A Novel Framework for DTCO: Fast and Automatic Routability Assessment with Machine Learning for Sub-3nm Technology Options

Chidi Chidambaram[‡], Andrew B. Kahng[†], Minsoo Kim[†,*], Giri Nallapati[‡], S.C. Song[‡] and Mingyu Woo[†]
[†]UC San Diego, La Jolla, CA, USA, [‡]Qualcomm Technologies Inc., San Diego, CA, USA (*E-mail: mik226@ucsd.edu)

Abstract

We report, for the first time, block-level area scaling that is reversed from cell-level scaling in cell height <150nm regime, using a machine learning (ML)-assisted design-technology co-optimization (DTCO) methodology to optimize block-level area accounting for routability. >400 unique standard-cell architectures are studied, combining cell height (CH, 80∼150nm), contacted poly pitch (CPP, 39∼57nm), metal pitch (MP, 16∼30nm) and use/non-use of buried power rails (BPR). Cell- and design-level routability assessments are performed to obtain routability index (a.k.a. K_{th}) and utilization of designs. CH <120nm with four available tracks increases block-level area due to routing difficulty, showing diminishing return from further pushing of ground rules. BPR improves block-level scaling with CH <120nm, but benefit slows down with CH <100nm. A 1:2 gear ratio (MP:CPP) improves block-level area compared to a 2:3 ratio. This DTCO flow is automated, and the ML-based prediction expedites the process.

Introduction

Design-technology co-optimization (DTCO) is an essential, high-value element of technology enablement. However, DTCO has proved to be a very expensive process; large turnaround times and engineering efforts are needed to develop standard-cell libraries and perform comprehensive block implementation experiments. Turnaround time of DTCO at advanced nodes is weeks to months with hundreds of engineers. Meanwhile, standard cells have been optimized to aim for minimum cell area (cell height) for decades. However, cell-level area scaling does not necessarily result in block-level area scaling at advanced nodes. As standard-cell area is scaled down, local routing and pin accessibility of standard cells become significantly difficult. Various *scaling boosters*, such as buried power rails (BPR) [7], backside PDN and supervias, are proposed to improve power, performance, area and cost (PPAC). Scaling boosters also have an impact on routability. Due to significant routability impacts of standard-cell architectures and scaling boosters, block-level area must be estimated at an early stage of technology development.

Recently, DTCO methods for technology definition have been proposed in many aspects. [1] proposes a machine learning (ML)-based approach to find optimal combinations of design, technology and other ingredients for high-performance CPU designs. However, there is no automatic design enablement stage (i.e., standard-cell library generation), and the methodology requires four to six weeks of turnaround time. [8] proposes an ML-based modeling framework for devices. The framework can generate compact models of novel devices without prior knowledge. On the other hand, the work does not consider block-level evaluations.

In this work, we apply the PROBE2.0 framework [2] to evaluate various technology options and configurations for sub-3nm technology nodes. We generate 448 unique standard-cell architectures, combining cell height (*CH*, 80∼150nm), contacted poly pitch (*CPP*, 39∼57nm), metal pitch (*MP*, 16∼30nm) and use/non-use of buried power rails (BPR). We study achievable block-level area with automatically generated standard-cell libraries across four main axes: (i) available M2 routing tracks (RT) on standard cells, (ii) use of BPR, (iii) gear ratios for MP to CPP, and (iv) different available M2 RT with same CH. We also apply ML-assisted routability prediction to expedite assessment.

DTCO Framework and Configuration for Sub-3nm Nodes

In this section, we summarize PROBE2.0 [2], the framework used for routability assessment in DTCO, along with the configurations for sub-3nm standard-cell libraries that we study. As shown in Fig. 1(a), the PROBE2.0 framework includes automatic standard-cell layout generation anddesign enablement generation, and (cell- and design-level) routability assessments. The basic idea of PROBE2.0 is to measure inherent routability, represented by K_{th} metric, through neighbor-swap operations applied to canonical placements. Fig. 2 illustrates the neighbor-swap operation. For a given cell, we randomly choose a neighboring cell and swap the locations of the cell pair. Neighbor-swaps progressively increase routing difficulty by "tangling" the placement. We let K denote the number of neighbor-swap operations, normalized to the number of instances (standard-cells) in the design. As K increases, the number of post-routing design rule check violations (#DRCs) likely increases; we define K_{th} as the maximum K for which #DRCs is less than a prescribed threshold. PROBE2.0 advances over the previous PROBE [5] with a satisfiability modulo theory (SMT)-based "automatic" standard-cell layout generation [3][6] and an ML-based K_{th} prediction, as illustrated in Fig. 1(b).

In this work, we generate nine standard-cell libraries to study routability and achievable block-level cell density at sub-3nm technology nodes. Table II shows sets of parameters for the nine libraries, as follows. (i) Lib{1,2,5,6} have 4 RT while Lib{3,4,7,8} have 5 RT. (ii) Lib{1,3,5,7,9} have BPR for power and ground pins while Lib{2,4,6,8} have M1 pins. (iii) Lib{1,2,3,4} have a 1:2 ratio for MP to CPP (20nm MP and 40nm CPP) and Lib{5,6,7,8} have a 2:3 ratio for MP to CPP (26nm MP and 39nm CPP). (iv) Lib{3,9} have 120nm CH but Lib3 has 5 RT and Lib9 has four tracks according to MP of Lib{3,9}. Fig. 3 illustrates OAI21_X1 gates in the nine libraries.

Routability Assessment and Block-level Area Case Study

We perform cell- and design-level routability assessments to evaluate generated standard cells. Cell-level assessment is used to assess inherent routability for each cell in a standard-cell library. Design-level assessment, i.e., at the level of an entire place-and-route block, is used to evaluate a set of standard cells in the same library. We also perform case study of cell-level and achievable block-level area. Cell height of a standard-cell library is linearly related to cell-level area. However, cell-level area benefit does not always result in block-level area benefit, due to routability effects. Our case study shows how we can evaluate technology options with respect to block-level area density using the PROBE2.0 framework.

Fig. 4 shows results for the *cell-level* routability assessments. We study 15 standard cells with size X1, in each of four libraries (Lib{1,2,3,4}). Results in Fig. 4(c) show that standard cells with M1 have better routability than those with BPR, and standard cells with five tracks have better routability than those with four tracks. Further, standard cells with larger numbers of input pins have worse routability than those with smaller numbers of input pins. Figs. 4(a) and (b) show #DRCs vs. K plots for NAND and OAI gates, respectively.

Fig. 5 shows results for *design-level* routability assessments with four designs, AES, LDPC, JPEG and VGA. Fig. 5(b) shows K_{th} metric per standard-cell library per design while Fig. 5(a) shows #DRCs vs. K plots for the AES design. Fig. 6 shows results of our achievable utilization study for AES and LDPC designs. We observe strong positive correlation between K_{th} and achievable utilization.

In addition, we perform case study for block-level area benefits across four main axes of technology options.

- Case 1: Available M2 routing tracks (RT) (Lib1 vs. Lib3).
- Case 2: Use of buried power rails (Lib1 vs. Lib2).
- Case 3: Gear ratios for MP to CPP (Lib1 vs. Lib5).
- Case 4: Different RT with the same CH (Lib3 vs. Lib9).

In Case 1, Lib3 has 20% larger cell area than Lib1. However, based on the achievable utilization (0.71 and 0.92 for Lib1 and Lib3, respectively), the achievable block-level area with Lib1 is 7% larger than with Lib3. It is important to note that in this case, reduction of RT brings less cell-level area, but the block-level area actually increases. In Case 2, Lib2 has M1 PGpin while Lib1 has BPR, so that Lib2 also has 20% larger cell area than Lib1. Applying the same calculation as with Case 1, we obtain that the achievable area with Lib2 is 19% greater than with Lib1. Thus, in Case 2, the cell-level area benefit from use of BPR is reflected as a design-level area benefit. In Case 3, Lib1 has a 1:2 gear ratio for MP to CPP and Lib5 has a 2:3 gear ratio. Since Lib5 has 26nm MP, Lib5 has 30% more cell area than Lib1. But, block-level area of Lib5 is 14% larger than that of Lib1 since the respective Lib1 and Lib5 achievable utilizations are 0.71 and 0.62. Last, in Case 4, Lib9 has the same CH (120nm) as Lib3, but MP of Lib9 is 24nm instead of 20nm. Larger MP brings benefits of reduced resistance, but Lib3 incurs 15% area penalty at design-level. Table III summarizes our case study. As shown in Table III and Fig. 6(b), block-level area no longer changes linearly with CH at advanced nodes. Even when CH increases, block-level area can decrease.

Machine Learning (ML)-Assisted Routability Assessment

PROBE2.0 [2] uses ML-based K_{th} prediction to reduce the number of place-and-route (P&R) implementation runs needed for accurate routability assessment. However, large training sets (e.g., 80% of libraries) are used. In this work, we apply Latin hypercube sampling (LHS) [4] to select an asymptotically more efficient training set for ML-based K_{th} prediction. In a case study, we create 448 standard-cell libraries as follows: (i) CPP = 39, 42, 45, 48, 51, 54, 57nm; (ii) MP = 16, 18, 20, 22, 24, 26, 28, 30nm for M1 and M2; (iii) 4 and 5 RT; (iv) BPR and M1 PGpin; and (v) M3 pitch:CPP ratios of 1:2 and 2:3. We choose training sets of 64, 128 and 256 data points (i.e., standard-cell library and K_{th}) by LHS and use all remaining (resp. 384, 320, 192) data points for model testing. Fig. 7 shows comparisons for golden