The IC Processes of the Future
Advances in Device Structures and Device Simulation Techniques

Siegfried Selberherr
and
Claus Fischer, Stefan Halama, Christoph Pichler, Gerhard Rieger, Gerhard Schrom, Thomas Simlinger

Institute for Microelectronics
Technische Universität Wien
AUSTRIA

Abstract
The state of the art of Technology Computer-Aided Design (TCAD) frameworks for the development of new device structures and their fabrication processes is discussed. A framework must contain support for both integration of existing simulation programs and development of new tools. The complexity and scope of a rigorous TCAD framework requires special effort to create a system which is comprehensible for both users and programmers. The implementation of the task level and user interface of the Viennese Integrated System for Technology CAD Applications (VISTA) is presented, and some important aspects of the integration of independently developed tools are discussed. Particular examples highlight the capabilities of present tools. A \(0.25 \mu m\) fully-planarized CMOS process with shallow trench isolation is studied. The interaction of fairly heterogeneous tools within one common environment becomes thereby transparent. As a second example, the feasibility and the limitations of ultra-low-power CMOS technologies are investigated by simulation. A set of possible ultra-low-power CMOS processes is developed and analyzed for their performance on the gate level. To demonstrate capabilities of new device structures and materials, an analysis of a High Electron Mobility Transistor (HEMT) is presented.

1 Introduction
Process and device simulation is commonly used for the design of new VLSI technologies. Simulation programs serve as exploratory tools in order to gain better understanding of process and device physics. On the other hand, simulations are also carried out after the design phase to optimize certain parameters of a technology, e.g., to improve device performance and reliability or to increase the yield, e.g. [7].

For all these tasks the term TCAD, short for technology computer-aided design, was coined. TCAD includes both physically rigorous as well as simplified process and device simulation in one to three spatial dimensions. Furthermore, links to layout-oriented CAD and circuit simulation are required.

Depending on the particular application of TCAD tools, different demands arise. On the one hand, the development of new technologies and the prediction of the behavior of new devices
Table 1: List of TCAD systems and tool integration approaches, in alphabetical order of institutions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Status</th>
<th>Data level</th>
<th>Task level</th>
<th>Presentation Level</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>MECCA</td>
<td>AT&amp;T</td>
<td>I</td>
<td>awk/sed, C++</td>
<td>UNIX Shell, Tcl</td>
<td>Tk</td>
<td>[38] [39]</td>
</tr>
<tr>
<td>PREDITOR</td>
<td>CMU</td>
<td>E</td>
<td>CDB/HGDB</td>
<td>Tcl</td>
<td>Tk, Motif</td>
<td>[68] [69]</td>
</tr>
<tr>
<td>pdFab</td>
<td>“Superviser”</td>
<td>E</td>
<td>C</td>
<td>“Superviser”</td>
<td></td>
<td>[43]</td>
</tr>
<tr>
<td>VATS</td>
<td>IBM</td>
<td>I</td>
<td>Converter</td>
<td>“Superviser”</td>
<td></td>
<td>[36]</td>
</tr>
<tr>
<td>EASE</td>
<td>Intel</td>
<td>I</td>
<td>PIIF derivative</td>
<td>UNIX Shell, FASST/TEL</td>
<td>Motif</td>
<td>[13] [41]</td>
</tr>
<tr>
<td>CAFE</td>
<td>MIT</td>
<td>I</td>
<td>BPIF/Gestalt</td>
<td>MITPF</td>
<td></td>
<td>[5]</td>
</tr>
<tr>
<td>P&amp;D Workbench</td>
<td>NEC</td>
<td>I</td>
<td>Converter</td>
<td>ESCORT + MEDLEY</td>
<td></td>
<td>[66]</td>
</tr>
<tr>
<td>UNISAS</td>
<td>Oki</td>
<td>I</td>
<td>GCOS</td>
<td>UNICOLL</td>
<td></td>
<td>[47]</td>
</tr>
<tr>
<td>IDDE</td>
<td>Philips UK</td>
<td>IP</td>
<td>ASCII PIIF</td>
<td>DeckBuild</td>
<td>Apollo DIALOGUE</td>
<td>[23] [24]</td>
</tr>
<tr>
<td>PRIDE</td>
<td>Philips USA</td>
<td>IP</td>
<td>ASCII PIIF</td>
<td>none</td>
<td>SunView</td>
<td>[61]</td>
</tr>
<tr>
<td>SATURN</td>
<td>Siemens</td>
<td>IP</td>
<td>SATURN</td>
<td>none</td>
<td></td>
<td>[31] [32]</td>
</tr>
<tr>
<td>MASTER</td>
<td>Silvaco</td>
<td>C</td>
<td>C</td>
<td>none</td>
<td>DeckBuild</td>
<td>[29]</td>
</tr>
<tr>
<td>STORM</td>
<td>ESPRIT</td>
<td>P</td>
<td>DAMSEL</td>
<td>none</td>
<td>STORM UI/IDAS</td>
<td>[40]</td>
</tr>
<tr>
<td>Alladin, SEWB</td>
<td>Stanford University</td>
<td>R</td>
<td>SWR</td>
<td>none</td>
<td></td>
<td>[12]</td>
</tr>
<tr>
<td>CAESAR</td>
<td>TMA</td>
<td>C</td>
<td>TIF</td>
<td>Module/Step</td>
<td>CAESAR</td>
<td>[3]</td>
</tr>
<tr>
<td>ITS</td>
<td>Texas Instruments</td>
<td>IP</td>
<td>none</td>
<td>none</td>
<td>Motif</td>
<td>[22]</td>
</tr>
<tr>
<td>“System controller”</td>
<td>Toshiba</td>
<td>IP</td>
<td>ASCII topography</td>
<td>none</td>
<td></td>
<td>[34]</td>
</tr>
<tr>
<td>VISTA</td>
<td>TU Vienna</td>
<td>P</td>
<td>PA/PIF</td>
<td>XW.P + VUI</td>
<td></td>
<td>[14] [25]</td>
</tr>
<tr>
<td>PROSE</td>
<td>UC Berkeley</td>
<td>I</td>
<td>PIIF, SWR</td>
<td>Tcl</td>
<td>Tk + VEM</td>
<td>[74] [75]</td>
</tr>
<tr>
<td>SIMPL-IPX</td>
<td>UC Berkeley</td>
<td>I</td>
<td>Converter (BTU)</td>
<td>none</td>
<td>SIMPL-DIX</td>
<td>[58] [70]</td>
</tr>
</tbody>
</table>

Status Meaning
I Internal: The system is not used outside the institution listed.
P Production: The system is known to be in use at all.
C Commercial: The system is commercially available.
E Experimental: The system is an experimental implementation.
R Industrial Prototype: The system is a prototype implementation.

require both accuracy and robustness of the tool. In this case, very sophisticated physical models and numerical techniques must be used, usually at high computational costs. An example of such a model is the two-dimensional simulation of the transient-enhanced diffusion during rapid thermal annealing [28].

On the other hand, for statistical simulations [6] or post-design process optimization [72] speed is the most crucial issue, as physical models can be calibrated to an existing manufacturing process and hence do not pose a reliability problem.

Independently of the progress in advanced physical modeling, the fast and simple “tuned” models will still remain in broad use; there is no unique “best model” for all simulation problems.

TCAD involves a number of scientific disciplines in addition to electrical engineering and computer science. This has also had an impact on the characteristics of the software which has been produced by that heterogeneous community during the past 20 years. From a historical perspective, the evolution of TCAD started with single applications (so-called point tools)
which were implemented independently by different developers. A typical, traditional point tool solves a specific TCAD problem, it has usually been designed without consideration of a (software) environment in which it will be used.

During the past years these tools have been integrated with other services (like visualization, interactive design editing facilities, a tool control level) to form Technology CAD systems. To support the users, most of these TCAD systems are currently equipped with converter-based tool coupling, homogenizing user interfaces and TCAD-specific task level environments which are often centered around typical design tasks like optimization. This is very effective in facilitating the current use of the system, but it does not address the real software-related difficulties retarding the progress of Technology CAD.

As stated, TCAD systems which provide process simulation, device simulation, common services and miscellaneous task control facilities are nowadays used worldwide. Some of them were developed by extending existing integrated Electronic CAD (ECAD) systems or Computer Integrated Manufacturing (CIM) systems. Thus it is very hard to distinguish between advanced ECAD, CIM, and TCAD systems; a review and classification of ECAD systems is given in [35]. Furthermore, there are many TCAD approaches which emphasize certain aspects (like optimization) and even integrate some simulators, but entirely lack to fulfill any of the requirements and architectural criteria presented here. Hence, despite their practical utility, these systems have been discounted. Other remarkable achievements which unfortunately lack a persistent implementation have also not been considered. For these reasons and because of the sparse published data, the overview of existing systems given in Table 1 is incomplete and may not accurately represent the current status in all cases. Only a minority of systems is publicly available, hence a detailed classification of the approaches taken for data level, task level, and presentation level is not possible within the scope of this text. Where in doubt, fields in the table have been left blank. Note that SSF is a derivative of the SUPREM-4 file format, designed by Silvaco, and TIF is a derivative of the SUPREM-4 file format, by TMA.

The systems listed exhibit an impressive variety of architectures and interesting features. Despite their shortcomings from a methodology point of view, each of them (and of many more unlisted TCAD components) is a remarkable achievement in terms of practical utility. An in depth overview of different existing TCAD systems is given in [15].

2 User Interfaces in TCAD Systems

The state-of-the-art interfaces to computing devices, i.e. workstations, are capable of displaying many virtual screens, which are usually termed windows, on a large physical screen. Therefore, several workstation-based TCAD systems can be found which address the issue of multi-tool integration into a unified user interface, mostly based on the X Window System [33]. PRIDE [61], based on SunView which is a vendor specific variation of the X Window System, exhibits a user interface and task level architecture which is strongly influenced by the preprocessing – computation – postprocessing task model of TCAD. SIMPL-IPX [58], based directly on Xlib [50], features a central interactive graphical editor which has menu-oriented facilities for running simulators. In both cases, implicit or explicit assumptions about the design cycle have an impact on the (top-down) design of the software and restrict the design tasks which can be performed or implemented. A more flexible and extension-oriented user interface architecture has been accomplished in PROSE [75], which is mainly due to the use of the generic Tcl interpreter [52] and Tk toolkit [53].
The VISTA solution

The structure of the VISTA user interface [27] is shown in Fig. 1. The bottom layer is the X Toolkit [1], an object-oriented subroutine library, designed to simplify the development of X Window System applications. The X Toolkit defines methods, so called Intrinsics [51], for creating and using widgets, which appear to the user as pop-up windows, scrollbars, text-editing areas, labels, buttons, etc. Basic functionality is provided by the generic Athena widgets, which are part of the public domain X Window System distribution from the Massachusetts Institute of Technology. This widget set was used rather than any other open standard, because a migration from these generic widgets to another widget set (like OSF/Motif [57]) is significantly easier than vice versa.

A widget-wrapping layer has been put on top of these widgets in order to achieve some widget-set independence. All widgets are created and modified via specific functions rather than via the generic interface of the X Toolkit. This facilitates the potential migration of the entire user interface onto another X Toolkit-based platform.
In addition, specialized VISTA widgets have been developed on top of the widget-wrapping layer for supporting TCAD-related information flow [54] [55]. The VISTA widgets are also created and accessed via specific functions, so that they can more easily be replaced by other widgets, should the need arise.

![Monte Carlo Ion Implantation](image)

Figure 2: This widget aggregate is the user interface for editing the parameters of a PROMIS Monte Carlo ion implantation step.

Batch-mode applications which do not require direct user interaction (like all classical simulation programs) can also be equipped with a user interface, either for direct control (which is less often required) or for operation supervised by the Simulation Flow Control (SFC) module as described later. The situation is depicted in Fig. 1, where the user interface is implemented as a task-level extension language (LISP) macro, entirely decoupled from the tool code. The widget macro shown in Fig. 2 is such a tool control panel. Most tool control panels are created from formal specifications of the tools by simple interface generators which are implemented in the extension language. This relieves the application engineer from the need to use low-level programming to create new tool control panels. The separation of the user interface code from the tool itself is an important means to stabilize and unify the user interface behavior and to relieve the application code from the burden of the user interface implementation.

### 3 The Task Level in TCAD Systems

Looking at existing systems in CAD areas, it seems obvious that an interpreter is the classical solution for the task level environment ([2] [35] [36] [39] [42] [75]). Both single interactive actions and more complex flow control can be executed in an interpreted environment.

A UNIX- (or any other operating system) shell based solution does not fulfill the portability
requirement, whereas the use of an integrating master application (like an interactive device editor) alone does not nearly offer enough flexibility.

A conceivable, fashionable alternative is an entirely object-oriented concept where both CAD data, e.g. wafer model, and simulation tools are objects (e.g. [8] [73]). The necessary sequence of operations to accomplish a given design goal can be deduced by rule inference based on the properties and relationships of these TCAD-objects and are realized by calling the appropriate methods. At a closer look, however, the two concepts are not mutually exclusive. In fact, the use of an interpreter is merely an implementation-oriented early architectural choice which does not at all preclude the latter introduction of object-oriented concepts.

Although there are many TCAD systems which successfully use Tcl [52] and Tk [53] as task level interpreter and user interface (see [36], [39], [46], [69]), a solution based on a LISP like interpreter is also feasible. It can be extended and customized for TCAD purposes by both adding C coded primitives and by loading LISP code at run time.

The appropriateness of such a choice is confirmed by several other remarkable, LISP-based implementations of task level environments in related fields. Examples are the well-known GNU Emacs [62] text editor which uses LISP as extension and top-level implementation language, the generic CAD system AutoCAD [2] which derives much of its success from third-party applications implemented in the SCHEME-like extension language AutoLISP, or Winterp [44] ("Widget Interpreter"), an experimental user interface prototyping environment which is part of the public domain X Window System distribution of the Massachusetts Institute of Technology. Other integrated CAD systems which allow for the definition of complex, data-driven control flows often use LISP as major implementation language [8] [21].

A required task-level (LISP) function can be implemented as

- a LISP coded function loaded at run time,
- a C coded function which is linked with the LISP interpreter, or
- an external application, implemented as separate executable

Whichever method is used does hardly make a difference at the interpreter level. The functionality is always presented as a single LISP function.

For all batch mode applications (these are applications that do not require user interaction) the task level integration and the presentation integration (according to [30] and [71] that is the provision of applications with a homogeneous user interface) are accomplished within a single context. LISP functions are used to create a layer of virtual applications on the task level for one or more physical applications (executables). Fig. 3 shows two characteristic examples.

A typical example for a one-to-many mapping ("Application A" in Fig. 3) is a device simulator, where for every desired type of device characterization a function that performs the required simulation step and extracts the requested parameter is implemented ("Virtual Applications W" and "X" in Fig. 3). For process simulation the situation is inverted. Several physical applications are run consecutively under control of a single presentation.

Note that this feature is only possible with a dedicated task level that separates the applications from the presentation level. This is an important prerequisite for the portability and continuity of the user interface. Furthermore, the intermediate task level layer allows the user interface to be tailored to design tasks instead of application peculiarities.

The VISTA solution

VISTA's task level environment is predominantly based on XLISP [4], a comprehensible and compact LISP interpreter. The highly portable C source code of this public-domain product is
Figure 3: A set of physical applications is mapped to a set of virtual applications on the task level. The user interface for these virtual applications is implemented in the task level programming environment, leaving the physical applications entirely unaffected.

freely available.

To preserve the consistency and simplicity of XLISP and in order to provide a homogeneous procedural interface and programming environment, the X Window System interface in Fig. 1 for XLISP was implemented from scratch in VISTA. As there are other c-coded parts of the framework which need to be accessible on the extension language level, a generic automatic method for linking given functions with the XLISP interpreter has been implemented.

The object-oriented callback concept of the X Toolkit has been generalized in a very straightforward manner and successfully applied to those parts of the TCAD framework where a strict decoupling of functional modules and high flexibility of the control flow is desirable. It is obvious that this is of special value for a flexible task level implementation.

Events coming from the X Window System system are passed to the XLISP interpreter. If a LISP expression has been associated with the activated widget at creation time, this expression is then evaluated by the interpreter and can be used to change parameter values, trigger the execution of a simulator or start the evaluation of a LISP program.

The same callback concept is also used for the control of simulator execution. If a simulation tool terminates, it signals the termination to the parent process, which again causes an associated callback expression to be evaluated. Callbacks can be triggered by the user interface, error handler, network layer, or by the termination of child processes.

Many computationally extensive design tasks, like statistical computations, exhibit intrinsic parallelism and can hence be effectively and easily parallelized using the callback technique. Whenever there is no data dependency (which is the case for, e.g., stationary multiple operating point analysis) the simulations can be done in parallel and "only" need to be distributed on different workstations or servers and synchronized at the end.

Fig. 4 shows two (simulation) processes which are run in parallel. The termination of simulation process 1 causes the termination callback 1 to be executed and can be used to trigger the computation of the next step. The choice between a purely sequential simulation flow and a parallel simulation flow is merely a matter of callback configuration and does not require any
major change or adaptation of the task-level programs. The subprocesses can be distributed on different hosts to utilize the latent computational power of common workstation clusters.

![Diagram of two independent simulations run in parallel under control of the XLISP interpreter.](image)

Figure 4: Timing diagram of two (independent) simulations which are run in parallel under control of the XLISP interpreter.

The *Simulation Flow Control* (SFC) module [54] of VISTA is a high-level utility of the task level responsible for the definition, management, and execution of simulation sequences. The main field of application is the reproduction of fabrication processes, which usually involves several independent process simulators. The task flow is stored in a simulation flow description using symbolic names to call simulation tools. Larger process flows can be assembled from predefined sequences, the results of intermediate steps may remain available upon request for later analysis. The SFC functionality includes dependency analysis, interfaces for process parameter variation, and lot splits. The SFC is entirely coded in LISP, uses the callback concept for simulator synchronization, and is executed by the XLISP interpreter.
4 Tool Integration in TCAD Systems

The main motivation for TCAD systems is the integration of different simulators. It is of course always possible to integrate another tool into an existing set of tools, but the decisive criterion to assess the support offered by the framework is the initial effort and the maintenance effort involved in the integration. Another criterion for a proper application framework architecture is whether adaptations of already integrated tools are required when other, new tools are integrated or changed.

From a software point of view, a significant amount of TCAD methodology is technology-independent. This generic functionality must be provided as part of the framework, usually in form of libraries and applications. The scope of these services ranges from integrated CASE tools (e.g. [20]) for application development, scientific visualization, interactive geometry editing facilities, to the more traditional TCAD methodology like grid generation, interpolation, and many more.

There are many situations in TCAD, especially when coupling different simulators, where interesting semantic problems occur on the data level. Among the most intriguing challenges are potential inconsistencies between the geometry of the simulation domain and one or more grids with attributes defined on them, altogether describing the current state of a wafer sub-domain.

These inconsistencies are inevitable, and are not a consequence of “misbehaving” applications or of an insufficient data representation or architectural concept. Given the fact that the simulators are granted “expert status” and that they are allowed to focus on their specific task, they will always use different abstractions of the wafer state and produce and affect just the data relevant to the specific problem they are modeling. It is reasonable and justified that they neglect all other issues, and hence these latent inconsistencies must necessarily arise. It is the duty of the framework to provide means for resolving those conflicts and inconsistencies. Once these semantic gaps are bridged by the framework, simulators can work together constructively without having to care for potential grid- and geometry-related conflicts they may create.

There are numerous grid-related problems and conflicts that do arise when multiple state-of-the-art simulation tools are used to simulate practical device fabrication steps. Some of these problems can be solved by interpolation services.

However, several problems require a grid handling in between two calls of simulation tools that goes beyond simple interpolation. Amongst them are the mapping of non-geometry-conforming grids onto geometry-conforming ones, splitting of grids which extend across several “segments” (layers), superposition of attributes (e.g. an additional boron implant in a structure where boron had already been present), a change in the geometry to which the doping information must be adapted, or the extraction of a subdomain of the geometry for a simulation tool.

In a more general formulation, it is required to map a set of grids of different types, with or without spatial overlap, geometry conforming or not, and with several “attributes” (dopants or such) defined on them, onto a new set of geometry conforming grids. The most important constraint of this mapping is the qualitative statement that there should be very little information or accuracy lost.

The VISTA solution

To solve this generic problem in VISTA rigorously, a framework tool (VORONOI [27]) has been implemented which is used to resolve all of the conflicts listed above. The re-gridding problem is sub-structured in several smaller problems which lead to fairly independent functional modules that perform operations on common data.
The basic idea to overcome any of the grid/geometry consistency problems rigorously is to treat grid and geometry points equally by merging the set of geometry points and the set of grid points and by finally (re-)triangulating the resulting point cloud using a constrained Delaunay triangulation.

Such functionality is a minimum requirement of a TCAD system, and the implementation of grid conversion and related problems in general is a key to the evaluation of TCAD systems.

5 A Shallow Trench Isolated CMOS Technology

A major challenge of multi-tool TCAD lies in the robust reproduction of increasingly complex manufacturing processes. This challenge is particularly prevalent in modern fabrication processes with a tight interaction between structuring and doping techniques, like planarization and trench isolation processes. A fully planarized, trench-isolated 0.25µm CMOS process has been presented in [9] [10] [11] and is used here to demonstrate the coupling of several specialized process simulation tools by means of the VISTA framework.

The Monte Carlo simulation module of PROMIS [63] is used for the simulation of all ion implantation steps. It is a traditional FORTRAN simulator and has been integrated into VISTA directly using the PAI/PIF [17]. It reads only the device geometry and produces the resulting doping profiles on a single, non geometry-conforming tensor product grid which covers the bounding box of the entire device geometry.

TSUPREM-4 [67] is used to simulate the diffusion steps. It has been integrated by means of a wrapper which converts the wafer state from VISTA PIF to TIF (TMA’s Technology Interchange Format) and vice versa. TSUPREM-4 requires and produces doping information on several boundary-conforming triangular grids, one for each geometry segment.

The simulator ETCH [64] [65], newly developed using VISTA as implementation basis, utilizing high-level libraries [16], is used for the simulation of etching and deposition steps. It reads only the geometry and produces a new, changed geometry. As ETCH uses a cellular data structure to perform purely geometrical operations, it cannot change the wafer state mesh accordingly. Hence, after etching and deposition steps, the (old) wafer state mesh does not match the new geometry produced.

The auxiliary tool SKETCH is used to perform simple changes of the wafer geometry due to spin-on, exposure, and mask strip steps.

The VORONOI re-gridding and interpolation service is used by the SFC to achieve and maintain consistency between grid and geometry, and to transform the structural and doping information between tool-specific and wafer state compliant form.

Table 2 shows the simulation steps executed by the SFC. This “virtual process” is a simplification of the real fabrication process. The virtuality of the simulation enables us to, e.g., “expose and develop” a nitride mask directly (step 10). This will be justified as long as the lithography and resist development of this mask layer is not a critical process step.

The desired final device structure is shown in Fig. 5, together with the one-dimensional (symbolic) mask information used in steps 3, 11, 16, 22, 30, and 35.

Fig. 6 shows the phosphorus doping of the n-well after the trenches for the device gates and for the passing gates have been etched. This etch step has removed part of the geometry at the wafer surface. The doping in the void has been automatically removed, and the grid has been made geometry-conforming to produce a consistent wafer state after completion of step 24.

Fig. 7 is a screen dump of the xpi2d visualization client showing the final planarized geometrical structure of the n-channel transistor along with the phosphorus concentration (after step 28). The visualization uses the vector graphics widget of the VISTA user interface.
<table>
<thead>
<tr>
<th>step</th>
<th>simulator</th>
<th>SFC operation</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SKETCH</td>
<td>create-subdomain</td>
<td>“Si”, $5.6 \times 2.5\mu m$</td>
</tr>
<tr>
<td>2</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Resist”, $0.2\mu m$</td>
</tr>
<tr>
<td>3</td>
<td>SKETCH</td>
<td>expose</td>
<td>Mask 1</td>
</tr>
<tr>
<td>4</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Exposed”</td>
</tr>
<tr>
<td>5</td>
<td>ETCH</td>
<td>plasma-etch</td>
<td>misc. etch rates, 320 seconds</td>
</tr>
<tr>
<td>6</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Resist”</td>
</tr>
<tr>
<td>7</td>
<td>ETCH</td>
<td>iso-depo</td>
<td>“SiO2”, 60 seconds at $125nm/s$</td>
</tr>
<tr>
<td>8</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Resist”, $0.25\mu m$</td>
</tr>
<tr>
<td>9</td>
<td>ETCH</td>
<td>iso-etch</td>
<td>misc. etch rates, 330 seconds</td>
</tr>
<tr>
<td>10</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Si3N4”, $1\mu m$</td>
</tr>
<tr>
<td>11</td>
<td>SKETCH</td>
<td>expose</td>
<td>Mask 2</td>
</tr>
<tr>
<td>12</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Exposed”</td>
</tr>
<tr>
<td>13</td>
<td>PROMIS</td>
<td>monte-carlo-implant</td>
<td>“Phosphorus”, $5 \cdot 10^{11} , 1/cm^2, 60\text{keV}$</td>
</tr>
<tr>
<td>14</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Si3N4”</td>
</tr>
<tr>
<td>15</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Si3N4”, $1\mu m$</td>
</tr>
<tr>
<td>16</td>
<td>SKETCH</td>
<td>expose</td>
<td>Mask 3</td>
</tr>
<tr>
<td>17</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Exposed”</td>
</tr>
<tr>
<td>18</td>
<td>PROMIS</td>
<td>monte-carlo-implant</td>
<td>“Boron”, $5 \cdot 10^{11} , 1/cm^2, 60\text{keV}$</td>
</tr>
<tr>
<td>19</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Si3N4”</td>
</tr>
<tr>
<td>20</td>
<td>TSUPREM-4</td>
<td>diffusion</td>
<td>inert, 3.5 minutes at $1000\text{°C}$</td>
</tr>
<tr>
<td>21</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Resist” $0.5\mu m$</td>
</tr>
<tr>
<td>22</td>
<td>SKETCH</td>
<td>expose</td>
<td>Mask 4</td>
</tr>
<tr>
<td>23</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Exposed”</td>
</tr>
<tr>
<td>24</td>
<td>ETCH</td>
<td>plasma-etch</td>
<td>misc. etch rates, 200 seconds</td>
</tr>
<tr>
<td>25</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Resist”</td>
</tr>
<tr>
<td>26</td>
<td>ETCH</td>
<td>iso-depo</td>
<td>“SiO2”, 60 seconds at $1.1nm/s$</td>
</tr>
<tr>
<td>27</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Poly”, $0.25\mu m$</td>
</tr>
<tr>
<td>28</td>
<td>ETCH</td>
<td>iso-etch</td>
<td>misc. etch rates, 310 seconds</td>
</tr>
<tr>
<td>29</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Si3N4”, $1\mu m$</td>
</tr>
<tr>
<td>30</td>
<td>SKETCH</td>
<td>expose</td>
<td>Mask 5</td>
</tr>
<tr>
<td>31</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Exposed”</td>
</tr>
<tr>
<td>32</td>
<td>PROMIS</td>
<td>monte-carlo-implant</td>
<td>“Arsenic”, $2.5 \cdot 10^{14} , 1/cm^2, 33\text{keV}$</td>
</tr>
<tr>
<td>33</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Si3N4”</td>
</tr>
<tr>
<td>34</td>
<td>SKETCH</td>
<td>spin-on</td>
<td>“Si3N4”, $1\mu m$</td>
</tr>
<tr>
<td>35</td>
<td>SKETCH</td>
<td>expose</td>
<td>Mask 6</td>
</tr>
<tr>
<td>36</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Exposed”</td>
</tr>
<tr>
<td>37</td>
<td>PROMIS</td>
<td>monte-carlo-implant</td>
<td>“Boron”, $2.5 \cdot 10^{14} , 1/cm^2, 30\text{keV}$</td>
</tr>
<tr>
<td>38</td>
<td>SKETCH</td>
<td>strip-material</td>
<td>“Si3N4”</td>
</tr>
<tr>
<td>39</td>
<td>TSUPREM-4</td>
<td>diffusion</td>
<td>inert, 1 minute at $970\text{°C}$</td>
</tr>
</tbody>
</table>

Table 2: Process simulation flow for the $0.25\mu m$, fully planarized, shallow trench isolated CMOS process.
Figure 5: Final CMOS structure and symbolic mask information.

Figure 6: n-well doping profile (phosphorus) after etching the trench for the planarized poly gate. Result of step 24.
Figure 7: \textit{xpif2d} (the \textit{VISTA} visualization client) showing a detail of the n-well and a nearby passing gate structure. Result of step 28.
Fig. 8 shows a typical VISTA session where the simulation flow editor is used to edit the process flow. The parameters of the last step (the TSUPREM-4 wrapper in step 39) of the process flow described is being edited. The parts of the user interface shown are all created and executed by the task level shell.

Figure 8: Screen dump of VISTA session.
An Ultra-Low-Power CMOS Technology

The application of TCAD is perfectly suited to investigate specific constraints of technologies. The following example demonstrates a technology design study for Ultra-Low-Power applications. By decreasing the supply voltage and the threshold voltages, a great reduction of the power consumption can be achieved at the expense of an increase in gate delay. This can be compensated to a certain extent by employing parallelism in the systems design so that, for the same overall performance, the total power consumption is drastically reduced compared to conventional CMOS techniques [37]. We demonstrate the feasibility of ultra-low-power CMOS structures and determine a lower limit for the supply voltage, depending on the type of digital circuit technique.

The processes under consideration are recessed-well dual-gate processes with a very thin gate oxide (5nm and below) to obtain controllably low threshold voltages. The source/drain dopings are formed by single shallow implants and a conventional furnace anneal. The G/S and G/D overlap capacitances can be controlled with a spacer formed prior to the S/D implants.

As a consequence of the low voltages, ultra-low-power processes differ from conventional CMOS processes in several points: Because of the low $V_{DD}$, the hot-carrier problem does virtually not exist, and therefore an LDD (lightly doped drain) process is not necessary. Also, no GIDL (gate induced drain leakage) can occur.

As for very low $V_{DD}$ the devices must operate in the weak-inversion regime, the difference of the carrier mobilities $\mu_n, \mu_p$ can be roughly compensated by adjusting the threshold voltages to achieve symmetric inverter transfer characteristics. This compensation does not work, however, in the transient case because the speed is mainly determined by the strong-inversion part of the input characteristics. The sub-threshold behavior is crucial because it determines the achievable ratio of $I_{on}/I_{off}$, which is limited by $e^{V_{DD}q/kT}$ and decreases as $V_{T,n,p}$ are made smaller. Therefore, “zero-$V_T$” transistors are not desirable. On the other hand, if $V_{T,n,p}$ are too high the speed becomes unacceptably low. A major challenge is to achieve controllably low threshold voltages. Although the adjustment of $V_{T,n,p}$ with a bulk bias seems very attractive, this method is very likely not accepted for digital circuit design because of the significant overhead in layout due to the additional lines for bulk biasing.

Another problem can arise from the very thin gate insulator. If one uses very thin thermally grown gate oxides (below 5nm), boron diffusion can considerably degrade the device behavior. Also, boron segregation causes a deleterious effect, especially in the sub-threshold regime. On the other hand, there are several options for the gate insulator. Nitrides or oxinitrides may prove good alternatives to conventional (pure SiO$_2$) gate oxides. Silicon nitride can be used as an effective diffusion barrier, and ultra thin Si$_3$N$_4$ layers with low defect densities are easier to fabricate [45] than oxides. The tunneling barrier, lower than that of SiO$_2$, is still acceptable for ultra-low-power CMOS devices because of the low voltages and also because a controllably small gate current is allowed.

Both process and device simulation were done using VISTA to allow for quick process design and evaluation. For the electrical characterization of the devices MINIMOS [19] [59] was used to calculate a matrix of drain currents $I_D(V_G, V_D)$ over a range of $V_G$ and $V_D$ for the p-channel and n-channel transistor. Based on these data, a fast and accurate table-driven DC analysis of simple gates and inverters is possible.

The bulk effect could also have been included, but for the given devices and voltages it was found insignificant. The dynamic behavior was estimated from capacitance data obtained by AC analysis with MINIMOS.

The simulated processes were a 0.35$\mu$m process (A) for static logic and a 0.5$\mu$m process...
(B) for dynamic logic. The processes were designed for proper DC characteristics but were not optimized for speed. The device characteristics for process A are shown in Fig. 9 and 10. Fig. 11 and Fig. 12 show the inverter transfer curves.

Fig. 13 and Fig. 14 show the noise margins and the inverter delay as a function of the supply voltage. From Fig. 13 it can be seen that a ring oscillator built with process A would work even at $V_{DD} = 80\,\text{mV}$, and by using additional inverters at the gate inputs and outputs one could also design digital circuits for $V_{DD} < 100\,\text{mV}$, but the overhead of the additional components would be considerable.

<table>
<thead>
<tr>
<th>process</th>
<th>$V_{DD}$</th>
<th>$V_{T,n}$</th>
<th>$V_{T,p}$</th>
<th>$I_{off,n}$</th>
<th>$I_{off,p}$</th>
<th>$I_{on,n}$</th>
<th>$I_{on,p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>0.067</td>
<td>-0.059</td>
<td>$0.14 \cdot 10^{-6}$</td>
<td>$0.27 \cdot 10^{-6}$</td>
<td>$16.7 \cdot 10^{-6}$</td>
<td>$9.4 \cdot 10^{-6}$</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>0.26</td>
<td>-0.24</td>
<td>$0.7 \cdot 10^{-9}$</td>
<td>$2.8 \cdot 10^{-9}$</td>
<td>$25.6 \cdot 10^{-6}$</td>
<td>$16.7 \cdot 10^{-6}$</td>
</tr>
</tbody>
</table>

Table 3: Simulated device characteristics.
The threshold voltage was defined as $|I_D(V_T)| = 1\,\mu\text{A}/\mu\text{m}$.
All voltages are in V, all currents are in A/\mu m.

<table>
<thead>
<tr>
<th>process</th>
<th>$N_{MH,inv}$</th>
<th>$N_{ML,inv}$</th>
<th>$N_{MH,gate}$</th>
<th>$N_{ML,gate}$</th>
<th>$t_d$</th>
<th>$t_l$</th>
<th>$E_s$</th>
<th>$P_{stat}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28</td>
<td>23</td>
<td>13</td>
<td>39</td>
<td>0.29ns</td>
<td>7.2ns</td>
<td>0.65fJ</td>
<td>41nW</td>
</tr>
<tr>
<td>B</td>
<td>38</td>
<td>44</td>
<td>31</td>
<td>49</td>
<td>0.55ns</td>
<td>1.3µs</td>
<td>4.3fJ</td>
<td>0.88nW</td>
</tr>
</tbody>
</table>

Table 4: Noise margins (in %$V_{DD}$) for a simple inverter and a 3-input NAND gate, and inverter delay, leakage time, switching energy, and static power consumption.

For Process B, the ratios of $I_{on}/I_{off}$ in Table 3 are in the order of $10^4$ and the ratio of $\tau_l/\tau_d$ (leakage time to delay time) is about 2300, which is rather low for dynamic logic. For process A, it can be seen from Table 4 that for 3-input NANDs with minimal transistors the high-noise-margin $N_{MH}$ is already very low.

From these data we conclude that the limits for the supply voltage will be at $200\,\text{mV}$ for static logic and $500\,\text{mV}$ for dynamic logic with a fan-in of 3 at $T = 300\,\text{K}$.

The ultimate limit for the supply voltage is given by the thermal voltage as $V_{DD} > \frac{X \cdot kT}{q}$ where $X$ is a factor depending on the type of digital circuit technique and on the process technology. We found that $X_{stat} < 8$ and $X_{dyn} < 20$ is sufficient for a fan-in of 3.
Figure 9: Input characteristics, process A ($V_{DD} = 200\text{mV}$).

Figure 10: Output characteristics, process A ($V_{DD} = 200\text{mV}$).
Figure 11: Inverter transfer characteristics for (from left to right) 
\( W_n/W_p = 10.0, 6.3, 4.0, 2.5, 1.6, 1.0, 0.63, 0.4, 0.25, 0.16, 0.1, \) 
process A at \( V_{DD} = 200\text{mV}. \)

Figure 12: Inverter transfer characteristics for (from left to right) 
\( W_n/W_p = 10.0, 6.3, 4.0, 2.5, 1.6, 1.0, 0.63, 0.4, 0.25, 0.16, 0.1, \) 
process B at \( V_{DD} = 500\text{mV}. \)
Figure 13: Noise margins and delaytime vs. $V_{DD}$, process A.

Figure 14: Noise margins and delaytime vs. $V_{DD}$, process B.
7 A High-Electron-Mobility-Transistor Technology

One of the major issues of future TCAD applications is the investigation of novel device structures, amongst which the heterostructure devices have a prominent position. The following discussion of HEMT (high electron mobility transistor) simulation also highlights some aspects of TCAD in the simulation of such devices.

Figure 15: Structure of a HEMT.

Fig. 15 gives the structure of a low-noise HEMT which is currently in industrial use. The device consists of several layers, of which the undoped channel ($In_{0.2}Ga_{0.8}As$) is responsible for current transport, whereas the doped supply layer ($Al_{0.2}Ga_{0.8}As$) isolates the channel from the gate due to its large band gap. The supply layer doping stops about 2 nm above the channel to reduce scattering induced by dopants. The channel sits on top of an undoped substrate (GaAs), and source and drain contacts are put on special cap layers (GaAs) on both sides. The doping of the supply layer is typically some $10^{17}$ to some $10^{18}$ cm$^{-3}$.

It is of crucial importance for the simulation of modern devices that a framework be able to handle not only the multilayer geometry (consisting of 15 layers and more) of such devices, but also the material information. The latter will in general not be restricted to homogeneous material alloys within the single layers, nor will it suffice to support only a handful select material types. Thus, a powerful data format must provide a solid basis for such simulations.

As for the material types, VISTA provides a material database which is used in the simulations shown here. It provides a simulation tool with basic material parameters and is extensible for the needs of specific users. Several mobility functions that can be chosen independently for the various layers are also contributed to this simulation.

To obtain reasonable results, a proper model for the carrier current and energy flux at the heterojunction is indispensable, since carrier heating in the channel region is one main effect to determine the current. The model must be physically correct as well as numerically
stable, which implies that the handling of the segments must take the segment boundaries into special account. Unless one is interested in the quantum physical situation at the boundary, the boundary should be simulated as an abrupt junction in the band edge energies and effective densities of states. These assumptions lead to an electron concentration which is discontinuous at the hetero-junction, having two distinct values at the boundary, each of which is only valid for one side of the junction.

![Figure 16: IV-Characteristics of the HEMT.](image)

The IV-characteristics of the HEMT (Fig. 16) show a behavior similar to a MOS-transistor within the usual operating area, except for the better high frequency properties and different breakdown mechanisms. Alas, without a detailed account for the specific structure it is impossible to obtain these characteristics.

It is, however, not the only task of the framework to provide material parameters. Important are tasks like design optimization for the device in question or parameter fitting to adjust the various model parameters to the local process data. In these points, the notion of a framework shows its general superiority above single simulation runs, each of them triggered and evaluated separately by the user.

Fig. 17 and 18 show the electron concentration and temperature in the HEMT presented above. One might note the different concentration values in the different layers, which are connected at the abrupt hetero-junction by a model that takes into account the different band edge energies on both sides. The temperature in the channel reflects the influences of the high field under the gate and the reduction due to doping in the source and drain regions. A hydrodynamic model was applied in the channel region to obtain these results.
Electron Concentration [cm$^{-3}$]

Figure 17: Electron concentration in the HEMT.

Electron Temperature [K]

Figure 18: Electron temperature in the HEMT.
8 Conclusion

It is only due to the dedicated application-framework architecture that the simulation tools integrated into a TCAD system can still focus on a specific technological problem and may even use entirely different methods and internal problem representations. The rapid evolution of these simulators is driven by fabrication technology. The generic, technology-independent functionality of the framework helps to bridge the gaps between the simulators and provides the device and process engineer with a more stable, homogeneous representation of simulation capabilities.

Whereas the users of an integrated TCAD system will not necessarily notice architectural software shortcomings, both tool developers and tool integrators will experience the retarding effect of the prevailing application-driven TCAD system approach. It is hence vital for the TCAD framework that it addresses properly not only the application of methodology, but also the entire creation process of TCAD methodology. This definitely includes support for tool integration and tool development.

The study of Ultra-Low-Power device technology, performed within the VISTA system, predicts convincing features of this technology. It also gives evidence of the practical usability of a TCAD framework in the development and verification of new designs. According to the data obtained from this study, the possibilities of low power concepts are by far not exhausted in present applications.

The HEMT transistor which has been analyzed in the last example emphasizes yet another aspect of TCAD in the future: Frameworks will no longer be allowed to content themselves with restricting to a minimum concept of a data format. Flexibility in device geometry must come with flexibility in the physical data, as are materials and their properties. The power of a framework, however, shows best in the typical tasks required in future device characterization: Design optimization and parameter fitting are two prominent examples.

Acknowledgments

The presented work has been sponsored by the research laboratories of AUSTRIA MIKRO SYSTEME at Unterpremstätten, Austria; DIGITAL EQUIPMENT at Hudson, USA; HITACHI at Tokyo, Japan; MOTOROLA at Austin, USA; NATIONAL SEMICONDUCTOR at Santa Clara, USA; SIEMENS at Munich, FRG; and SONY at Atsugi, Japan, and by the "Forschungsförderungsfonds für die gewerbliche Wirtschaft", project 2/285 and project 2/299, as part of ADEQUAT (JESSI project BT1B), ESPRIT project 7236, and by ADEQUAT II (JESSI project BT11), ESPRIT project 8002.
References


