A NOTATION FOR DESIGNING RESTORING LOGIC CIRCUITRY IN CMOS

by
Martin Rem
Eindhoven University of Technology
and California Institute of Technology

and
Carver Mead
Professor of Computer Science,
Electrical Engineering and Applied Physics
California Institute of Technology

Technical Report #4600

Computer Science Department
California Institute of Technology
Pasadena, California 91125

Sponsored by
Defense Advanced Research Contracts Agency
ARPA Order Number 3771

Monitored by
Office of Naval Research
Contract #N00014-79-C-0597

A NOTATION FOR DESIGNING RESTORING LOGIC CIRCUITRY IN CMOS

Martin Rem
Eindhoven University of Technology
and California Institute of Technology
and
Carver Mead
Professor of Computer Science, Electrical Engineering
and Applied Physics
California Institute of Technology

1. INTRODUCTION

As the underlying silicon fabrication technology has become capable of producing chips with transistor counts in excess of 1,000,000, problems associated with correct design are assuming ever greater importance. Exhaustive checking of mask artwork for errors becomes prohibitive. Technologies and design styles which obviate large classes of potential errors are enormously preferable to those that do not.

A modular, hierarchical design style can, with proper restriction, confine many types of checks to one level of the hierarchy within each module. A set of such restrictions is given in this paper, together with a mechanism for their enforcement. These restrictions capture a substantial fraction of the design style given in [1].

As feature sizes are scaled below one micron, ratio logic processes like nMOS and I^2L become progressively less attractive. Straightforward scaling to smaller sizes results in a linear increase in current per unit chip area. Technological tricks such as high resistivity polysilicon pullup devices or very small injector current can be used to decrease current drain, but the resulting devices become increasingly vulnerable to "soft error" problems from alpha particles, etc. Fully restored "static" logic using a complementary process is the natural choice for systems with submicron components. Present bulk CMOS processes have a number of very ugly analog rules associated with the 4-layer nature of the process. As a result, the designer must be aware of details of the technology to an alarming degree. CMOS on an insulating substrate is, on the other hand, a conceptually clean process: it requires no analog rules whatsoever if proper timing conventions are observed. There are recent signs that it may become reliably producible as well.

We introduce a programming notation in which every syntactically correct program specifies a restoring logic component, i.e., a component whose outputs are permanently connected, via "not too many" transistors, to the power supply. It is shown how the specified components can be translated into transistor diagrams for CMOS integrated circuits. As these components are designed as strict hierarchies, it is hoped that the translation of the transistor diagrams into layouts for integrated circuits can be accomplished mechanically.

In this paper we do not address the dynamic behavior of the logic components. The "proper timing conventions," alluded to above, are left for a subsequent paper.

2. SWITCHES IN CMOS

The CMOS technology uses two types of transistors: the N-channel enhancement transistor (1a) and the P-channel enhacement transistor (1b).

Fig. 1

Both of them act as switches but they are "on" and "off" for complementary values on their gates. Denoting a high voltage by "1" and a low voltage by "0", switch 1a is on if the gate is 1 and 1b is on if the gate is 0. When the switches are on, however, they do not convey a 1 and a 0 on their paths (in Fig. 1 the horizontal connections) equally well. Switch 1a conveys a 0 virtually perfectly, but it is not a perfect switch for a 1. Switch 1b, conversely, is a good conveyor for a 1 only.

Using these CMOS transistors we want to make two types of switches, a "normally-off" switch (2a) and a "normally-on" switch (2b).

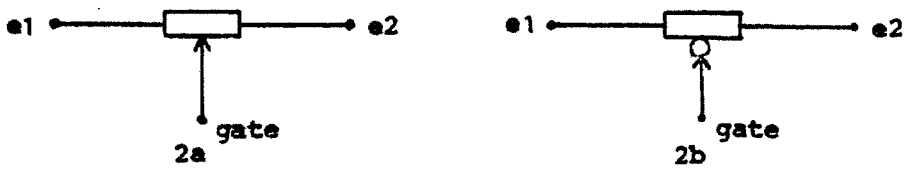


Fig. 2

If the gate is 0 switch 2a is off (nonconveying) and 2b is on (conveying). Otherwise 2a is on and 2b is off. The points e1 and e2 are called the end points of the switch. We call the connection between the end points its path. If nothing is known about the values conveyed through its path, except that they are 0's and 1's, the realization of a switch requires two transistors: (the complement of g is denoted as g')

3. RESTORING LOGIC COMPONENTS

A restoring logic component (RL) has external ports. The purpose of an RL is to establish a relation between the values it communicates via its external ports. We restrict ourselves to the values 0 and 1.

We design components in a hierarchical fashion. A typical RL is shown in Fig. 5.

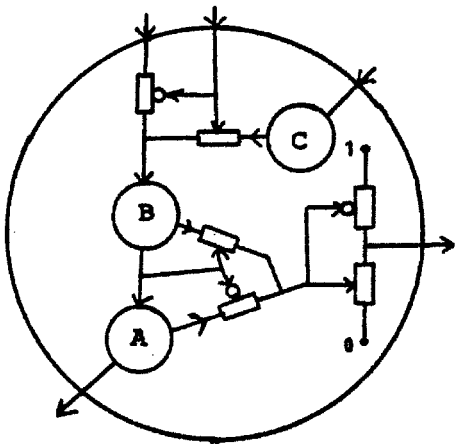


Fig. 5

It consists of subcomponents A, B, and C, which are also RL's, and a pattern of connections between them. We restrict the possible connection patterns to guarantee that the composite is again an RL. Such restrictions are only useful if they can be formulated in terms of the connection pattern, i.e., independent of the internal structures of the subcomponents thus connected. Before we can formulate these connection rules we have to give a few definitions. Each port is either an input port or an output port. The connection pattern of an RL specifies connections between its external ports and the external ports of the subRL's. We call the external ports of a subRL internal ports of the RL. An external output port of a subRL is an internal input port of the RL. Conversely every external input port of a subRL gives the RL an internal output port. The rules on connection patterns will be stated in terms of external and internal ports of the RL.

We assume that the distribution of power and ground to all components is taken care of by the compiler. Johannsen [1] has outlined a method for the distribution of power and ground over hierarchically defined components. In our nomenclature: each RL has two constant internal input ports, denoted by 0 and 1. These constants are the power supply rails which must be present in every component.

In Section 2 we have introduced the term path for the connection between the two end points of a switch. We now generalize that term. We say that there is a path between two ports p1 and p2 if either they are connected by a wire (a "wire path") or there is a switch such that there are paths between p1 and one end point of the switch and between p2 and the other end point. In the latter case we say that the switch is on the path. A path is called a conveying path if all switches on

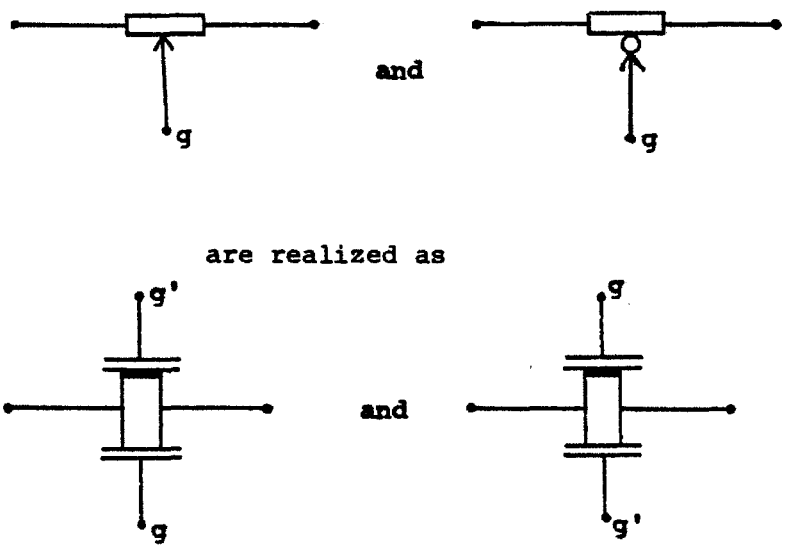


Fig. 3

These double transistors make our switches good conveyors for both 0's and 1's, which allows the use of longer strings of switches. These strings of switches, however, should not be too long: the distance to the "power supply" must not be excessive, otherwise the signal will become inaccurate and the circuit slow. To do justice to the nature of restoring logic we disallow the driving of external outputs by long strings of switches. This shall be reflected in the composition rules to be formulated in Section 3.

The gate inputs are run in two-rail logic to accommodate both the g and the g' signals. For switches that are known to convey always the same value there are two instances in which they can be realized by just one transistor:

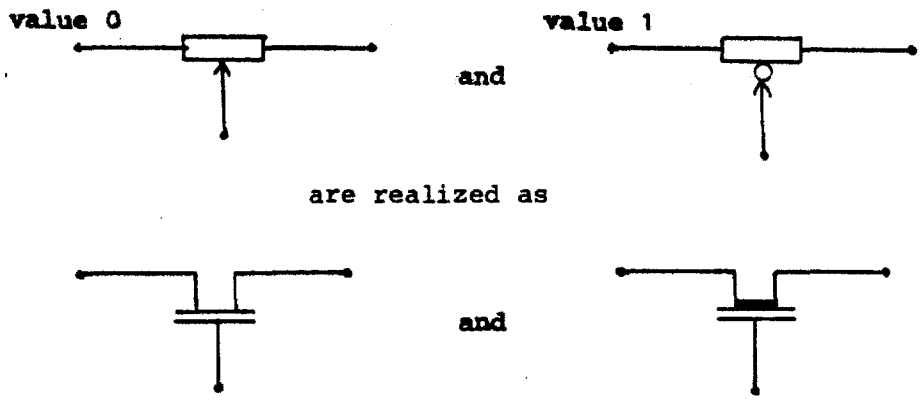


Fig. 4

In that case, the two-rail representation of the gate signal is not necessary. It is assumed that the compiler can recognize instances in which one transistor suffices. From now on we shall simply design in terms of switches and apply the above knowledge only if we wish to count the number of transistors a component requires.

the path are on. The values on the input ports (external or internal) determine which switches are on and which are off, and hence between which ports there are conveying paths. (Whenever we do not specify whether a port is external or internal, that is done intentionally.)

Two input ports are said to be fighting if there exists any assignment of values to all input ports such that there is a conveying path between the two input ports.

We introduce three rules the connection pattern must satisfy:

Rule 1. [no fighting]: No two input ports are fighting.
Rule 2. [restored external outputs]: Every external output port
(a) has a wire path to an internal port, or
(b) has a conveying path to 0 or 1 for every assignment of values to all input ports.
Rule 3. [nonfloating internal outputs]: For every internal output port p and for every assignment of values to all input ports there is a conveying path between p and an input port.

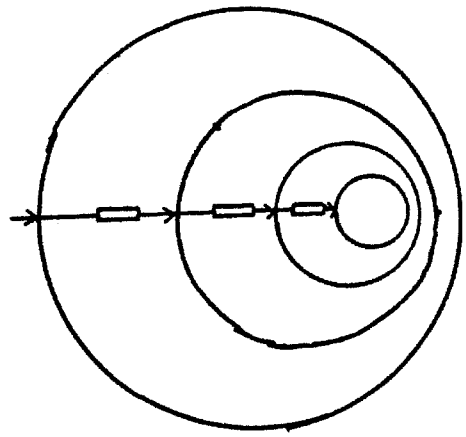
Notice that Rule 1 includes 0 and 1 (the two constant internal input ports). Remember that internal outputs are regarded as (external) inputs of the subcomponent and that the subcomponent's external outputs are internal inputs for the component.

The justification of Rule 1 is obvious. The result of Rule 2 is that all external outputs are driven by power or ground. They may be driven via a number of switches, but such a string of switches is confined to one component, viz. the component in which the actual connection to 0 or 1 is made.

The rules for internal outputs, i.e., outputs to subcomponents, are more liberal. We allow that inputs from subcomponents and inputs from the environment are directed through switches before they are output to subcomponents. For inputs from subcomponents this is reasonable: they are restored by the subcomponents. With inputs from the environment we have to be more careful. We have to allow that such a signal from an external input port goes through a switch to an internal output port. Otherwise we would be unable to make the flip-flop to be shown in Example 3. But it does allow long strings of switches "going into" the hierarchy, as sketched in Fig. 6.

We do not consider this a serious drawback. One may expect a subcomponent to have (physically) shorter connections than the component itself. Restoring in the "inward" direction, therefore, seems less vital than in the "outward" direction. Still, if we wish to bound the lengths of such inward strings of switches we could have the compiler insert amplifiers into them to restore their signals.

The consequence of allowing the switches in the outputs to subcomponents is that Rule 2 has to be stronger than one might expect. In Rule 2 we could not allow wire paths between external input ports and external output ports. This may seem to disallow running through a



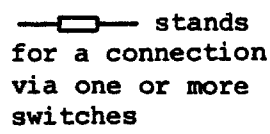


Fig. 6.

component wire whose signals are not used by the component. In fact, it does not. Such a wire is just not part of the component. (On the chip a wire between two components may run through the "area" of another component, but that is a matter of chip layout. It is a physical property, not a functional one.) Allowing wire paths between external input ports and external output ports would have given rise to the possibility of ill-restored outputs. Fig. 7 sketches an RL that is allowed by Rules 2 and 3. Now assume that each S_i is just a wire path from its input to its output, which would be allowed if we weakened Rule 2. The output of the RL is then not restored. Imagine now that each S_i actually has the same structure as the whole RL. It is clear that this would violate our goal of having restored external outputs.

In one respect is Rule 3 stronger than necessary. It requires that all subcomponents receive well-defined inputs, even a subcomponent whose ouputs are not used. We could have restricted the rule to subcomponents whose outputs are actually used in the computation, but that would have made both the rule and the checking whether it is obeyed more complicated.

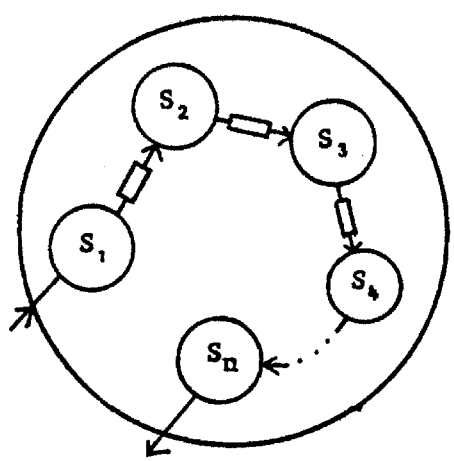


Fig. 7

4. THE PROGRAMMING NOTATION

In this section we introduce a programming notation in which connection patterns can be specified that satisfy the three rules of the preceding section. There are two properties a good notation should enjoy. First, it should be relatively simple for the compiler to check that a program is syntactically correct. If this mechanical check is simple, it will probably be simple for programmers to convince themselves that their designs satisfy the rules. We shall show how the syntactic checking can be performed. Second, it should be possible to give a formal definition of the semantics of our programs. We have not yet achieved the second goal, but ultimately we must be able to prove that a component performs a certain computation. That seems a much better technique than a demonstration of its effect with an a posteriori simulation. (Besides, how do we know that the simulation is correct if we do not have a rigorous definition of the meaning of our statements?) It will not be simple, but remember: a program of more than, say, 20 lines is probably too long, we then have not chosen the right subcomponents.

For the formulation of connection patterns we introduce the term node. Every port is a node, but the program may introduce additional (interior) nodes. For each node n we shall introduce a connection condition C(n) and a connected-to-constant condition CC(n). We shall, furthermore, distinguish a directly driven set D, which is a subset of the set of nodes. These concepts will be used in the syntax checking. A formal definition of how they depend on the connection pattern specified will be given later. Intuitively, C(n) will be the condition on the input values under which node n is connected to an input, and CC(n) will be the condition under which it is connected to a constant. The C(n)'s will be used to enforce the no-fighting rule. The set D will comprise all nodes that are connected by a wire path to an internal input port.

The program consists of a sequence of statements. Each statement introduces a number of connections and switches between nodes, and thereby affects the C(n) and CC(n) of each node involved and the set D. Initially, i.e., prior to the first statement, D is the set of all internal input ports, C(n) is 1 for each input port and CC(n) is 1 for the two constant internal input ports. The C(n) and CC(n) are 0 for all other nodes. ("1" should be interpreted as "true" and "0" as "false.")

The program is complete if finally we have:

for every external output port p : $p \in D \vee CC(p) = 1$
for every internal output port p : $C(p) = 1$

(These completeness conditions correspond to Rules 2 and 3. The observing of Rule 1 is discussed below.)

EXAMPLE 1

```
comp inverter (in?,out!):
begin in' → out = 1; in → out = 0 end
```

Prior to the statement

$$BE \rightarrow x = y$$

we should have

$$\text{for all nodes n in BE : } C(n) = 1, \text{ and}$$

$$(C(x) \wedge C(y) \wedge BE) = 0$$

The first requirement is introduced to permit the syntax checking to be done incrementally at each statement of the program. A consequence, however, is that not every order of the statements in the program is permissible. It is still an open question whether this serializability requirement is not too strong. If we succeed in designing our components under this regime it will certainly enhance both the readability and the checkability of our programs.

The second requirement guarantees the observance of the no-fighting rule. The statement does not have an effect on the set D. The effect on C(n) and CC(n) is

$$Z(x) := (Z(x) \vee (Z(y) \wedge BE))$$

$$Z(y) := (Z(y) \vee (Z(x) \wedge BE))$$

in which Z stands for C or CC.

The set D is affected only by a statement that specifies a direct connection, i.e., one that does not go through a switch. We obtain such a statement by dropping the conditional part "BE→":

$$x = y$$

As for the effect on C(n) and CC(n) this statement is like a switch specification with "1" as its boolean expression. Prior to the statement the condition

$$(C(x) \wedge C(y)) = 0$$

should hold, and its effect is that Z(x) and Z(y) both become Z(x) ∨ Z(y) (Z still standing for C or CC). The effect on the set D is that if either node x or node y was a member of D then D is extended with the other node.

In the example of the inverter we initially have out $\notin$ D. As the program leaves the set D unchanged we have to show that it establishes CC(out) = 1. The first statement is legitimate as we initially have C(in) = 1 and

$$C(out) \wedge C(1) \wedge in' = 0 \wedge 1 \wedge in' = 0$$

The effect is that both C(out) and CC(out) become in'. The second statement is legitimate as well: C(in) is still 1 and

The above is a simple example of an RL, it does not have subRL's. The first line specifies the name of the component and its external ports. A question mark or an exclamation point indicates that the port is an input port or an output port, respectively. In the connection pattern two switches are specified, textually separated by a semicolon. The first statement expresses that the output port out is connected to the constant input port 1. The condition in front of the arrow specifies under which circumstances the switch in the connection should be on. In this case a normally-on switch whose gate is connected to the input port in (or a normally-off switch with its gate connected to in') is specified. The second statement specifies the second switch.

For the more pictorially inclined reader we observe the resemblance of the program and the following diagram.

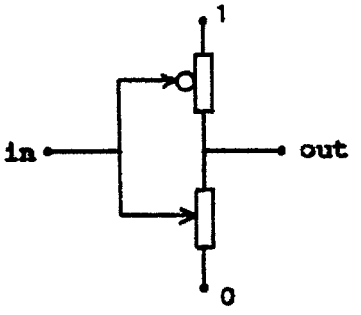


Fig. 8

Why is the program syntactically correct? In order to be able to show that the only output port out satisfies

$$out \in D \vee CC(out) = 1$$

we have to be more precise as to how a statement affects C(n), CC(n) and D.

In a program switches are introduced by statements

$$BE \rightarrow x = y$$

in which x and y are nodes, and BE is a boolean expression in terms of nodes, more precisely: BE is a production of the grammar

<boolean expression>::= <term> { ∨ <term> }

<term>::= <factor> { ∧ <factor> }

<factor>::= <primary> | <primary>'

<primary>::= <node> | (<boolean expression>)

$$C(out) \wedge C(0) \wedge in = in' \wedge 1 \wedge in = 0$$

It establishes CC(out) = in' ∨ in, which is 1. Hence, it is a complete program.

Notice that both switches in the inverter are of the type that can be implemented by one transistor. The inverter, consequently, requires only two transistors. We shall use this inverter as a subcomponent in our third example.

EXAMPLE 2.

```
comp nor(a?, b?, out!):
begin a ∨ b → out = 0; a' ∧ b' → out = 1 end
```

In the first statement the boolean expression is a disjunction of two nodes. This gives rise to a diagram in which two switches are placed in parallel. The boolean expression of the second statement specifies two switches that are placed in series. The whole component requires four transistors. The following picture shows a diagram of the component.

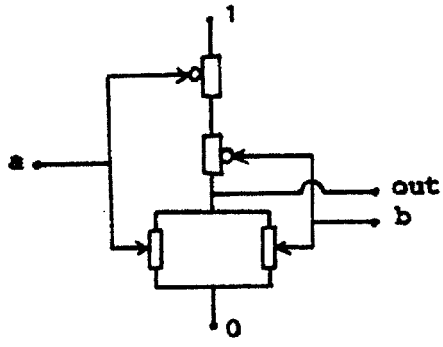


Fig. 9

A new node is introduced by mentioning it in the right-hand side (in the part to the right of the arrow) of a statement. There is no example of this in the paper.

EXAMPLE 3.

```
comp flip-flop(in?, ld?, q!, qbar!):
begin sub i1,i2: inverter;
      i2.in = i1.out;
      ld' → i1.in = i2.out; ld → i1.in = in;
      q = i2.out; qbar = i1.out
end
```

The second line of the program specifies that the component flip-flop has two subcomponents, named i1 and i2, of type inverter. As each inverter has two external ports, this declaration provides the component with four internal ports. An internal port that corresponds to the external port p of a subcomponent S is denoted as S.p. As both i1 and i2 have an external output port out, the component flip-flop has the internal input ports i1.out and i2.out. Likewise, it has the internal output ports i1.in and i2.in.

The reader is encouraged to check that the component satisfies the rules by formally deriving that all statements are legitimate and that the program establishes

$$q \in D,\ qbar \in D,\ C(i1.in) = 1,\ C(i2.in) = 1$$

A possible diagram of the component is

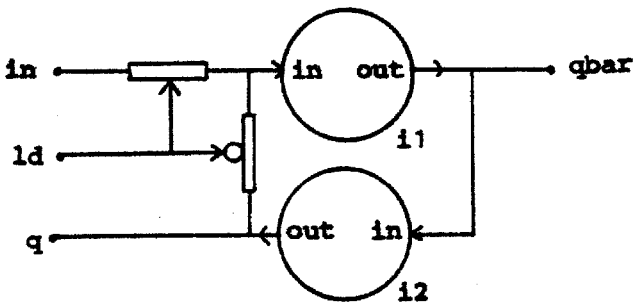


Fig. 10

5. BUSES

If we want to design a random access memory out of inverters, we must be able to connect their inputs and outputs via buses to the inputs and outputs of the memory. We want to connect the outputs of many subcomponents (inverters) to the same bus. Just connecting these outputs (internal inputs to the memory) to the bus would violate the no-fighting rule. We shall remedy this by putting switches in these connections.

To indicate when the memory cell has to drive the bus ("reading") and when it has to receive a value from the bus ("writing") two inputs, r and w, go into the cell:

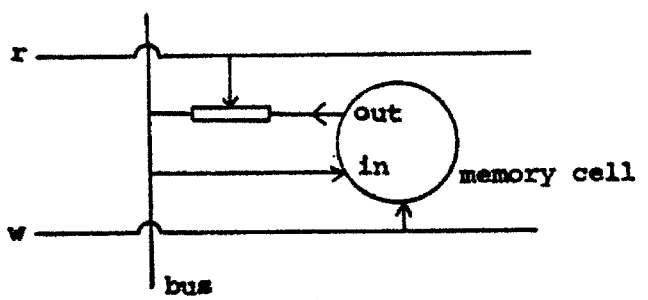


Fig. 11

We attach a number of cells to the same bus. Such a composition will only be an RL if we guarantee that, at most one of the cells can have its r equal to 1. The signals r come from another subcomponent of the memory, usually called the "decoder." The purpose of the decoder is to assure that at most one r equals 1. Given that the outputs of the decoder satisfy that requirement, we can show that the composition is again an RL. This is a new phenomenon: a condition on the values output

by a subcomponent has to be taken into account to prove that a connection pattern specifies an RL. We call such a check a semantic check.

The following program is a 1-of-2 decoder.

```
comp 1-of-2 decoder(in?, out1!, out2!):
begin in → out1 = 1; in → out2 = 0;
      in' → out1 = 0; in' → out2 = 1
end
```

By a syntactic check, as described in Section 4, we can show that this is a legitimate RL. In this case it is also simple to check that the output values satisfy (out1 ∧ out2) = 0, but that is a semantic check.

The moral is that we will design components that are only "conditional RL's," i.e., they are RL's under the condition that the output values of other components satisfy certain constraints. When such components are put together we will have to see to it that such semantic constraints are indeed satisfied.

6. A GLANCE INTO THE FUTURE OF COMPUTING

In this paper we have not addressed the dynamic behavior of components, i.e., how they react to transitions on their inputs. That is obviously the next step. By adopting proper timing and signaling conventions (cf. Chapter 7 of [2]) one should be able to address the dynamic behavior in an equally discrete fashion. The purpose of such conventions is to generate "data valid" inputs that signal that the input data are well-defined and may be inspected. Such a data valid signal may come from a clock or it may be an asynchronous acknowledge signal.

After that there are two roads we can follow. We can make a machine. That machine will accept programs and execute them. We then concentrate on the programs and if we wish to have a certain computation performed, we write a program for it. That is the traditional road.

We are led to the other, more promising, road if we observe that we are already designing programs, programs that can be compiled into transistor diagrams for CMOS. We make components out of subcomponents. Every time they will be more "powerful" or "sophisticated" than their subcomponents. We can inspect how a component is implemented by looking at its program text to see how it is composed out of subcomponents. Every component is again an implementation of a "higher level" concept. We can, e.g., introduce components that communicate other data types than just 0's and 1's. If we look at the implementation of that concept, we may notice that it is achieved by multiplexing or by the use of multiple ports. In that way the components we introduce will give us new modes of expression so that we can formulate our programs in terms of concepts that are more appropriate to our computations. After a while, we will have a mode of expression that one would customarily call a "higher level programming language."

Throughout all the levels of the hierarchy we have maintained that we program by composing components out of communicating sub-components. But by expressing a program in such a notation we have also specified an implementation for it, we have actually specified for the program a transistor diagram in CMOS. From there, the step to a complete silicon compiler is a (nontrivial) matter of generating the proper geometric representation of the transistor diagrams.

Of course, we do not have to translate all our programs into silicon to have them executed. We could also compile them into machine code, e.g., into code for a machine designed by taking the other aforementioned road. Our choice will depend on such external factors as the speed with which the computation has to be performed or the expected frequency of its use. It is also possible that we want to make a translation into machine code first in order to get some experience with the program and that we do not have it compiled into silicon until it is in a form that suits us.

POSTSCRIPT

Is this an article about machine design or about programming? The answer to that question is definitely "Yes!".

ACKNOWLEDGEMENTS

The research described in this paper was sponsored by the Defense Advanced Research Projects Agency, ARPA Order Number 3771, and monitored by the Office of Naval Research under contract number N00014-79-C-0597.

REFERENCES

[1] Johannsen, Dave, "Hierarchical Power Routing." Display file 2069, Computer Science Department, California Institute of Technology, Pasadena, CA, October 1978

[2] Mead, Carver & Lynn Conway, "Introduction to VLSI Systems." Addison-Wesley Publishing Company, Reading MA, 1980