Chapter 9
Leakage Power Reduction

9.1 Introduction

As was described in Chapter 3, in a short-channel nano-CMOS transistor, several forms of leakage current exist: reverse-biased diode leakage, subthreshold leakage, gate-oxide tunneling current, hot-carrier gate current, gate-induced drain leakage and channel punch-through current [308]. Of all these leakage mechanisms, gate-oxide (direct) tunneling current that flows during both active and sleep modes of a device is the most significant component for low-end nano-CMOS technology of 45 nm and below. Thus, the major sources of power dissipation in a nano-CMOS circuit can be summarized as dynamic and short-circuit current along with subthreshold and gate-oxide leakage [211]. In this chapter, reduction of gate-oxide leakage and subthreshold leakage of a CMOS data path circuit during high-level synthesis is discussed.

Why Gate-Oxide or Subthreshold Leakage Reduction During High-Level Synthesis? To make high-level synthesis suitable for nano-CMOS circuits, the aim is to develop models to capture gate leakage and optimize it during the high-level synthesis process. The behavioral level is not as highly abstracted as the system level nor as lowly abstracted as the gate or transistor level. Hence, at the behavioral level there is a balanced degree of freedom to explore power reduction mechanisms; and it can help in investigating lower-power design alternatives prior to circuit layout in actual silicon. Moreover, correct design decisions at early phases of circuit abstraction (like high level) will ensure that design errors are not propagated to lower levels, which may be costly to rectify. In a similar philosophy, subthreshold leakage reduction is beneficial if performed during behavioral synthesis.
9.2 Gate-Oxide Leakage Reduction

9.2.1 Dual-$T_{\text{ox}}$ Technique

In this section, reduction of total gate-oxide leakage of a CMOS data path circuit based on a dual-$T_{\text{ox}}$ technique during high-level synthesis is presented [216, 217]. The algorithm incorporates time constraints as a performance (or delay) trade-off factor and offers the user the choice of predetermining the performance of a circuit, vis-à-vis, power requirements. The algorithms consider an unscheduled DFG schedule for each of their nodes at appropriate control steps, and simultaneously binds them to the best available resource while considering resource constraints so as to achieve the desired performance with the minimum gate-oxide leakage.

9.2.1.1 Motivation

As the gate-oxide thickness decreases, the gate-oxide leakage current increases exponentially and delay decreases. It is believed that using higher gate-oxide thickness resources in off-critical paths and lower gate-oxide thickness resources in critical paths can obviate this effect. The extent to which the reduction can take place depends on both resource and time constraints. Using a larger number of higher-thickness resources implies added silicon cost. Thus, three important elements used to support design decisions are gate-oxide leakage power (current) dissipation, circuit performance (or critical path delay) and silicon cost (number of resources in the data path circuit to be generated).

The following can be concluded from the related background literature:

- Low-power behavioral synthesis research works have mostly considered dynamic power reduction. Only few behavioral synthesis works address subthreshold leakage. No high-level synthesis work is available in the current literature to optimize gate leakage. This calls for a new behavioral synthesis approach considering tunneling current for the low-end nanoscale technology (45 nm and below) domain.
- Behavioral synthesis works either focus on operation scheduling or on resource binding only. Behavioral scheduling algorithms minimize subthreshold leakage current without considering resource constraint (i.e., ignoring the silicon cost), whereas resource-binding algorithms take an already scheduled graph (ignoring the circuit delay).

Thus, simultaneous scheduling and binding are necessary to minimize the gate-oxide leakage current of the overall data path circuit while keeping performance degradation under control. Although performance can be incorporated as time constraints in the form of a delay trade-off factor (or implicitly through a current-delay product), the cost of silicon is taken into account as resource constraints.