The Virtual Wafer Fab Technology for the Deep-Submicron ULSI Era

Xing Zhou

School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore

Abstract

This article describes a new technology, the virtual wafer fab technology, which has become increasingly popular in the semiconductor industry for the deep-submicron era. The first section, ***Introduction****, which is taken from the author's lecture notes for the 4th-year design course, is intended for readers who are not familiar with the field. The second section is on the motivation and ideas of the* ***Virtual Wafer Fab*** *and some related concepts. The last section describes an on-going joint project with the local wafer fab,* ***Project DOUST****, its goals, ideas, and potential impact.*

Contents

INTRODUCTION

Have you ever imagined this: Whenever you push the button of an electronic system (e.g., the keyboard of a computer, or the remote control of a VCR), you are actually manipulating the motion of individual electrons inside the system.

Since the invention of the transistor 50 years ago and the advent of the modern VLSI technology, we have witnessed the dramatic advancement in the microelectronics industry. It is all based on the foundation of the semiconductor industry for our ability to make electrons move faster, transistors smaller, and more transistors on a single chip. As this trend leads us into the ultra-large-scale integration (ULSI) era, the deep-submicron technology calls for new design methodologies and new electronic design automation (EDA) tools to cope with the complexity and coupling among different stages of a design.

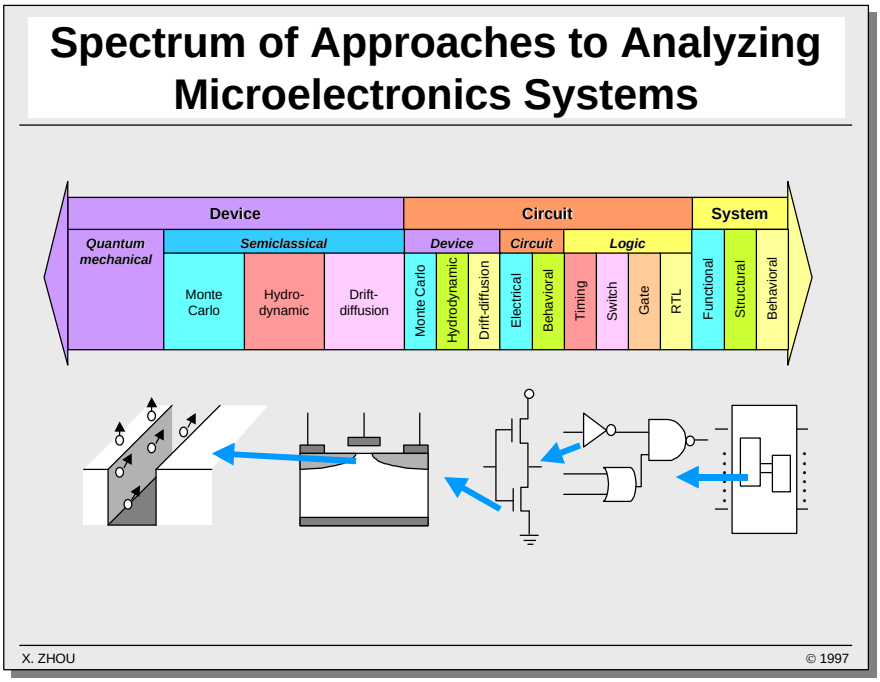
Spectrum of Approaches to Analyzing Microelectronic Systems

There exists a wide spectrum of approaches to analyzing microelectronic systems, which can be largely classified into three levels: *system*, *circuit*, and *device*, with the corresponding three walks of people: system designers, circuit engineers, and device physicists.

At the "engineering end" of the spectrum, system designers are concerned with the behavior and functionality of a complex system. In general, a top-down synthesis approach is employed in a system design. Design methodologies evolve with the rapid advancement of the technology as well as design tools. New approaches, such as semi-custom design (standard cell, gate array, field-programmable gate array, etc.) for the application-specific integrated circuits (ASIC), have been widely used. Nowadays, it is totally unimaginable that a Pentium chip can be "designed" bottom-up without the use of advanced EDA tools.

Traditionally, integrated circuit design has been centered at the *de facto* industry standard — the Berkeley SPICE circuit simulator. Higher-level logic simulators (such as switch-level timing analyzers, gate-level or register-transistor-level simulators) are aimed at increased simulation speed at the expense of accuracy. On the other hand, circuit designers expect a set of SPICE parameters to be extracted from the lower layout or technology level for use in circuit simulation as well as logic-level verification.

For ultra-small-size transistors or non-steady-state transport problems, the closed-form macromodels at the circuit level will not be sufficient or valid for the analysis of the problem at hand. Device electrical performance is characterized by 2D/3D numerical solutions of the coupled Poisson and current continuity equations, together with the drift-diffusion (DD) or energy-balance (EB) equations. When highly nonequilibrium and nonlocal effects are of importance, Monte Carlo (MC) technique is used to solve for the exact solution of the Boltzmann transport equation. Finally, at the "physics end" of the spectrum, many attempts have been made to formulate a rigorous theory of quantum transport.



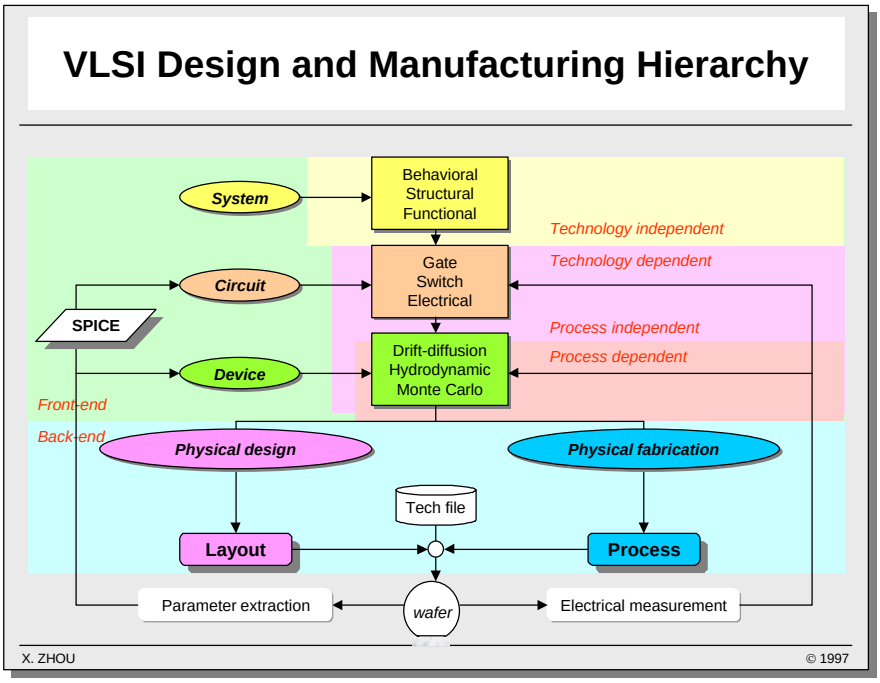
VLSI Design and Manufacturing Hierarchy

A microelectronic system, from concept to product, must go through the process of design and implementation. Before the deep-submicron era, a full-custom design (as opposed to other semi-custom ASICs) mainly consists of two sessions: *frontend* and *backend*.

In frontend design, it usually starts from the *system level* with a top-down synthesis approach, which is technology independent (i.e., at this level, it is "irrelevant" whether the design will be implemented in CMOS or bipolar technology). The design is then transformed into the *circuit level*, in which the logic functionality, timing delays, speed and power, etc., are the primary concerns. This level is technology dependent but relatively process independent. If more detailed study on the transistor performance is needed, it can be supplemented at the *device level* based on the device physics which, in general, requires process information.

At the backend, the final design must be translated into the physical layout representation, which is to be used to implement in wafer fabrication. In the "conventional" hierarchy, technology development (manufacturing) is relatively independent of the design. The "feedback" only occurs at the circuit level where the fab provides the circuit designer a set of SPICE parameters for the particular process through electrical measurement and parameter extraction of the fabricated transistors.

From the described VLSI design and manufacturing hierarchy, two streams of knowledge can be identified: vertically, a single design is represented at different levels of abstraction; horizontally, any design (in its final layout format) must be combined with a process recipe to be implemented, step by step, on a silicon chip.



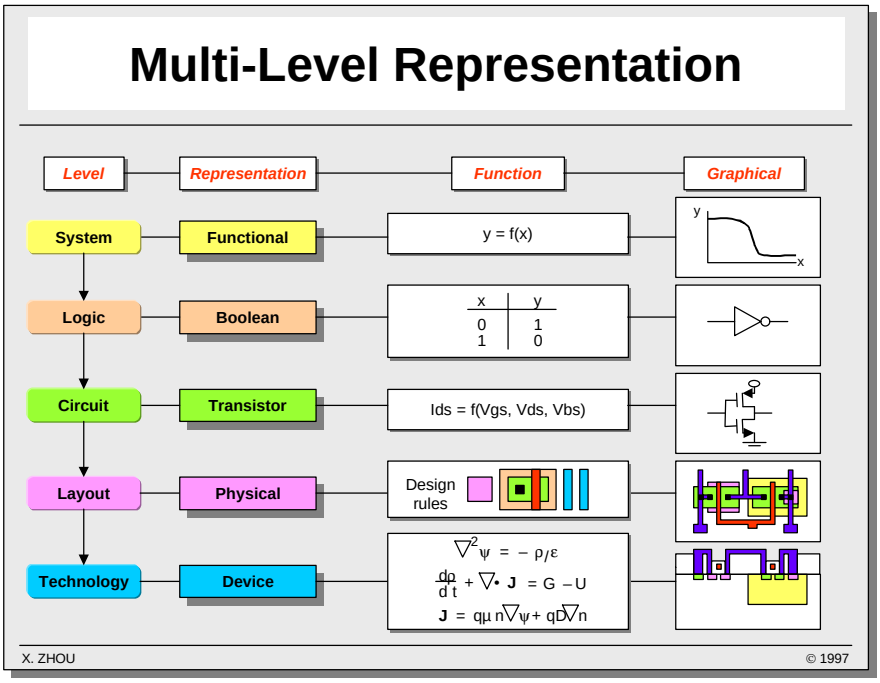
Multi-Level Representation

The design and implementation of an integrated circuit, be it the whole system on a chip or a single inverter, can be represented at five distinctively different levels of abstraction (take the inverter as an example):

- **System level**: *functional representation* where the behavior of the system is described by its transfer function;
- **Logic level**: *Boolean representation* where the system is represented by its logic function plus some limited modeling of the propagation delays;
- **Circuit level**: *transistor representation* where the system is characterized by closed-form nonlinear equations for each transistor I–V characteristics;
- **Layout level**: *physical representation* where the system is drawn by its equivalent geometric patterns subject to the given design rules;
- **Technology level**: *device representation* where the system is simulated by a set of physical differential equations on the cross-sectional plane of its final structure.

You may ask why there are so many levels of representations for the same design. The answer is that for a complex problem, a "divide-and-conquer" approach should be used. Since an integrated circuit is designed and implemented in a hierarchical way, it should also be analyzed and simulated at different levels of abstraction. The essence of modeling is what Albert Einstein described:

"*Everything should be made as simple as possible, but not any simpler.*"



Layout + Process = Chip

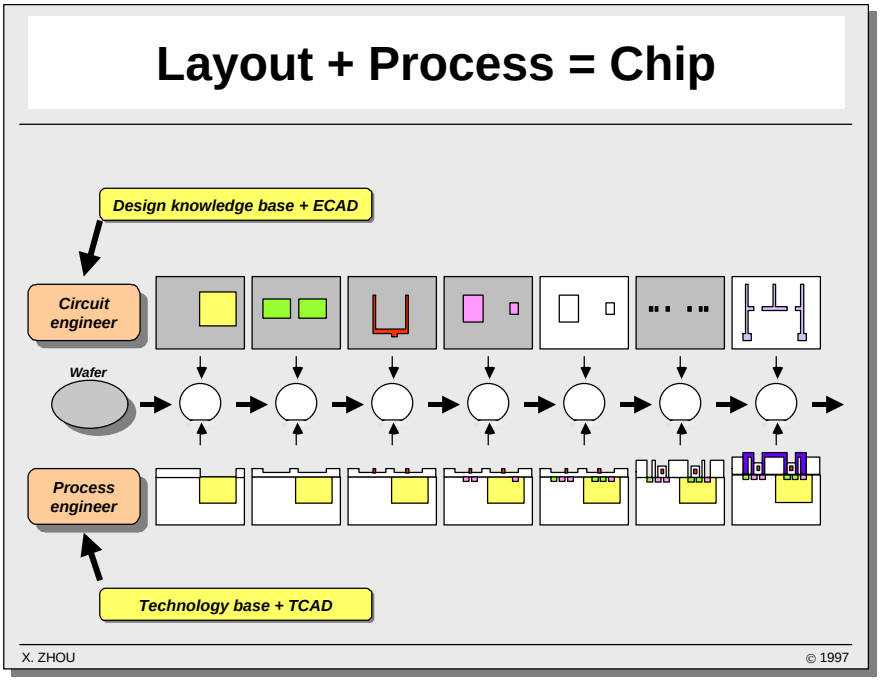
From another point of view, the final physical design (layout) of the system must be implemented into the silicon chip (or "committed to silicon"). This process involves the application of each individual mask during the processing of the silicon wafer. Again, take the inverter as an example, the final completion of the "device" consists of a sequence of processing steps, such as diffusion, oxidation, and ion implantation, for a given technology. The final performance of the system (the "device") not only depends on the physical design (layout), but also on the process in which the device structures and doping profiles are altered. Imagine that for the whole system, this combination of layout and process must work "laterally" for all layers of the masks as well as "vertically" for all layers of depths.

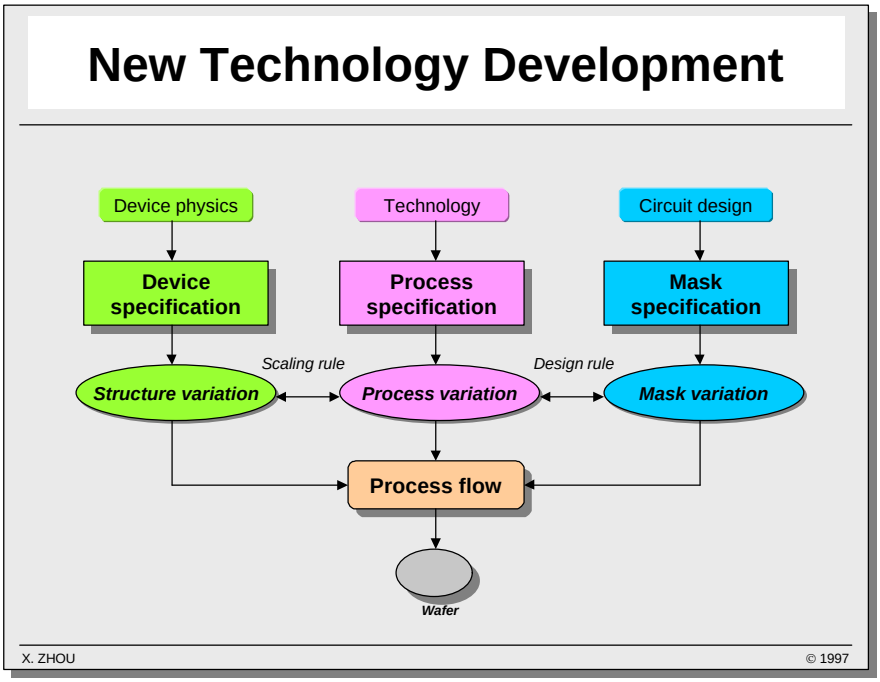
New Technology Development

A new technology development is the process of achieving the optimum device performance through process variations. Generally, it involves three streams of knowledge: device physics, processing technology, and circuit design. It is centered at a specified process recipe since the final target is a process flow to implement the optimum design into the silicon chip. However, process variation must be constrained by the design rule and the scaling rule. The transistor electrical characteristics are closely coupled with the doping profiles and layer structures, and very often, trade-off must be taken for different design targets. Processing technology, such as the self-aligned technology, is also linked to the mask design and limited by photolithography.

Traditionally, technology development is relatively independent of the circuit design. However, when going into the deep-submicron regime, the conventional scaling rule does not apply any more. Device performance is closely coupled with the fabrication parameters and

circuit design constraints. Circuit designers also need process information (such as interconnect delays) at the very early stage of the design. Many of the conventional scaling rules and design methodologies are changed.





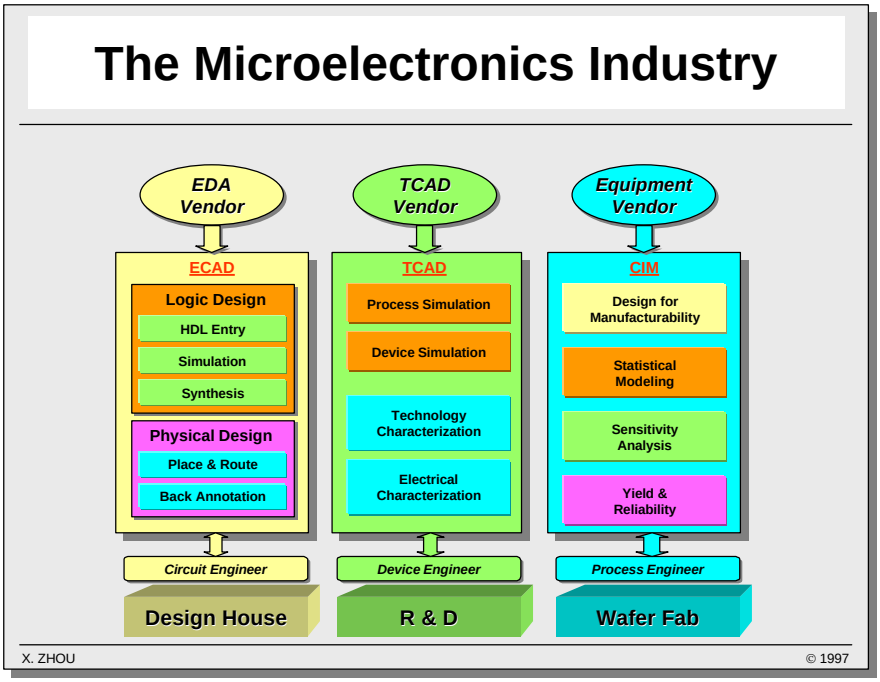
The Microelectronics Industry

Traditionally, the microelectronics industry consists of two major sectors: the *design house* where chips are being designed by system/circuit engineers, and the *wafer fab* where chips

are being fabricated by process/device engineers. In a design house, extensive electronic computer-aided design (ECAD) tools supported by the EDA vendors are used. A design session generally involves an iterative process from logic design (frontend) to physical design (backend). In a wafer fab, expensive equipment from the equipment vendors are used to fabricate the chips. The primary concerns are things like design for manufacturability, statistical modeling, sensitivity analysis, yield and reliability, all supported by computer integrated manufacturing (CIM).

The link between the design house and the wafer fab has been the mask information (GDS II files) and a set of SPICE parameters. The design house provides the mask information for the wafer fab, and the fab provides the SPICE parameters for the particular process to the design house. However, the SPICE parameters, which are needed in the circuit design, will not be available before the device is fabricated. For deep-submicron technologies, this information and, in particular, the interconnect delay information, is critical at the very early stage of a design (before fabrication).

With the rapid advancement of process and device models, technology computer-aided design (TCAD) tools are developed by the TCAD vendors to emulate wafer fabrication and device characterization. Realistic process information (layer thickness and doping profiles) can be obtained from process simulation, and electrical information (I–V and C–V characteristics) can be obtained from 2D device simulation. Then, SPICE parameters can be obtained through parameter extraction based on the simulated electrical characteristics, and delay information can be obtained through technology characterization. So far, the TCAD approach to technology development and transistor design has been widely used by all semiconductor companies, but it is mainly restricted to the device engineer and R&D community.



VIRTUAL WAFER FAB

One of the challenges of the deep-submicron technology is due to the fact that as transistor dimensions are getting smaller, their performance is more closely coupled with the fabrication parameters, and the interconnect delay is becoming dominant. This means that the transistor structure and processing information need to be considered in the early stage of a design before it is fabricated. This new trend places demands for new design methodologies as well as new design tools.

Chip Design and Wafer Fabrication

The semiconductor industry or, to a larger extend, the microelectronics industry, involves chip design and wafer fabrication, which is a complex, iterative process of "design – manufacturing – characterization – simulation – verification."

A chip design starts with the product specification, followed by the frontend and backend designs. In this phase, ECAD tools have been developed so powerful that the logic design can be synthesized from a high-level hardware description language (HDL), the circuit netlist can be extracted from the logic functional description, and the layout can be extracted from the circuit- and logic-level descriptions.

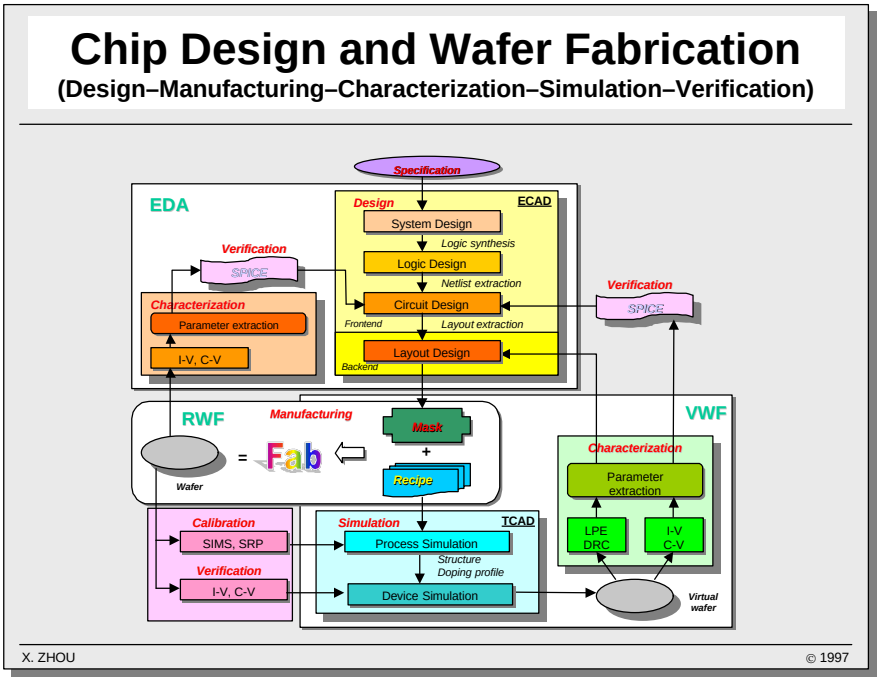
Once a set of mask has been designed, it is combined with a given process recipe in the manufacturing phase in a "real wafer fab" (RWF). Electrical and technological characterization is then performed on the fabricated device to extract the parameters for back-annotation and verification.

Although the EDA tools (including design and verification tools) are already quite advanced, this "design – manufacturing – characterization – verification" loop can be very costly if a "first-time silicon success" cannot be achieved. Moreover, it would be extremely expensive and time consuming if a new technology is to be developed using this iterative experimentation.

With the maturity of TCAD tools, real wafer fabrication can be emulated by process simulation, from which realistic device structures and doping profiles can be generated, and transistor performance can be characterized through device simulation with reasonable accuracy. Interconnect delays can also be extracted through technology characterization with 3D accuracy, which can provide information for design rule checker (DRC) and layout parasitic extraction (LPE) tools in the physical design. SPICE parameters can also be extracted from the "virtual device" I–V characteristics for back-annotating circuit simulators and timing analyzers, which provides the notion of "calibrating" ECAD tools based on TCAD tools. It is obvious that there is great incentive to develop this "virtual wafer fab" (VWF) technology to supplement the RWF experimentation.

Of course, how effective this approach will be in aiding first-time silicon success depends on how well the process and device simulators are calibrated to the RWF results. This calibration involves another loop — "manufacturing – calibration – simulation – verification." The process models must be calibrated to the experimental doping and carrier

profiles obtained from secondary ion mass spectroscopy (SIMS) and spreading resistance profiling (SRP); and the device models must be verified with the measured I–V and C–V characteristics.



Virtual Wafer Fab

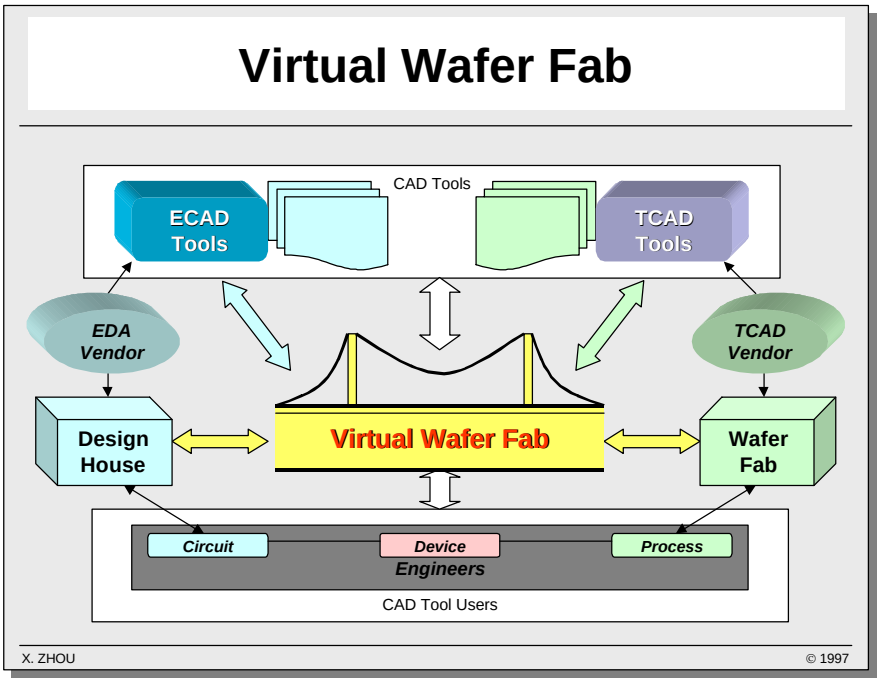
EDA vendors provide ECAD tools for the design house and TCAD vendors provide TCAD tools for the wafer fab. These tools are used by the designers (circuit, device, and process engineers), or collectively, the CAD tool users. The "lost link" between the design house and the wafer fab has not been significant before the deep-submicron era.

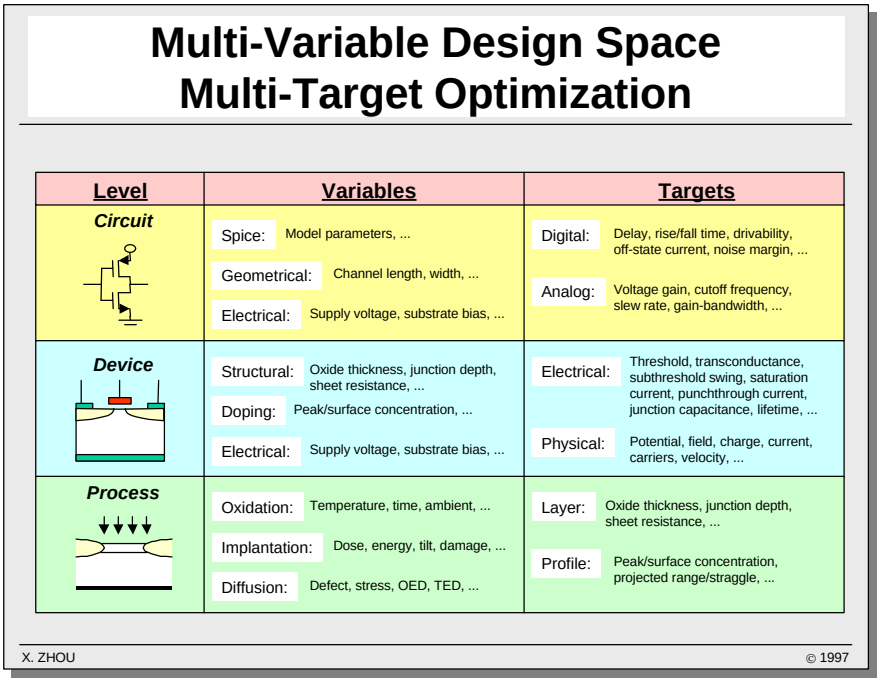
The notion of "virtual wafer fab" is to bridge the gap between the two. This "virtual wafer fab" technology will be proven to be increasingly vital, if not a must, in the deep-submicron ULSI era. Basically, it serves as a bridge between the design house and the wafer fab, between the ECAD and TCAD tools, and between the CAD tools and CAD tool users in general. It will have a major impact on the way we develop a new technology. It will eventually change the way we design in the same way as the EDA has shaped the whole microelectronics industry.

Multi-Variable Design Space, Multi-Target Optimization

The "virtual wafer fab" technology is based on the use of TCAD tools centered at the device level. The problem is very complex in the sense that it has a multi-variable design space and it is a multi-target optimization problem. Just take a process simulator as an example. For a particular process, there may be tens of different models to choose from, hundreds of parameters to adjust, and thousands of factors to influence the final outcome. In this sense, the question is not *whether* the tools can be used or not, it is *how* they should be used. TCAD

should not be "tweaking coefficients for all days." It should be used for "testing concepts by analysis and design."





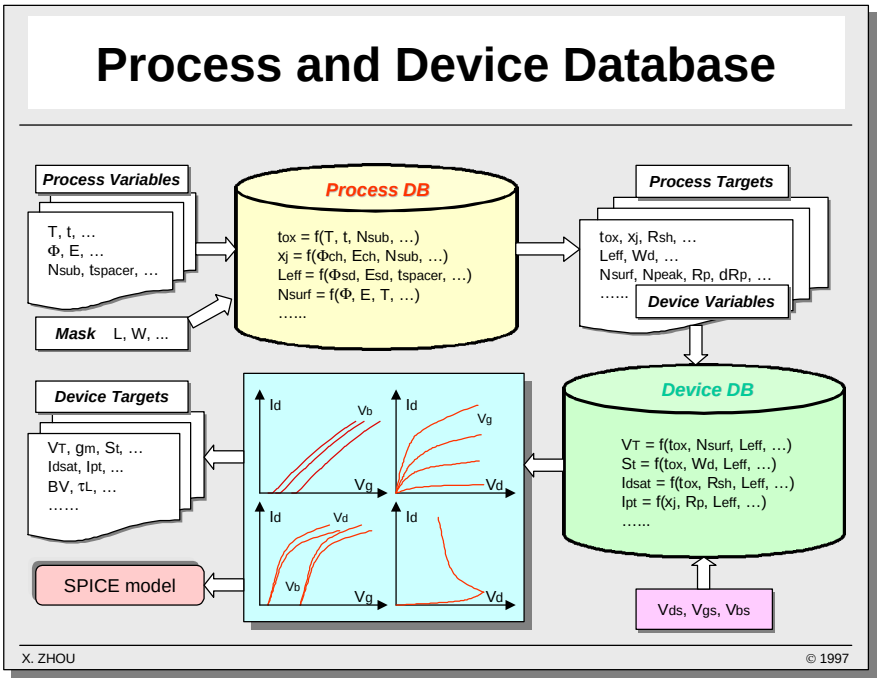
The multi-target optimization problem should be decomposed at different levels: process, device, and circuit. The target values of a specified gate delay, threshold voltage, or subthreshold swing, etc. depend on different variables at different levels. Transistor characteristics are more related to device parameters such as gate oxide thickness, junction depths, effective channel length, etc., which are indirectly related to processing variables such as oxidation time and temperature, implant dose and energy, etc. It is important to break

these dependencies at different levels so that they can be tackled more accurately and efficiently.

Process and Device Database

The first step is to identify target parameters to be designed or optimized. The specification will be different for different applications. For example, threshold voltage is one of the most important parameters in CMOS circuits, which is influenced by many variables. Any process variations, such as gate oxidation time or temperature, threshold adjustment implant dose or energy, will have direct impact on the final threshold voltage. However, from the device modeling point of view, it is the gate oxide thickness and the channel doping profile which influence the device electrical characteristics.

Hence, it is advantageous to construct separate databases for the processes and devices. For the process database, the input will be the processing variables and mask information, and the output will be the layer structures and doping profiles. The database stores all the input–output dependency data and, if possible, compact models linking the input and output can be developed. The device database, on the other hand, takes the device structure and biasing condition as input and stores the generated I–V characteristics, from which device targets can be extracted. Again, it would be extremely useful and efficient if compact analytic models at the device level can be developed based on the detailed TCAD results. Finally, circuit-level parameters and interconnect delays can also be extracted from the "virtual device" I–V characteristics.



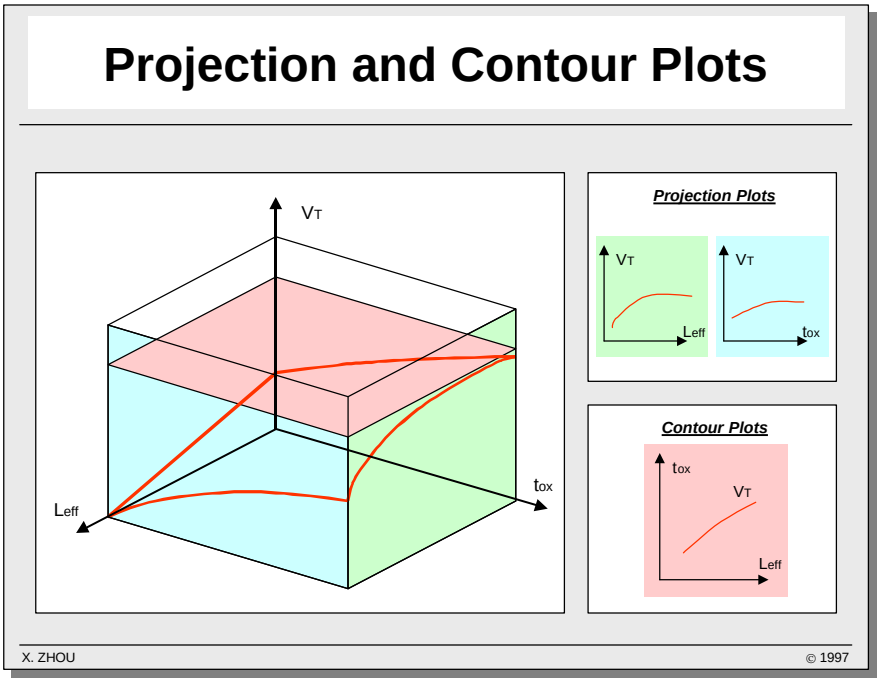
Projection and Contour Plots

Once a complete process and device database is constructed, a specified set of multi-target objects can be readily accessible from the multi-variable design space. These targets can be

visualized graphically by 3D plots, e.g., a plot of the threshold voltage (V_t) as a function of the effective channel length (L_{eff}) and the gate oxide thickness (t_{ox}).

These 3D data can be viewed in different ways at the designer's discretion. One type is the *projection plot* in which the target is plotted against one major variable, optionally with a secondary variable as a parameter. The other type is the *contour plot* in which constant values of the target are plotted against two major variables.

These plots can be generated directly from the database. If compact models are developed, it would be extremely efficient to generate these plots, which can be used for multi-target optimization.



Conceptual Database

A conceptual database is described as one which stores the target–variable dependency data in a form easily accessible to the user. Each target in the database can be related to a number of variables. These data are generated from a complete set of simulations based on the TCAD models, which includes all the nonuniformities in chemical/electrical distributions and the nonlinearities of carrier transport. Each target–variable dependency represents a particular design of experiment (DOE). If a compact model is developed for a particular dependency, an analytic functional relation for that dependency is also defined.

Conceptually, data can be manipulated to obtain any view of a specified target. For example, projection plots of the *m*th target as a function of variable *i* for different values of variable *j* can be generated at the choice (*m*; *i*, *j*), and the *n*th target can also be plotted in the same way for comparison. Likewise, by specifying the target values for the *m*th and *n*th targets, contour plots can be obtained as a function of two variables *i* and *j*. These plots and data can be used

for nonlinear regression, response surface modeling (RSM), and simultaneous optimization of multiple targets.