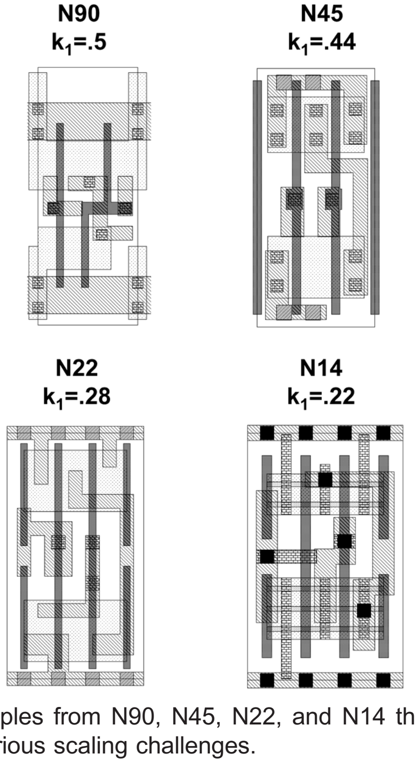
2.2.2 Impact of patterning-driven design constraints

Following explanations of the various resolution barriers that were penetrated in the semiconductor industry's relentless pace of scaling, and a review of some of the topological design objectives that concern standard cell logic designers, Fig. 2.9 illustrates the net impact on the NAND layout from five nodes of scaling across three unique resolution domains, as previously shown in Fig. 1.1; N65 and N32 are not shown because they reside in the same resolution domain as N45.

A few qualitative differences between the cell images in Fig. 2.9 stand out:

- In response to the increasing impact of corner rounding, the poly and diffusion shapes have become rectangular (no notches). This change initially forced the power connections from the diffusion (where they could be shared across the vertical cell boundaries) onto the metal and finally onto the local-interconnect levels. Similarly, signal connections to the poly have been moved from contacts that provide simple vertical connectivity to local interconnects that can also provide some degree of lateral connectivity.
- In response to increasing routability challenges stemming from more constraints on upper-level metal layers, the designer put more emphasis on pin access, resulting in larger and more spread-out pins.



N90 **k_1=.5**

N45 **k_1=.44**

N22 **k_1=.28**

N14 **k_1=.22**

Figure 2.9 NAND examples from N90, N45, N22, and N14 that show the net effect of designer responses to various scaling challenges.

- In response to the increasing impact of proximity effects, the poly was forced onto a fixed pitch, i.e., full-size dummy neighbors are designed into the layout (the only attribute distinguishing them from actual poly gate shapes is that they are not connected and do not form a functional transistor). The proximity effects gradually extended from lithography to other processes, such as etch, and then, with the advent of strained silicon as a performance boost, into device engineering. Selectively adding compressive and tensile stress layers to the transistor allowed device engineers to improve mobility in the channel and provide a performance boost when it became challenging to scale the channel length as the main means of improving device performance. Engineering and modeling these stress layers in strained silicon devices became very difficult without control over the exact dimensions of the source and drain regions. Having already inserted dummy poly for lithographic reasons, device engineers implemented "diffusion tuck under," i.e., all source drain regions had to terminate under a dummy poly at a fixed space to the active poly. (Chapter 4.1 explains how this diffusion tuck under, in combination with a "double diffusion break," i.e., an empty poly track to the next active transistor's dummy poly, increased the cell width by one poly pitch. However, this increase in width was the only scaling detractor for the NAND over the entire N130-N10 range.
- As discussed earlier, in response to diminishing diffusion efficiency, FinFET was introduced to provide a substantial boost in that scaling parameter.

While the design styles of the NAND cell shown in Fig. 2.9 clearly evolve towards increasingly restricted geometries, the primary impact on cell area was an increase in the cell width to accomodate the diffusion tuck-under and double-diffusion break. Other than this one poly pitch growth in width, the cell area scaled at the same ratio as the critical pitches. Therefore, one might assume from this simple comparison that logic scaling was largely unaffected by all of the pattering constraints over these five technology nodes. However, it is important to consider that, for example, it took a single contact layer in N90 to connect the first metal to the devices; in N14 it takes seven mask levels on four different process layers to make that connection. Additionally, the process-complexity increase in moving from planar devices with poly gates to FinFET with high-k metal gates is staggering. Ultimately, the scaling impact must be measured not only by the loss of transistor density but also the process cost and complexity.

The NAND used in these discussions is part of a class of logic cells referred to as combinatorial logic cells that execute Boolean logic functions such as AND, OR, NOR, and AOI (i.e., "and or invert"). As illustrated in Fig. 2.10, during the operation of a logic circuit a signal is sent through a series of combinatorial logic cells before it reaches a latch in which the state of the logic signal is memorized before it is launched through another set of

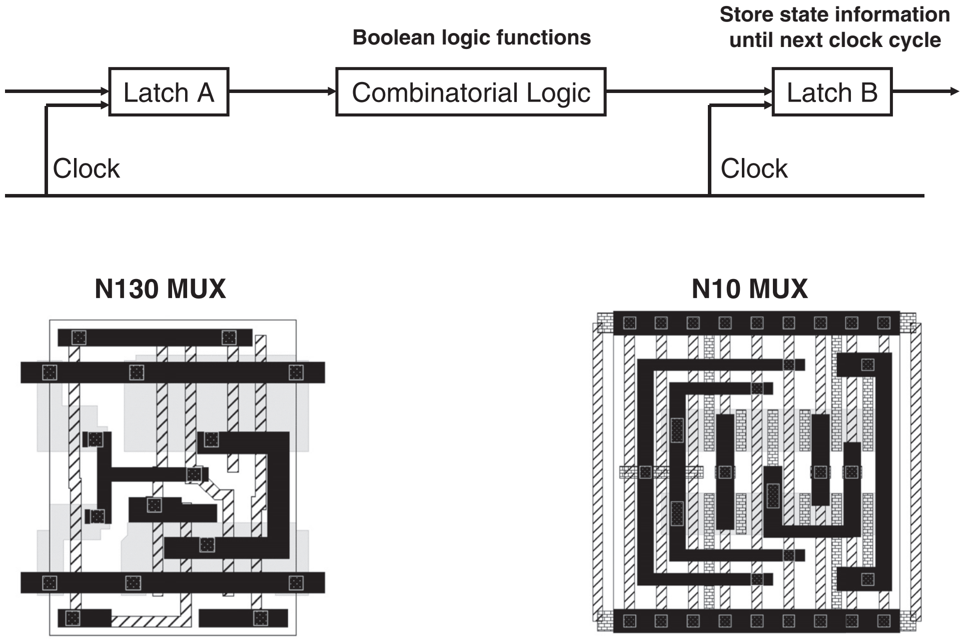


Figure 2.10 A large portion of a logic block is occupied by latches. The challenge of scaling these complex sequential logic cells is illustrated by comparing a N130 and N10 multiplexer (MUX) layout.

combinational logic cells in the next clock cycle. Because these latches contain memory as well as logic functions, they are referred to as sequential logic cells, more importantly; they tend to be the most complex layouts in a logic block. A typical logic block consists of 30–40% latches by area, so it is very important to scale these logic cells very efficiently. The bottom half of Fig. 2.10 compares the N130 and N10 node renderings of a multiplexer (MUX), a critical design element in a latch.

Although a patterning engineer will marvel at the regularity exhibited by the N10 MUX layout, the scaling challenges encountered in the complex logic cells are undeniable. Figure 2.11 breaks down the cumulative scaling impact into individual steps based on the restrictions incurred by the most fundamental construct in the MUX, the poly gate over the diffusion intersection that forms the transistor. Figure 2.11(a) shows three transistors as they might have been used somewhere in a complex logic cell. Dense packing is achieved by staggering the transistors vertically, which is made possible by a complex diffusion shape that provides some local wiring capability in addition to forming source/drain regions and by the freedom to put poly on a range of pitches.

After corner rounding became an issue (due to the loss of diffracted orders at low k_1), diffusion corners had to be moved far away from active gates, and wiring on the diffusion became very inefficient; the resulting loss of *stacked devices* is shown in Fig. 2.11(b). As discussed earlier, restricting poly to a fixed

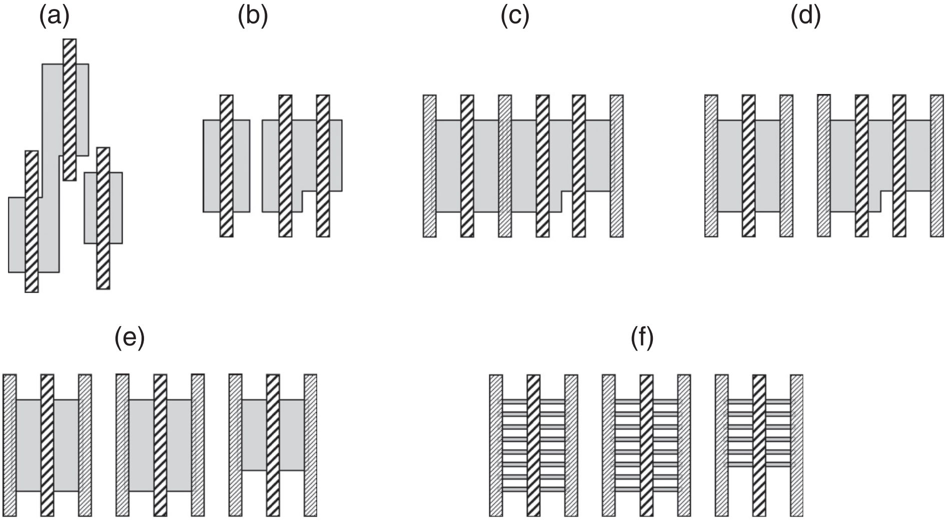


Figure 2.11 Increasing the constraints on transistor layouts in complex logic cells: (a) densely packed transistors using diffusion routing, (b) loss of stacked devices, (c) dummy poly with diffusion tuck-under, (d) double diffusion break, (e) loss of tapered devices, and (f) introduction of fins.

pitch (to achieve better linewidth control at low k_1) was quickly exploited by device engineers in the era of strained silicon, and *diffusion tuck-under* was introduced, as shown in Fig. 2.11(c). The formation of a robust isolation between separate diffusion shapes tucked under the same poly became a yield and reliability concern and lead to the introduction of *double diffusion break*, as seen in Fig. 2.11(d). Diffusion-shape corner rounding, especially in combination with the process complexity introduced by FinFET, eliminated the possibility of having two devices of different width share the same diffusion shape leading to the loss of *tapered devices* in Fig. 2.11(e).

Finally, the price to pay for the high diffusion efficiency provided by FinFET is the coarse granularity with which the device width can be controlled (to balance power/performance for a given circuit). While it was previously possible to adjust the device width in single design-grid steps—often as small as 1 nm—FinFET changes the effective device width in integer multiples of the fin count. Cumulatively, the design restrictions outlined in Fig. 2.11 cause 20–40% less area scaling on complex logic[23] than possibly achievable based on the pure linear scaling of critical dimensions. Although there is a clear correlation between patterning, device, and integration challenges and the resulting loss of scaling, the exact technology node at which scaling penalties were incurred in the path from N130 to N10 varies by design. Product designs focused on high-performance, early yield on large chips and extreme reliability tend to adopt design restrictions earlier than product designs competing on the basis of cost, density, and low power consumption.

Furthermore, integrated device manufacturers (IDMs) tend to negotiate design restrictions with their internal design teams more effectively than foundries competing for fabless design customers. Many leading-edge fabless design companies have acquired deeper process expertise to not only negotiate more-aggressive design rules but also assess the risk of not adopting design restrictions in time. More discussion of construct-based technology scaling is covered in Section 4.1 as part of the DTCO overview.

The scaling impact assessment in this section focused primarily on the loss of area scaling and the increase in process complexity. The following sections provide a qualitative view of the increase in design complexity encountered with the introduction of double patterning.

2.3 Standard Cell Layout in the Era of Double Patterning

With the introduction of double patterning in the N14 node, designers were confronted with two-color decomposability (some felt a sense of déjà vu for when the semiconductor industry experimented with altPSMs). As outlined in Section 1.5, LELE double patterning requires the layout to be cleanly separable, or *decomposable*, into two masks. As Fig. 2.12 shows, even for a moderately complicated layout it is difficult to judge whether a particular collection of shapes is decomposable or not. After the layout is colored following simple same-color versus different-color spacing rules, as shown at the bottom of Fig. 2.12, it is easy to identify un-decomposable layouts by the presence of shapes that violate the same-color-space rule. Although many academic and industry papers have been written on efficient decomposition algorithms, it is more important to create a set of design rules, checking tools, and methodologies to prevent decomposition errors.

The topic of how much color information designers need to see in LELE became a rich source of material for design and patterning conference evening panel discussions. The different foundries' marketing teams became divided between advocates of "colored" and "colorless" design flows even though the fundamental differences in these flows were minor compared to the overall complexity introduced by double patterning. Figure 2.13 shows three variants of the LELE-enhanced cell-level design flow. Providing the designers with split-level design rules, i.e., simply stating that the space between shapes on different masks is n while the space between shapes on different masks is $2n$, allows designers to create colored designs that pass DRC without any further complications as shown in Fig. 2.13(a). To assist this split-level design methodology, EDA tool suppliers developed interactive tools that resemble real-time spell checkers and automatically color shapes as they are placed into the context of other colored shapes. In contrast to this explicitly colored design methodology, the color-aware spacing rules can be provided to an automatic decomposition tool that runs under the covers just prior to DRC;