS | Digital Radio | TSMC | Graphene

More from Extreme



How Do Solar Panels Work?

04/27/2024 By Jessica Hall

Boston Dynamics Dumps Hydraulics, Unveils All-Electric Atlas Robot

04/17/2024 By Ryan Whitwam

What Is Dark Matter?

04/01/2024 By Adrianna Nine, Graham Templeton

How Do SSDs Work? The Inner Details, Explained

03/24/2024 By Joel Hruska, ExtremeTech Staff

What Is Artificial Intelligence? From Generative AI to Hardware, What You Need to Know

03/22/2024 By Jessica Hall

Popular in the Community

Conversation 11 Comments

Welcome to ExtremeTech comments! Have fun, please be kind, and keep conversations on-topic. Feel free to criticize ideas, not people.

Log in Sign up

What do you think?

Sort by **Best**

Edward Casey

28 February, 2012

I particularly like articles that show things like this. Two many people, in my opinion, think religiously that science can solve any problem, given time. Some problems will be have no solution that is worth the cost; some things that show no promise will improve to become mainstream and many thing...**See more**

Reply ▪ 5 ▪ Share

1 reply

geoffcunningham

2 March, 2012

'Technology isn't a ladder - it is a finite plane'
Iain M. Banks

Might just be that silicon is the optimum for electronic ICs. Then, to get better computation you'd need a different principle for storing information and processing. Problem is, at the moment electrons just seem so damn good!
...**See more**

Reply ▪ 5 ▪ Share

2 replies

Sapan Bhuta

3 March, 2012

Truly excellent article!

Reply ▪ 4 ▪ Share

johnybizzaro

8 May, 2014

Again the main problem is finding new materials and new techniques for manufacture. I have said it over and over. The current process is just B.S. We have 120W TDP cpus. We need a new material . That is not easy. It costs money but as usual the private sector won't do it because the system is brok...**See more**

Reply ▪ 2 ▪ Share

1 reply

LarchOye

20 January, 2014

Intel killed Moore's law, but yet still reference it? So strange.

Reply ▪ 1 ▪ Share

Subscribe Today to get the latest ExtremeTech news delivered right to your inbox.

Email Address... Sign Up

Aerospace
Cars
Computing
Defense
Electronics
Energy
Gaming
Phones
Science
Security
Deals
Shop

askmen
Offers.com
IGN
PCMAG.COM
lifehacker
RetailMeNot
Mashable
SPEEDTEST

ExtremeTech supports Group Black and its mission to increase greater diversity in media voices and media ownerships.

About Ziff Davis
Terms of Use
Accessibility