Andrew Kahng on PPAC Scaling Below 7nm

Comments(0)

Last week Dr. Andrew Kahng came to town. He was at CDNLive, where his presentation *Toward New Synergies Between Academic Research and Commercial EDA* won the **best** paper award for the academic track.Then the following day, he presented at the (internal) Cadence Distinguished Speakers Series, where he talked about *PPAC Scaling at 7nm and Below.* I first met Andrew back when I was at Cadence around 2000, when we were both on the Cadence Technology Advisory Board. He hasn't presented since that era since he opened by saying the last time he was in the River Oaks Cafeteria with ice cream, instead of the building 10 auditorium with pizza. At various times my office was in buildings 1 and 2, which for those of you too recent to know, were tilt-up buildings on the corner of River Oaks and Seely where condominiums now stand.There were four buildings there, which is why buildings on the current campus start from building 5, and, like Spinal Tap's amplifiers, **go up to 11**.

BREAKFAST BYTES

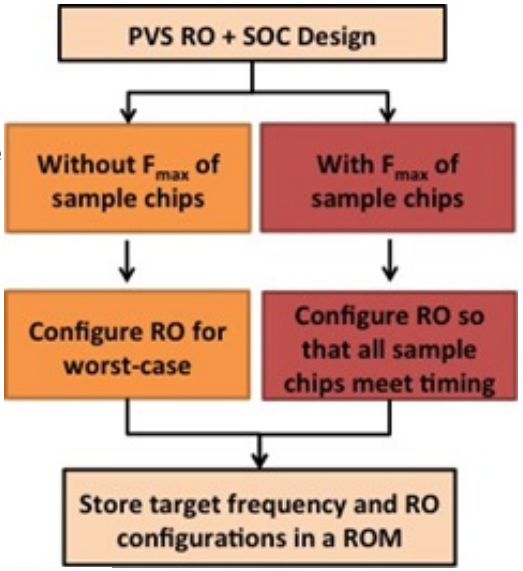
By the way, PPAC stands for power-performance-area-cost. The industry has talked about PPA for a long time, with the A for area also being a surrogate for cost. But with different process choices, multiple patterning vs. EUV someday and other options, area alone is not the only parameter that feeds into cost.

Andrew sees two megatrends that are driving all the issues.

The first megatrend is what he calls "the race to the end of the roadmap." This is advancing Moore's Law to the end of what we know as fast as we can. Despite economic challenges, the technical challenges are being addressed despite a few major issues: the lack of EUV and the lack of a replacement back-end technology to replace copper, restricted design rules, reliability limitations. The result is volume production of 7nm in 2018. Another major issue is guard-banding with excessive pessimism, making design excessively hard and impacting yields.

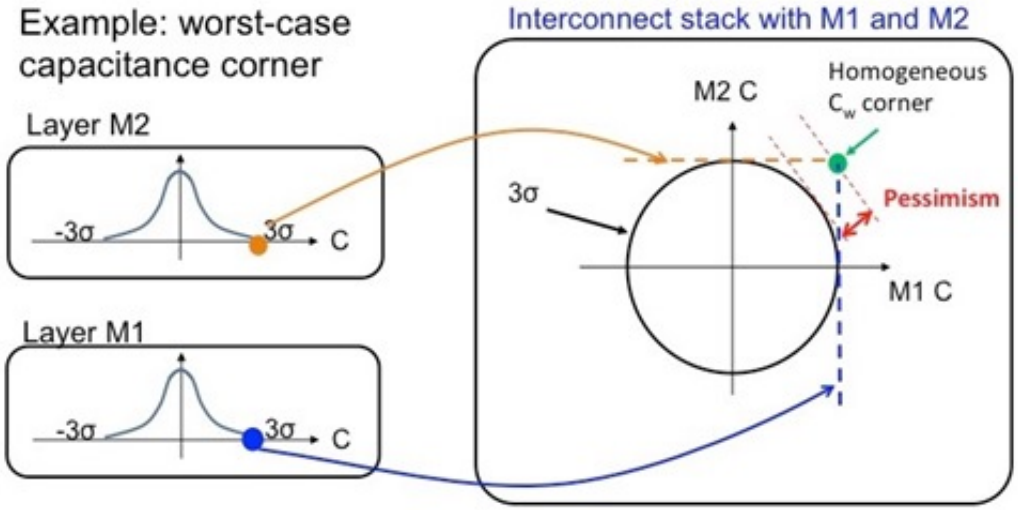
The second megatrend is keeping power under control. Low power is essential in all markets, from mobile to big data to cloud. We have done a lot of the easy stuff in previous process generations and now need more extreme approaches.

One major tradeoff is optimization versus schedule. Moore's Law is 1% per week, meaning that schedule trades off with PPAC. There are other rules of thumb, such as a mV reduction in voltage margin translates into another 5MHz of operating frequency.

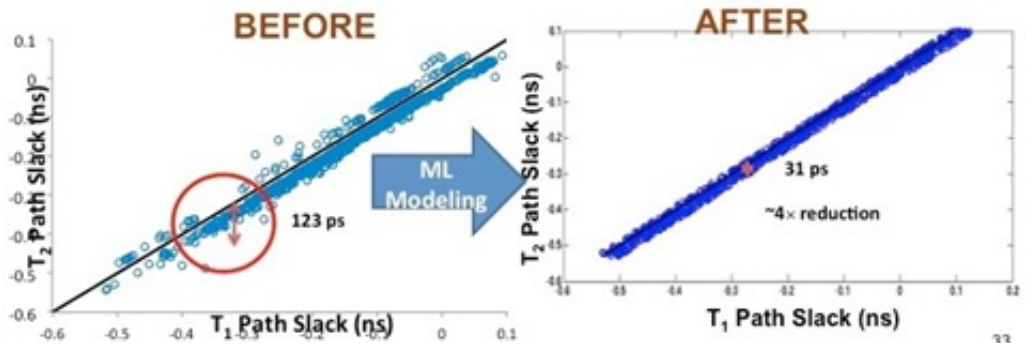


Going deeper into the details, the first area that Andrew has been working is on adaptive voltage scaling. There are a number of forms of this, but the basic idea is the same. Instead of fixing the supply voltage based on pre-tapeout analysis, take measurements off the actual silicon and lower the supply voltage to reflect the actual margin. This requires an on-chip monitor, typically a ring oscillator (RO). Andrew's group at UCSD has been looking at ways to do this such that all sample chips meet timing. Since threshold voltages change with aging, voltage scaling can also be used to compensate over time, with the design signed off using aged parameters but performance improved for most of the life (which also slows aging).

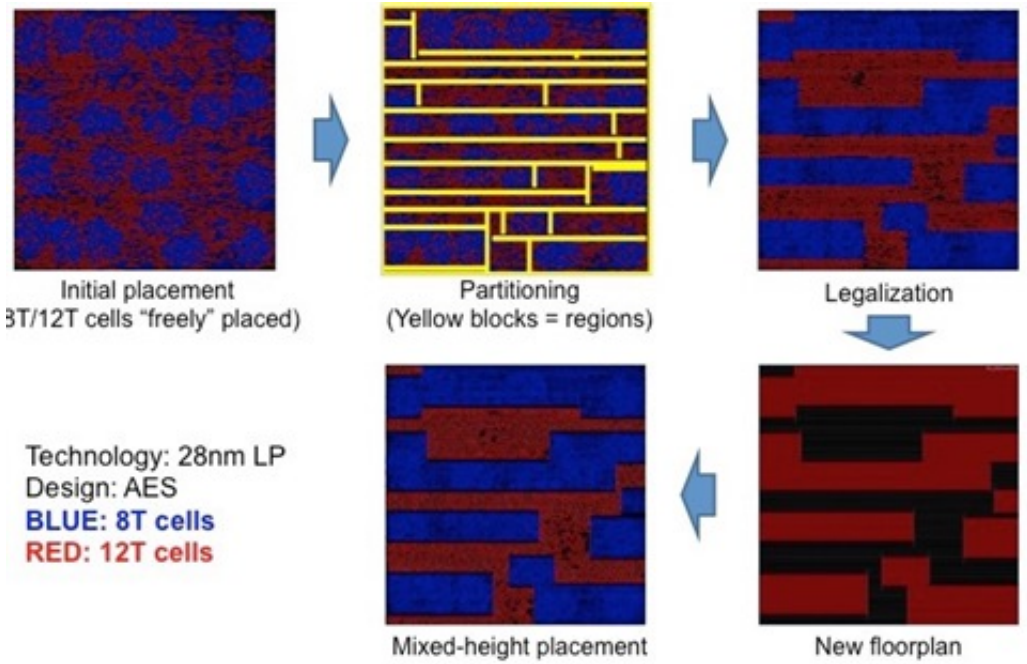
The next area Andrew looked at is that combining worst-case corners derived from 3σ distributions can end up with corners that are much more pessimistic than necessary. In essence, two 3σ distributions combine to form a square when the true 3σ limit is a circle. Since most critical nets are routed on multiple metal layers, and variation in those layers is largely uncorrelated, this creates a lot of opportunity for tightening the corners.



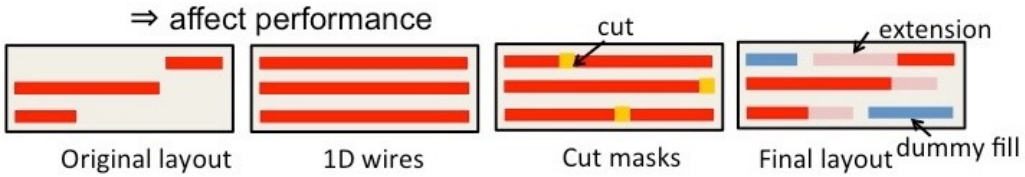
There is "free" margin that can be recovered. Libraries are characterized a certain way, but the reality is that there are tradeoffs between, for example, clock-to-Q, setup, and hold times for a flop. The right tradeoff is not constant and so each FF can use its own "best" set of values compatible with how the silicon actually behaves. There is also a lot of free margin to be picked up when different timing engines produce different results. For example, in one experiment done at UCSD there was as much as 123ps slack divergence resulting in 20% performance difference, which is a whole node of Moore's Law scaling. Applying big data machine-learning approaches can up modeling, reducing that 123ps divergence to just 31ps, a 4X reduction.



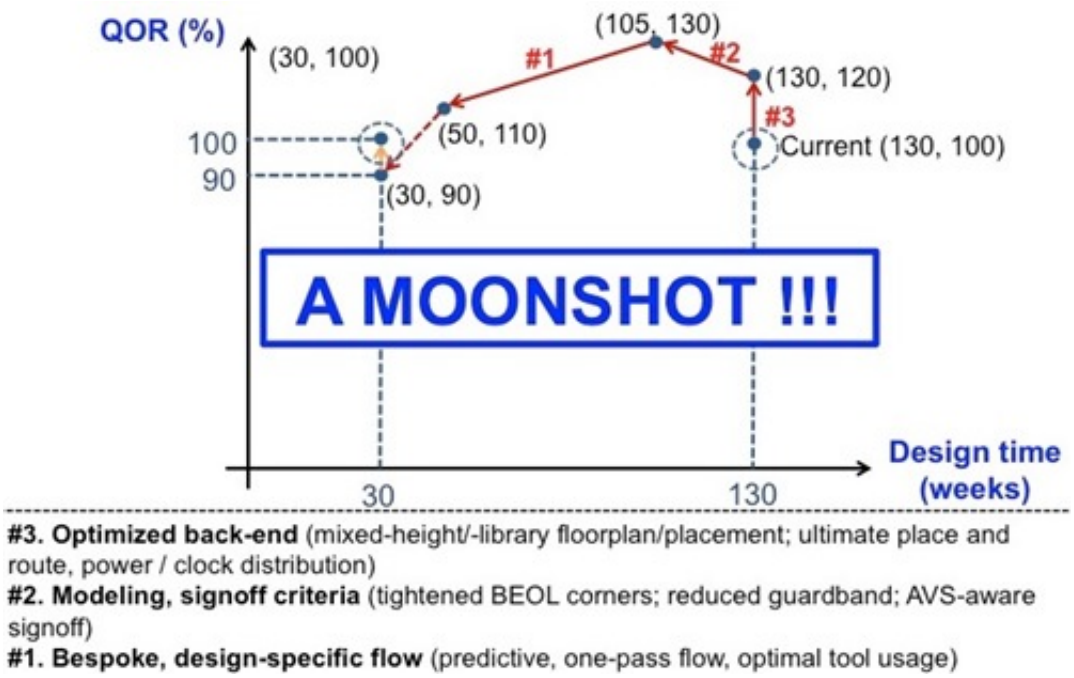
Most foundries have multiple height libraries: large height with better timing but, obviously, more area and more power. Small height has smaller power and area but longer delays and often a requirement for more buffers. Andrew's group has done work on mixing cell height, which normally is not done, and getting better results. Of course, cells cannot be mixed arbitrarily due to things like power supply architecture, but can be partitioned and legalized to end up with mixed areas with better overall results.



At the 7nm and 5nm nodes, layers are one-dimensional grids with divisions done using separate cut masks. This produces more controllable layout that attempting to use few masks and much larger spacing (assuming no EUV for now). The\ interconnect isn't all that needs to be colored; the vias and the cut masks do, too. This creates more opportunity for further optimization when trading off timing against metal density rules and resolution enhancement technology (RET) rules. Optimizing end-of-line extension can produce better tradeoffs.



Andrew's final call to arms was for a massive "moonshot" to predict tool outcomes, find the sweet spot for different tools and flows, and thus design in specific tool and flow knobs to the overall methodology. This would combine all the ideas already discussed (and others that I haven't had space to cover) and so end up with a fully predictive, one-pass flow with optimal tool usage. With modern massively parallel, big data architectures, it is not unreasonable to use tens of thousands of machines if it could "get us to the moon" of a non-iterative flow.



Previous: **Phil Moorby and the History of Verilog**

ucsd roadmap ITRS 5nm 7nm andrew khang power khang

Post