

Physical Synthesis for Power under Process Variation

Jason Cong
cong@cs.ucla.edu
UCLA Computer Science

Tony F. Chan
chan@math.ucla.edu
UCLA Mathematics

Lieven Vandenberghe
vandenbe@ee.ucla.edu
UCLA Electrical Engineering

March 15, 2006

1 Introduction

Power is widely considered the only real limiter for Moore’s Law in the next decade [Gelsinger, 2004]. Indeed, IBM CTO B. Meyerson has claimed that CMOS process scaling has already stopped between 130nm and 90nm due to power limitations [Clarke, 2004]. Existing methods for physical synthesis, however, have primarily targeted timing closure and are inadequate to meet the challenge of low-power physical design under process variation. The current lack of a general formulation to simultaneously consider and balance all objectives and constraints leads to potentially gross suboptimality.

The goal of the research proposed here is the formulation and implementation of a *complete* placement-centric physical synthesis flow to minimize total power consumption of an IC design scalably under evolving timing and netlist constraints in the presence of uncertain wire-and-device characteristics. *Complete* means simultaneous optimization over the following sets of design variables under tight timing constraints.

- (i) locations $z_i = (x_i, y_i)$ for modules $i = 1, \dots, N$
- (ii) supply voltages V_{dd_i} and threshold voltages V_{th_i} assigned to modules from a fine-grain voltage-island power network
- (iii) module sizes l_i
- (iv) gate oxide thickness t_{ox_i}
- (v) effective channel length L_{eff_i}

For brevity, these variables are referred to collectively as $q = (z, v, l, t_{ox}, L_{eff})$.

Variables V_{dd} and l impact dynamic power, while variables V_{th} , L_{eff} , and t_{ox} impact static power through current leakage. Although each module will be given a distinct location z_i and possibly a different size l_i , the variables V_{dd_i} , V_{th_i} , t_{ox_i} and L_{eff_i} are ultimately assigned not independently to individual modules but rather to subregions based on the overhead introduced by level converters and the granularity afforded by the available manufacturing technology. Nevertheless, fine-grain voltage islands have been demonstrated based on the novel design of low-cost level converters [Puri et al., 2003]. Similarly, given the demands of aggressive power minimization, it is expected that the best assignments of V_{th} , L_{eff} , and t_{ox} will vary over many regions on the chip, producing much higher complexity for optimization. It is important to consider the optimization of these variables with placement, because placement (a) defines the interconnects and provides the load and density information for sizing; and (b) is needed for assuring the spatial locality required (or preferred) for multiple V_{dd} , V_{th} , L_{eff} , and t_{ox} assignment as well as for power and clock gating. It is also important to optimize these variables simultaneously, as they all compete for the same timing slack in power minimization. Previous work on power minimization has considered (i) adjusting only proper subsets of these

variables simultaneously, often under a fixed placement, or (ii) sequential optimization over multiple variables. Either of these previous approaches is likely to result in inefficient use of timing slack and a suboptimal power result.

Recent advances in both circuit modeling under uncertainty and optimization under uncertainty will also be used to explicitly incorporate uncertainty due to process variation into all objectives and constraints. This explicit modeling of uncertainty will enable *both* aggressive reduction in total power usage *and* reliable attainment of timing-yield targets.

The proposed algorithm will be built on a scalable and unified multilevel flow for a combination of strictly robust, scenario-based, and chance-constrained formulations of robust optimization under uncertainty. The multilevel flow for global optimization, successfully adapted to wirelength-driven placement by the PIs in SRC Tasks 686.001 and 1091.001, is ideally suited to the simultaneous incorporation of diverse and complex constraints. Early, explicit, and consistent handling of these constraints will ultimately match the performance of critical regions to their timing constraints by giving them sufficient power, while minimizing power everywhere else. Optimization subproblems at each level of the multilevel hierarchy will be formulated and solved rigorously, with (i) all objectives and constraints explicitly modeled in a mathematical-programming formulation, and (ii) the proposed algorithms explicitly targeting that formulation. The multilevel formulation will support thorough exploration of the trade-offs between multiple V_{dd} , multiple V_{th} , module sizes, module locations, t_{ox} , and L_{eff} in the optimization of total dynamic and leakage power.

The remainder of the proposal is organized as follows. Section 2 states the mathematical formulation of physical synthesis in both its deterministic and robust forms. Section 3 describes the outer multilevel strategy for building a scalable and robust solution method. Section 4 describes the optimization algorithms used at each level of the multilevel hierarchy. Section 5 describes how the robust, timing-yield constrained formulation will be reduced to deterministic form and solved.

2 Problem Formulation

Given a provisional netlist with timing constraints and net-switching statistics, we seek a complete geometrical specification of its layout such that total power consumption is minimized and the likelihood that all timing constraints are satisfied under the uncertainty of process variation is greater than a prescribed lower limit. The formulation under uncertainty is presented as an extension of the deterministic formulation below.

2.1 Deterministic Form

In terms of leakage power P_{leak} , dynamic power P_{dyn} , and an a-priori estimate K of the ratio $P_{\text{dyn}}/P_{\text{leak}}$, the proposed deterministic formulation of power-driven physical synthesis is

$$\begin{aligned} & \underset{q=(z,v,l,t_{\text{ox}},L_{\text{eff}})}{\text{minimize}} && P_{\text{leak}}(q) + KP_{\text{dyn}}(q) \\ & \text{subject to} && u_j(q) \leq \bar{u}_j \quad \text{for } j = 1, 2, \dots \\ & && t_i(q) + d_i(q) + e_{ij}(q) \leq t_j(q) \quad \text{all edges } (i, j) \\ & && t_k(q) \leq r_k \quad \text{all } k \in \text{PO} \end{aligned} \quad (1)$$

where $q = (z, v, l, t_{\text{ox}}, L_{\text{eff}})$ is the set of design variables defined in Section 1. The functions u_j with respective upper bounds \bar{u}_j represent *generalized Poisson-based density constraints* for modeling global features such as area utilization, routability, heat distribution, etc. [Eisenmann and Johannes, 1998, Chan et al., 2005c]. As usual, required arrival times r_k at primary outputs $k \in \text{PO}$ are back propagated to the modules in order to obtain a tractable formulation; $t_i = t_i(q)$ denotes the arrival time of a signal at node i , $d_i = d_i(q)$ is the delay at node i , and $e_{ij} = e_{ij}(q)$ is a precise *a-priori estimate* of the optimized propagation delay from node i to node j attainable after buffering and sizing [Cong and Pan, 2001].

Static power is dominated by leakage power P_{leak} . For a gate with effective channel length L_{eff} , a standard empirical model of leakage power [Kao et al., 2002, Mani et al., 2005] $P_{\text{leak}} = c_0 e^{-c_1 L_{\text{eff}} - c_2 V_{th}}$ with fitting constants c_0 , c_1 , and c_2 is used. A more explicit model incorporating drain induced barrier lowering (DIBL) is [Weste and Harris, 2005]

$$P_{\text{leak}} = v_T^2 e^{1.8} \mu \epsilon_{\text{ox}} \frac{W}{t_{\text{ox}} L} e^{\frac{V_{gs} - V_{th}}{n v_T}} (1 - e^{-V_{ds}/v_T}) V_{dd},$$

where V_{gs} denotes voltage between gate and source, V_{ds} denotes voltage between drain and source, v_T denotes thermal voltage, and μ denotes carrier mobility.

Dynamic power is dominated by net-switching power, which can be modeled as $kCV_{dd}^2\alpha f$, where k is a constant, $C = C(q)$ is the total parasitic and load capacitance due to wires and gate-input pins to be charged and discharged, and αf is the switching rate; i.e., the expected number of transition events per unit time [Cheon et al., 2005]. Capacitance is a function of module positions, module sizes, and net lengths; total weighted half-perimeter wirelength is a standard and effective approximation to the objective at early stages [Sarrafzadeh et al., 2002].

2.2 Statistical Form

In sub-100nm IC's, random and systematic variation in gate channel length, doping density, wire length and width must be explicitly modeled in order to obtain reliable yields. The statistical formulation of physical synthesis is obtained by (a) interpreting v_i , t_i , d_i , l_i , t_{ox} , and L_{eff} in (1) as spatially correlated random variables and (b) requiring that the arrival-time constraints hold only with a certain minimum yield probability Y_{min} .

$$\begin{aligned} & \underset{q=(z,v,l,t_{\text{ox}},L_{\text{eff}})}{\text{minimize}} && P_{\text{leak}}(q) + KP_{\text{dyn}}(q) \\ & \text{subject to} && u_j(q) \leq \bar{u}_j \quad \text{for } j = 1, 2, \dots \\ & && t_i(q) + d_i(q) + e_{ij}(q) \leq t_j(q) \quad \text{all edges } (i, j) \\ & \mathbf{P}(t_k(q) \leq r_k \text{ all } k \in \text{PO}) && \geq Y_{\text{min}}. \end{aligned} \quad (2)$$

Many of the random variables can be modeled as Normal or Log-Normal [Mani et al., 2005] with known means and variances; but in general, their distributions are unknown. Systematic effects in manufacturing are manifested as spatial correlations. Accurate formulation of the variations in v_i , t_i , d_i , l_i , t_{ox} , and L_{eff} as determined by fundamental physical processes is another goal of the proposed research. Proposed techniques for the deterministic reformulation and solution of (2) are described in Section 5.

3 Combined Multilevel Algorithm

The outer flow of the core algorithm combines optimization by simultaneous placement, module sizing, voltage assignment, and t_{ox} and L_{eff} optimization in a multilevel formulation. Multilevel optimization strongly supports (i) scalability and parallelizability; (ii) correct handling of complex constraints, including timing, routability, heat dissipation, noise, etc.; (iii) the incorporation of multiple, diverse, and complementary optimization heuristics; (iv) adaptability to rapidly changing formulations of multiple objectives and constraints.

Multilevel optimization consists of four main elements [Brandt, 1986, Cong and Shinnerl, 2003]: coarsening, relaxation, interpolation, and iteration flow.

Coarsening— recursive aggregation, or generalized clustering, produces order $\log(N)$ approximations or *cluster levels* of the original problem (1), each smaller than its predecessor by a prescribed factor. Each cluster level is defined by an explicit mapping of design variables, objective, and constraints at its adjacent finer level to corresponding variables and functions at the cluster level. Thus, at each level, all modules in the same cluster will have one location, the same V_{dd} , V_{th} , and t_{ox} , and the same scaling factor for gate sizes and L_{eff} , resulting a much simplified problem.

Relaxation— optimization within each level begins with a solution inherited from an adjacent level and adjusts all design variables either simultaneously or in some sequence in order to reduce the objective until a stopping criterion is met.

Interpolation— A solution at one level is transformed to a solution at the adjacent finer level by mapping values of the design variables associated with coarser-level aggregates to values of design variables associated with their finer-level components. Thus, values for clusters' locations, V_{dd} , V_{th} , l , L_{eff} , and t_{ox} are refined to values for the modules or sub-clusters composing them.

Iteration Flow— In the simplest V-cycle form of multilevel optimization, a single pass of recursive aggregation is followed by aggressive coarsest-level optimization and a single pass of recursive interpolation and relaxation. However, recursive correction by frequent recoarsening under evolving aggregation criteria is widely cited in the literature as a more effective approach for difficult problems [Brandt and Ron, 2003].

Proposed strategies for relaxation are derived from (i) generalized Poisson-based force-directed placement [Chan et al., 2005c] and (ii) recent advances in linear and geometric programming. These strategies are discussed in more detail in Section 4. Formulations of coarsening and interpolation specific to low-power physical synthesis are considered below.

Although placement, sizing, and voltage assignment may well be considered either separately or jointly within each cluster level, the multilevel flow ensures that solution of Formulation (1) proceeds jointly over all design variables.

3.1 Coarsening

The multilevel hierarchy is built by recursive aggregation [Brandt and Ron, 2003]. The aggregation algorithm first quantifies the *affinity* each module has for its netlist neighbors. Affinities between vertices can be based on logic dependency or logic hierarchy (the logic hierarchy is preferred, if possible, for ease of verification, test, and ECO), timing constraints (modules on the critical paths will be clustered), constraints on V_{dd} , V_{th} , L_{eff} , and t_{ox} assignments (some modules may be required or preferred to have the same assignment), netlist connectivity (such as by First-Choice [Karypis, 2003] or Best-Choice [Alpert et al., 2005]), and/or geometry derived from intermediate placement results (especially in a multi-V-cycle flow [Briggs et al., 2000]). The use of different affinity functions and their impact on the final quality of result will be investigated. Initial experiments with clustering schemes will focus on the logic hierarchy and timing constraints.

Although modules in the given netlist are generally assumed to have single outputs, aggregates of modules defining coarse-level variables must be given multiple outputs in order to maintain an accurate timing view at all levels of hierarchy. “Timing views” of each cluster will be constructed, and the constraints (1) will be aggregated to correctly propagate timing data both within clusters and between clusters.

3.2 Interpolation

Interpolation may be viewed approximately as the inverse of coarsening. However, rather than simply transfer each cluster’s variables’ values to all its components, a more sophisticated mapping can be used that takes the finer-level view of the circuit into account [Chan et al., 2003]. In order for the multilevel formulation to improve quality over the traditional, sequential application of placement, voltage assignment, sizing, etc. to a flat netlist, two key conditions must be met. First, the interpolation of a solution from one level to the next must satisfy any necessary restrictions, such as timing feasibility. Second, a solver at one level of hierarchy must make good use of a solution interpolated from an adjacent level as starting point; i.e., the ability of the solver to make “warm starts,” good initial guesses of the optimal solution, is crucial. If these conditions are not met, there is no reason to expect that the multilevel solution will outperform the traditional approach. Warm starts are considered in more detail in Section 4. Finally, when the variables V_{ddi} , V_{thi} , t_{oxi} and L_{effi} are refined for cluster i , the minimum allowed cluster or region sizes due to the restrictions from manufacturing and/or the design of the power supply networks will be considered. Assignments of these variables will not be interpolated beyond their minimum allowed sizes.

4 Intralevel Optimization

Relaxation at each cluster level consists of three major steps in the following sequence.

- (i) global placement under multiple density constraints (for area, routability, heat density, etc.);
- (ii) simultaneous module sizing and assignments of V_{dd} , V_{th} , L_{eff} , and t_{ox} based on a generalized geometric-programming formulation similar to that in [Boyd et al., 2005] in order to reduce power as much as possible by making efficient use of available slacks;
- (iii) legalization for exact module locations and use of feasible discrete V_{dd} , V_{th} , L_{eff} , and t_{ox} values.

At each intermediate configuration of an iterative algorithm, the change in power with respect to changes in the separate variables can be modeled following the Zyuban and Strenski marginal-cost/sensitivity model of hardware intensity:

$$\theta(X) = \sum_i -\frac{D}{E} \frac{\frac{\partial E}{\partial x}}{\frac{\partial D}{\partial x}} \Big|_{x=X}$$

Energy-efficient design is achieved when the marginal costs of all the tuning variables are balanced [Brodersen et al., 2002].

4.1 Global Placement

Global placement has three objectives (i) minimize the total interconnect capacitance weighted netwise by switching activity; (ii) maximize the total slack for subsequent optimization 1, and (iii) enhance spatial locality for V_{dd} , V_{th} , L_{eff} , and t_{ox} assignment as well as power and clock gating. How to balance the three factors will be important part of the research. It is expected that objectives (i) and (ii) are somewhat consistent and can be combined in a weighted sum. The third objective may be expressed in terms of the additional affinity functions to be used in coarsening and placement. Sequential circuit elements may be clustered near clock-tree leaves, and nets may be weighted by switching activity [Cheon et al., 2005].

The placement engine will build on the mPL6 package of the PIs’ current SRC Task 1091.001. mPL6’s approach to placement generalizes the analytical force-directed framework of Eisenmann and Johannes in two ways [Chan et al., 2005c]. First, mPL6 incorporates force-directed placement within a multilevel-placement engine as intralevel relaxation. This approach leads to improvement in both scalability and solution quality. Second, mPL6 reformulates force-directed placement within a systematic nonlinear-programming model. This reformulation gives a systematic means of scaling density-balancing forces before combining them with the wire-length gradients and removes the need for extensive ad-hoc tuning. For further details, please see the paper on mPL5 [Chan et al., 2005c].

Poisson-based methods for density-constraint satisfaction directly applies only to density constraints formulated as equalities. The density inequalities in (1) are therefore replaced by equalities via the introduction of nonnegative artificial “density slack” variables ds_{ij} , one for each bin of a uniform grid laid over the placement region. These can be interpreted as deriving from artificial unconnected “filler” cells added to underutilized regions in order to allow the given, interconnected cells to assume non-uniform configurations.

4.2 Incorporating Multilevel Linear and Geometric Programming

Scalable, high-quality placement under generalized density constraints has already been successfully developed by the PIs in mPL6, and its extension to maximize timing slack is not expected to cause difficulty. The challenge lies in incorporating sizing, voltage assignment, and interconnect performance optimization into the multilevel flow in a way that minimizes total power. Thus, although efficient, high-quality multilevel algorithms for device sizing, voltage assignment and performance estimation are a by-product of the proposed research, the focus here is on the incorporation of such algorithms into a unified multilevel placement flow as well as on the fast solution of the individual problems themselves. Fast approximating algorithms for large-scale linear-programming (LP) and

geometric-programming (GP) formulations of sizing and voltage assignment will be investigated, and multilevel algorithms for them will be developed. Of particular interest is the ability to construct loosely-coupled convex subproblems in which the coupling itself is optimized.

Classical decomposition methods

Decomposition methods in large-scale optimization decompose a large optimization problem into smaller problems that can be solved independently, in sequence or in parallel [Bertsekas and Tsitsiklis, 1989, Lasdon, 1970]. These techniques are useful for problems that possess a structure that makes them nearly separable, and where the subproblems (at the leaf nodes) can be solved very efficiently. Primal-dual formulations lead to convex subproblems for optimizing the choice of coupled variables. These techniques will be investigated both as a means of enhancing relaxation efficiency and as a means of improving clustering.

Multilevel approach to linear and geometric programming

As an example, consider optimal gate sizing where the goal is to minimize the critical path delay by varying individual gate sizes subject to limits on sizes, power and area. This problem can be formulated as a geometric program [Boyd et al., 2005]. If the entire problem is too expensive to solve using standard geometric programming algorithms, one can attempt to first solve a simplified problem obtain by clustering nodes in the graph and treating each of these clusters of nodes as a single gate. From the solution of the coarsened problem, one obtains an approximation of the solution of the original problem. This approximation can then be refined efficiently.

A rigorous implementation will require specialized solution methods for the subproblems. For example, standard interior-point methods for linear and geometric programming are not well suited for exploiting warm starts. Cutting-plane methods [Ye, 1997] are more attractive when a series of closely related problems need to be solved. Recent techniques for solving constrained optimization problems via smooth unconstrained minimization [Nesterov, 2004, Nesterov, 2005] are also very important in this context.

Gate channel-length L_{eff} control and gate oxide thickness t_{ox} control are also formulated as generalized geometric programs [Boyd et al., 2005] via approximation by posynomials of standard expressions for their impact on capacitance and leakage.

4.3 Legalization

Global placement may produce results with partial module overlap. The overlap may further increase after module sizing. In addition, simultaneous V_{dd} , V_{th} , L_{eff} , and t_{ox} assignment using geometric programming is based on the assumption that these design variables can vary continuously within their prescribed ranges, which is not realistic. Legalization includes module overlap removal and mapping continuous design variables to discrete variables. Legalization algorithms for overlap removal in mixed-size placement have been developed by the PIs [Cong et al., 2005, Cong and Xie, 2006] under current SRC Task 1091.001 and have been shown to be very effective. They will be extended to consider a power-optimization requirement (e.g. insertion of level converters and constraints due to power supply network, etc.). Legalization of V_{dd} , V_{th} , L_{eff} , and t_{ox} assignments will lever-

age existing techniques for approximate integer programming by linear and geometric programming. Initially, randomized rounding [Raghavan and Tompson, 1987] and sensitivity-based methods [Nguyen et al., 2003] will be considered.

Complete legalization can be performed only at the finest level of the multilevel minimization flow. However, experience with large-scale mixed size placement strongly suggests that it is important to carry out “progressive” legalization at every level for macros with much larger sizes than the average cluster sizes in that level [Chan et al., 2005a, Chan et al., 2005b]. It is expected that similar techniques can be applied to legalization of V_{dd} , V_{th} , L_{eff} , and t_{ox} assignments as well. For larger macros or clusters that will not be further refined for V_{dd} , V_{th} , L_{eff} , and t_{ox} assignments, it is likely beneficial to determine their discrete legal values at intermediate levels of the multilevel optimization flow.

5 Robust Algorithm

Maximizing total slack enables aggressive power reduction by slack redistribution. But maximal total slack may be attained by allowing many paths to become near-critical. In this case, the deterministic formulation loses usefulness, and enforcing yield constraints in the statistical formulation becomes imperative.

Variations in feature dimensions and doping density cause variables d_i , l_i , and t_i to be random variables. The variations are tightly correlated due to systematic error in the manufacturing process such as lithographic model angle, doping, chemical process, etc. These correlations in the random variables will be modeled explicitly and hierarchically in the multilevel flow. The precise relationship will be investigated.

In order to solve Formulation (2), its transformation to a related deterministic form is necessary. Several techniques drawn from leading paradigms for robust algorithms will be investigated. Because slight differences in formulations for optimization under uncertainty can amount to huge differences in applicability or results, some of these differences are explained here.

The general formulation of optimization under uncertainty can be written as

$$\begin{aligned} & \text{minimize} && h(x) \\ & \text{subject to} && f_i(x, \delta_i) \leq 0, \end{aligned} \quad (\text{UCP})$$

where x is the vector of deterministic variables, and δ_i is the vector of uncertainties for constraint i . Note that $h(x)$ is a deterministic objective. If each such constraint i is required to hold for each admissible value of δ_i , then there are infinitely many deterministic constraints for each f_i , one for each fixed value of each δ_i . Typically, the δ_i can vary continuously.

Robust Convex Programming (RCP) is formulated the same as UCP above and is based on a direct, deterministic minimization of $h(x)$ subject to all infinitely many constraints over all possible values of δ_i . In some cases, depending on the set of possible values δ_i , RCP is a tractable optimization problem. For example, if the coefficients of the inequality constraints in an LP are restricted to an ellipsoid, the robust counterpart of the LP is a nonlinear convex optimization problem (a second-order cone program) [Ben-Tal and Nemirovski, 1998, Ben-Tal and Nemirovski, 2000]. Realistic ellipsoidal descriptions of the uncertainty can be derived, for example, from the error covariance and confidence ellipsoids of a least-squares estimation [Goldfarb and Iyengar, 2003a, Goldfarb and Iyengar, 2003b]. A literal interpretation of RCP

is not generally applicable to IC design, because direct enforcement of all constraints under all possible uncertainties is overly pessimistic; i.e., RCP is also not naturally suited to yield optimization. However, RCP is still a useful model (i) for understanding the often unreasonable sensitivity of a deterministic solution to constraint perturbations and (ii) as an effective heuristic technique for incorporating correlated uncertainties into a deterministic optimization model.

Probability-Constrained Programming (PCP) [Charnes and Cooper, 1959] is more directly applicable to VLSI physical synthesis:

$$\begin{aligned} & \underset{x}{\text{minimize}} && h(x) \\ & \text{subject to} && \mathbf{P}(\text{violations } f_i(x, \delta_i) > 0) < \epsilon. \end{aligned}$$

PCP is, however, extremely difficult to solve exactly for most distributions. (An important exception is an LP with normally distributed coefficients. In this case the corresponding PCP problem is again a second-order cone program.) Relaxations form an “outer (infeasible) approximation” to the solution space. Even when the original feasible (constraint) region is convex, the corresponding PCP feasible region may not be convex. PCP is the preferred yield-driven model for most IC design problems, but a unified algorithm framework for it has yet to be introduced.

Scenario-based (sampling-based) convex programming (SCP_N) [Calafiore and Campi, 2005] can be written as follows.

$$\begin{aligned} & \underset{x}{\text{minimize}} && h(x) \\ & \text{subject to} && f_i(x, \delta_{i_k}) \leq 0 \quad k = 1, \dots, N \end{aligned}$$

Deterministic optimization is done over a random sample of the infinitely many constraints. The uncertainties are sampled at random. Equivalently, each of the deterministic constraints is replicated N times, once for each randomly sampled instance of its uncertainties δ_i . If known, the specific distributions for them can be exploited, but they are not required in an explicit parameterized form. It is sufficient to be able to generate samples from the distribution, so much more general and more realistic distributions can be handled than in PCP.

The value of this method is supported by general theoretical results on the required number of samples [Calafiore and Campi, 2005]. Compared to semi-infinite optimization, note that the required number of samples N for SCP_N is *independent* of the dimension of the parameter space Δ in which $\delta \equiv (\delta_i)$ lies. In semi-infinite optimization, the required number of discretization points grows exponentially with the dimension of the parameter space Δ .

Under both PCP and SCP_N , a well-motivated heuristic [Boyd et al., 2005, Mani et al., 2005] is used to transform the probabilistic timing-yield constraint into a deterministic constraint by replacing each primitive random variable r by an estimate of the form $\mu(s) + \kappa\sigma(s)$ based on its mean $\mu(s)$ and standard deviation $\sigma(s)$. This heuristic is simple and effective. The incorporation of correlations between its random parameters will be investigated.

Although sampling significantly increases the number of constraints present in the formulation, *adaptive constraint generation* can be used to limit computation at each iteration to those constraints closest to being violated. Recent methods suitable for constraint generation include the *Analytic Centering Cutting-Plane Method* (ACCPM) [Goffin and Vial, 1999, Luo and Sun, 1998, Ye, 1997], which is popular both as a general-purpose convex optimization algorithm, and, when

combined with decomposition techniques, in distributed optimization. Other examples are the analytic centering techniques for recursive parameter estimation in signal processing and control [Bai et al., 1999, Bai et al., 2000]. Sequential analytic centering methods can be analyzed rigorously using the techniques developed for interior-point methods (in particular, the convergence analysis of Newton’s method for logarithmic barrier functions) [Nesterov and Nemirovsky, 1994, Ye, 1997].

Modeling the uncertainty of the parameters in an optimization model involves a trade-off between tractability and accuracy. It is critical to use realistic models (e.g., models incorporating correlations) which can be handled efficiently. For example, when estimating a covariance matrix from the sample covariances, it is important to impose structural restrictions that make the resulting robust optimization problems easier to solve. Of particular interest are low-rank structure, bandedness, sparsity and sparsity in the inverse [Dempster, 1972].

Technology Transfer

Anticipated results of the proposed research include technical reports, published papers in major EDA conferences and journals, and a software prototype of a novel physical synthesis flow for power optimization under process variation.

The PIs’ research group has a strong track record in delivering results and transferring technology to SRC member companies. Their early work on interconnect optimization (TRIO: <http://cadlab.cs.ucla.edu/~trio>) and estimation (IPEM: http://cadlab.cs.ucla.edu/software_release/ipem/htdocs) has been used and customized by multiple SRC companies, including Intel and IBM. The group’s PEKO placement sub-optimality benchmarks [Chang et al., 2004, Goering, 2003c] have been downloaded by over 310 different individuals at EDA companies and research universities and have been covered by multiple EE Times articles [Goering, 2003c, Goering, 2003b, Goering, 2003a]. Their SRC placement project, Task 1091.001 ending June 2006, promised a Moore’s-law generation 30% reduction of wirelength and has already exceeded that goal considerably [Chan et al., 2005c]. The resulting mPL placement package [Chan et al., 2005c] has over 250 downloads, including many from SRC member companies such as Cadence, Intel and Mentor Graphics.

References

References and PI bios are available at <http://cadlab.cs.ucla.edu/cpmo/wp06refs.html> .

References

- [Alpert et al., 2005] Alpert, C., Nam, A. B. K. G.-J., Reda, S., and Villarrubia, P. (Apr 2005). A semi-persistent clustering technique for VLSI circuit placement. In *Proc. Int'l Symp. on Physical Design*, pages 200–207.
- [Bai et al., 2000] Bai, E., Fu, M., Tempo, R., and Ye, Y. (2000). Convergence results of the analytic center estimator. *IEEE Transactions on Automatic Control*, 45(3):569–572.
- [Bai et al., 1999] Bai, E., Ye, Y., and Tempo, R. (1999). Bounded error parameter estimation: a sequential analytic center approach. *IEEE Transactions on Automatic Control*, 44(6):1107–1117.
- [Ben-Tal and Nemirovski, 1998] Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, 23:769–805.
- [Ben-Tal and Nemirovski, 2000] Ben-Tal, A. and Nemirovski, A. (2000). Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming Series A*, 88:411–424.
- [Bertsekas and Tsitsiklis, 1989] Bertsekas, D. P. and Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation*. Prentice-Hall, Englewood Cliffs, New Jersey.
- [Boyd et al., 2005] Boyd, S. P., Kim, S.-J., Patil, D. D., and Horowitz, M. A. (2005). Digital circuit optimization via geometric programming. *Operations Research*, 53(6):899–932.
- [Brandt, 1986] Brandt, A. (1986). Algebraic multigrid theory: The symmetric case. *Appl. Math. Comp.*, 19:23–56.
- [Brandt and Ron, 2003] Brandt, A. and Ron, D. (2003). Multigrid solvers and multilevel optimization strategies. In Cong, J. and Shinnerl, J., editors, *Multilevel Optimization and VLSICAD*. Kluwer Academic Publishers, Boston.
- [Briggs et al., 2000] Briggs, W., Henson, V., and McCormick, S. (2000). *A Multigrid Tutorial*. SIAM, Philadelphia, second edition.
- [Brodersen et al., 2002] Brodersen, R., Horowitz, M., Markovic, D., Nikolic, B., and Stojanovic, V. (2002). Methods for true power minimization. In *Proc. Int'l Conf. on Computer-Aided Design*, pages 35–42.
- [Calafiore and Campi, 2005] Calafiore, G. and Campi, M. C. (2005). Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming Series A*, 102(1):25–46.
- [Chan et al., 2003] Chan, T., Cong, J., Kong, T., Shinnerl, J., and Sze, K. (Nov 2003). An enhanced multilevel algorithm for circuit placement. In *Proc. Int'l Conf. on Computer-Aided Design*, San Jose, CA.
- [Chan et al., 2005a] Chan, T., Cong, J., Romesis, M., Shinnerl, J., Sze, K., and Xie, M. (2005a). Enhanced robustness in multilevel mixed-size placement. Technical report, Computer Science Dept., University of California, Los Angeles. SRC Pub. P012921.
- [Chan et al., 2005b] Chan, T., Cong, J., Shinnerl, J., Sze, K., and Xie, M. (2005b). Highly scalable placement by multilevel optimization. Src task 1091.001 annual report, Computer Science Dept., University of California, Los Angeles.
- [Chan et al., 2005c] Chan, T., Cong, J., and Sze, K. (2005c). Multilevel generalized force-directed method for circuit placement. In *Proc. Int'l Symp. on Physical Design*.
- [Chang et al., 2004] Chang, C., Cong, J., Romesis, M., and Xie, M. (2004). Optimality and scalability study of existing placement algorithms. *IEEE Trans. on Comp.-Aided Design of Integrated Circuits and Sys.*, pages 537–549.
- [Charnes and Cooper, 1959] Charnes, A. and Cooper, W. (1959). Chance constrained programming. *Management Science*, 6(1).
- [Cheon et al., 2005] Cheon, Y., Ho, P.-H., Kahng, A., Reda, S., and Wang, Q. (2005). Power-aware placement. In *Proc. Design Automation Conf.*, pages 795–800.
- [Clarke, 2004] Clarke, P. (2004). Scaling died at 130-nm so innovate, says ibm cto. *EE Times*. <http://www.eetimes.com/news/semi/showArticle.jhtml;jsessionid=04C51F2VRZD54QSNDBECKICCJUMKJVN?articleID=19400132>.
- [Cong and Pan, 2001] Cong, J. and Pan, D. (2001). Interconnect performance estimation models for design planning. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 20(6):739–752.
- [Cong et al., 2005] Cong, J., Romesis, M., and Shinnerl, J. (Nov. 2005). Robust mixed-size placement under tight white-space constraints. In *Proc. Int'l Conf. on Computer-Aided Design*, pages 165–172.
- [Cong and Shinnerl, 2003] Cong, J. and Shinnerl, J., editors (2003). *Multilevel Optimization in VLSICAD*. Kluwer Academic Publishers, Boston.
- [Cong and Xie, 2006] Cong, J. and Xie, M. (2006). A robust detailed placement for mixed-size IC designs. In *Proc. Asia South Pacific Design Automation Conf.*, pages 188–194.
- [Dempster, 1972] Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–175.
- [Eisenmann and Johannes, 1998] Eisenmann, H. and Johannes, F. (1998). Generic global placement and floorplanning. In *Proc. 35th ACM/IEEE Design Automation Conference*, pages 269–274.
- [Gelsinger, 2004] Gelsinger, P. (2004). Gigascale integration for teraops performance — challenges, opportunities, and new frontiers. In *Proc. Design Automation Conf.* keynote address.
- [Goering, 2003a] Goering, R. (2003a). FPGA placement performs poorly, study says. *EE Times*. <http://www.eedesign.com/story/OEG20031113S0048>.
- [Goering, 2003b] Goering, R. (2003b). IC placement benchmarks needed, researchers say. *EE Times*. <http://www.eedesign.com/story/OEG20030410S0029>.
- [Goering, 2003c] Goering, R. (2003c). Placement tools criticized for hampering IC designs. *EE Times*. <http://www.eedesign.com/story/OEG20030205S0014>.
- [Goffin and Vial, 1999] Goffin, J.-L. and Vial, J.-P. (1999). Shallow, deep and very deep cuts in the analytic center cutting plane method. *Mathematical Programming*, 84(1):89–103.
- [Goldfarb and Iyengar, 2003a] Goldfarb, D. and Iyengar, G. (2003a). Robust convex quadratically constrained programs. *Mathematical Programming Series B*, 97:495–515.
- [Goldfarb and Iyengar, 2003b] Goldfarb, D. and Iyengar, G. (2003b). Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38.

- [Kao et al., 2002] Kao, J., Narendra, S., and Chandrakasan, A. (2002). Subthreshold leakage modeling and reduction techniques. In *Proc. Int'l Conf. on Computer-Aided Design*, pages 141–148.
- [Karypis, 2003] Karypis, G. (2003). Multilevel hypergraph partitioning. In Cong, J. and Shinnerl, J., editors, *Multilevel Optimization and VLSICAD*. Kluwer Academic Publishers, Boston.
- [Lasdon, 1970] Lasdon, L. S. (1970). *Optimization Theory for Large Systems*. MacMillan.
- [Luo and Sun, 1998] Luo, Z.-Q. and Sun, J. (1998). An analytic center based column generation algorithm for convex quadratic feasibility problems. *SIAM J. on Optimization*, 9:217–235.
- [Mani et al., 2005] Mani, M., Devgan, A., and Orshansky, M. (2005). An efficient algorithm for statistical minimization of total power under timing yield constraints. In *Proc. Design Automation Conf.*, pages 309–314.
- [Nesterov, 2004] Nesterov, Y. (2004). Rounding of convex sets and efficient gradient methods for linear programming problems. Technical report, core discussion paper, Université catholique de Louvain.
- [Nesterov, 2005] Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming Series A*, 103:127–152.
- [Nesterov and Nemirovsky, 1994] Nesterov, Y. and Nemirovsky, A. (1994). *Interior-point polynomial methods in convex programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA.
- [Nguyen et al., 2003] Nguyen, D., Davare, A., Orshansky, M., Chinnery, D., Thompson, B., and Keutzer, K. (2003). Minimization of dynamic and static power through joint assignment of threshold voltages and sizing optimization. In *Proceedings of the 2003 international symposium on Low power electronics and design*, pages 158–163. ACM Press.
- [Puri et al., 2003] Puri, R., Stok, L., Cohn, J. M., Kung, D. S., Pan, D., Sylvester, D., Srivastava, A., and Kulkarni, S. (2003). Pushing ASIC performance in a power envelope. In *Proc. Design Automation Conf.*, pages 788–793.
- [Raghavan and Tompson, 1987] Raghavan, P. and Tompson, C. (1987). Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374.
- [Sarrafzadeh et al., 2002] Sarrafzadeh, M., Wang, M., and Yang, X. (2002). *Modern Placement Techniques*. Kluwer, Boston.
- [Weste and Harris, 2005] Weste, N. and Harris, D. (2005). *Principles of CMOS VLSI design: a circuits and systems perspective*. Addison-Wesley/Pearson, New York, NY, USA.
- [Ye, 1997] Ye, Y. (1997). *Interior Point Algorithms: Theory and Analysis*. Discrete Mathematics and Optimization. Wiley, New York.