

Pascal Meinerzhagen · Adam Teman
Robert Gitterman · Noa Edri
Andreas Burg · Alexander Fish

Gain-Cell Embedded DRAMs for Low-Power VLSI Systems-on-Chip

Gain-Cell Embedded DRAMs for Low-Power VLSI Systems-on-Chip

Pascal Meinerzhagen • Adam Teman
Robert Gitterman • Noa Edri • Andreas Burg
Alexander Fish

Gain-Cell Embedded DRAMs for Low-Power VLSI Systems-on-Chip

Pascal Meinerzhagen
Intel Labs, Circuit Research Lab
Intel Corporation
Hillsboro, Oregon, USA

Adam Teman
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel

Robert Giterman
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel

Noa Edri
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel

Andreas Burg
EPFL STI IEL TCL
Lausanne, Switzerland

Alexander Fish
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel

ISBN 978-3-319-60401-5 ISBN 978-3-319-60402-2 (eBook)
DOI 10.1007/978-3-319-60402-2

Library of Congress Control Number: 2017943168

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Runa and Elias

To Hadas and Shalev

To Tom, Daniel and Jonathan

Contents

1	Embedded Memories: Introduction	1
1.1	Increasing Need for Embedded Memories in Low-Power VLSI SoCs	1
1.2	Memory Requirements of Various Low-Power VLSI SoCs	4
1.3	Brief Review of the State of the Art	8
1.4	Book Outline	9
	References	10
2	Gain-Cell eDRAMs (GC-eDRAMs): Review of Basics and Prior Art	13
2.1	Basics of GC-eDRAM	13
2.2	Advantages and Drawbacks of GC-eDRAM	14
2.3	Review of Prior-Art GC-eDRAM Circuit Techniques and Target Applications	16
2.3.1	Categorization of GC-eDRAM Implementations	16
2.3.2	Comparison of the State-of-the-Art Implementations	18
2.3.3	Circuit Techniques for Target Applications	19
2.3.4	Summary and Conclusions	23
	References	24
3	Retention Time Modeling: The Key to Low-Power GC-eDRAMs	27
3.1	Introduction	27
3.2	Choice of Basic 2T GC-eDRAM Bitcell	28
3.3	Analytical GC-eDRAM Retention Time Model	30
3.3.1	Definition of Retention Time	30
3.3.2	Analytical Model of Nominal EDRT	31
3.3.3	Statistical Distribution of EDRT	32
3.4	Model Validation Through Circuit Simulations	33
3.4.1	Nominal EDRT	33
3.4.2	Statistical EDRT Distribution	35

3.5	Model Validation Through Silicon Measurements of 0.18 μm CMOS Test Arrays	36
3.5.1	Test Chip Design	36
3.5.2	Measurement Results	37
3.6	Sensitivity Analysis of GC-eDRAM Retention Time	41
3.6.1	Plackett-Burman Design of Experiment (PB-DOE)	41
3.6.2	PB-DOE Applied to GC-eDRAM RT	42
3.6.3	Impact of Process Corner	44
3.7	Best-Practice 2T GC Design	45
3.8	Conclusions	46
	References	46
4	Conventional GC-eDRAMs Scaled to Near-Threshold Voltage (NTV)	49
4.1	Introduction	49
4.2	2T GC, Array, and Macrocell Optimized for NTV Operation	51
4.2.1	2T Two-Port GC and Array Architecture	51
4.2.2	Operation Principle	52
4.3	Impact of Voltage Scaling on GC-eDRAM Retention Time	54
4.3.1	Worst-Case Access	54
4.3.2	Retention Mode	56
4.4	Macrocell Implementation Results	57
4.5	Conclusions	58
	References	59
5	Novel Bitcells and Assist Techniques for NTV GC-eDRAMs	61
5.1	Introduction	61
5.2	Single-Supply Transmission-Gate (TG) 3T-Bitcell GC-eDRAM	62
5.2.1	Proposed 3T TG Gain-Cell	63
5.2.2	Peripheral Circuits	66
5.2.3	Macrocell and Test Chip Design	69
5.2.4	Lab Setup and Silicon Measurements	71
5.3	Impact of Body Biasing (BB) on Retention Time	73
5.3.1	Bitcell Design for Body Biasing Experiment	74
5.3.2	Macrocell Architecture and Test Chip Design	75
5.3.3	Silicon Measurements	76
5.4	Replica Technique for Optimum Refresh Timing	78
5.4.1	Conventional Design for Worst-Case Retention Time	78
5.4.2	Replica Technique Concept	81
5.4.3	Replica Technique Integration into Gain-Cell Array	82
5.4.4	Testing and Characterization Procedure	84
5.4.5	Silicon Measurements	85
5.5	Conclusions	87
	References	89

6 Aggressive Technology and Voltage Scaling (Down to the Subthreshold Domain)	91
6.1 Introduction	91
6.2 Retention Time Model Validation for 28 nm CMOS	92
6.3 2T Gain Cells Optimized for Subthreshold Operation	93
6.3.1 2T Gain-Cell Implementation Alternatives	93
6.3.2 Best-Practice Write Transistor Implementation	96
6.3.3 Best-Practice Read Transistor Implementation	99
6.3.4 Storage Node Capacitance and WWL Underdrive Voltage	100
6.4 Macrocell Implementation in 0.18 μm CMOS	104
6.5 Macrocell Implementation in 40 nm CMOS	106
6.6 Conclusions	109
References	110
7 Novel Bitcells for Scaled CMOS Nodes and Soft Error Tolerance	113
7.1 Introduction	113
7.2 4T GC with Internal Feedback (IFB) for Scaled CMOS Nodes	114
7.2.1 Cell Structure and Operating Mechanism	114
7.2.2 Implementation and Simulation Results	117
7.3 Redundant 4T GC for Soft Error Tolerance	120
7.3.1 Radiation-Hardened Memories	121
7.3.2 Proposed 4T CDMR Dynamic Memory Array	123
7.3.3 Implementation	126
7.4 Conclusions	132
References	133
8 Conclusions	135
8.1 Summary	135
8.1.1 Near- V_T GC-eDRAM Techniques	136
8.1.2 Sub- V_T and Deeply Scaled GC-eDRAM Techniques	137
8.2 Outlook	138
Glossary	141
Index	143

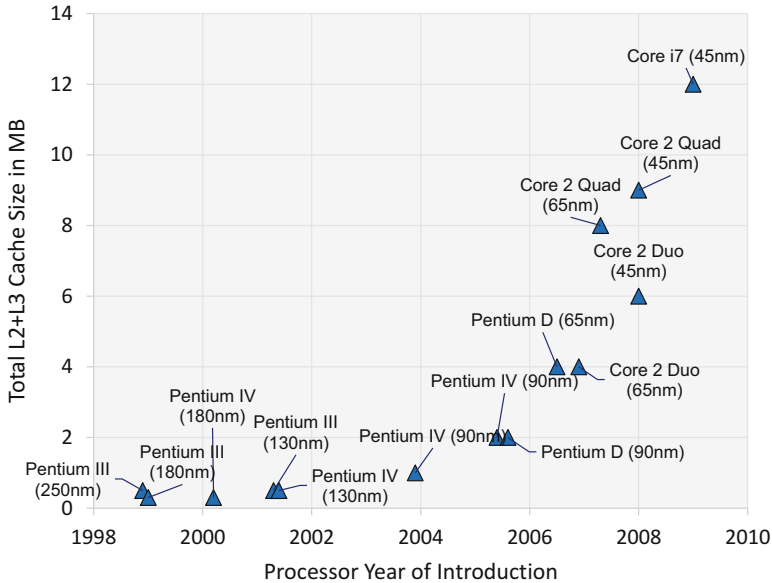
Chapter 1

Embedded Memories: Introduction

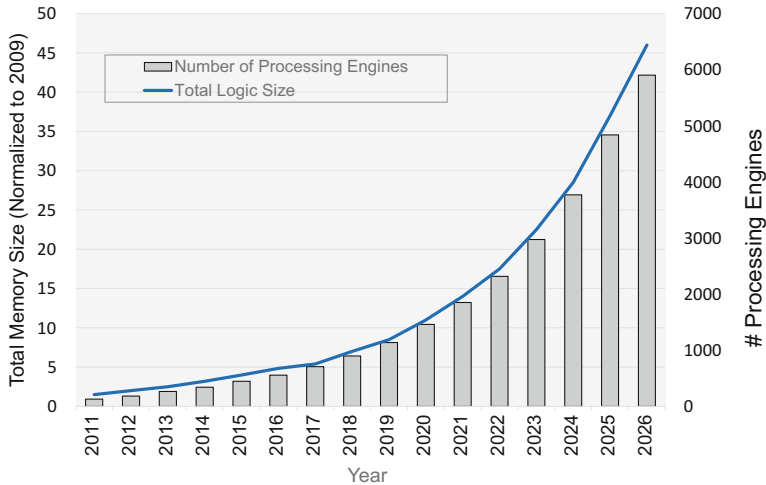
1.1 Increasing Need for Embedded Memories in Low-Power VLSI SoCs

There is a steadily increasing need for *embedded memories* in very large scale integration (VLSI) system-on-chip (SoC) designs targeted at microprocessors (used in servers, personal computers, laptop computers, tablets, and smartphones), biomedical implants, wireless communications systems, and many other applications. Such embedded memories are required to temporarily store data and/or instructions. From a system level perspective, it is clearly advantageous to always have more memory embedded directly on the compute chip, rather than relying on external memory chips. The primary reasons for this are: (1) embedded memories allow higher system-level integration densities, and (2) going off-chip through input/output (I/O) pads and capacitive lines on printed circuit boards (PCBs) entails severe speed and power penalties compared to on-chip connections [12]. As shown in Fig. 1.1a, the total cache size requirement in microprocessors has increased by around $5\times$ in a time interval as short as 4 years. In fact, back in 2005, an Intel® Pentium® D microprocessor used around 2 MB of cache memory, while the Intel® Core™i7, released in 2009, takes advantage of almost 10 MB of cache memory [17]. In accordance with this past trend of quickly increasing demand for embedded memories, the International Technology Roadmap for Semiconductors (ITRS) predicted in its 2011 Edition that the total embedded memory size for general SoC applications will increase by almost $50\times$ over the next 15 years [10], as shown in Fig. 1.1b.

As of today, embedded memories typically consume at least 50% of the total area and power budget of VLSI SoCs [10]. Figure 1.2 illustrates this by showing the layout pictures and the chip microphotographs of various VLSI SoCs, ranging from high-end microprocessors, through wireless communications systems, to ultra-low power (ULP) subthreshold ($\text{sub-}V_T$) microprocessors for health monitoring. The embedded memories, in the form of static random-access memory (SRAM)



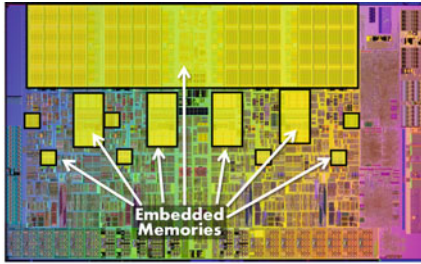
(a) Evolution of total cache size in microprocessors since 1998 [1].



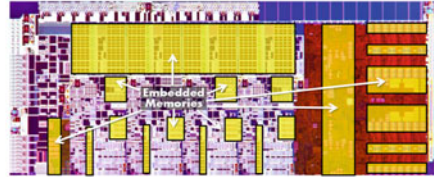
(b) Predicted evolution of total memory size in SoCs [2].

Fig. 1.1 (a) Past evolution, and (b) predicted future evolution of embedded memory size

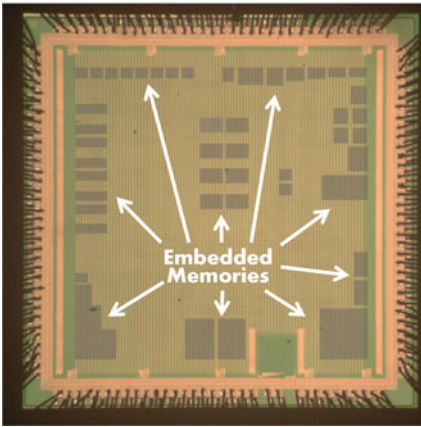
macrocells, are visible as regular layout tiles. Especially in case of the sub- V_T microprocessor, shown in Fig. 1.2d, the embedded memories, visible as yellow tiles, consume a dominant area share compared to the logic core which is in the center



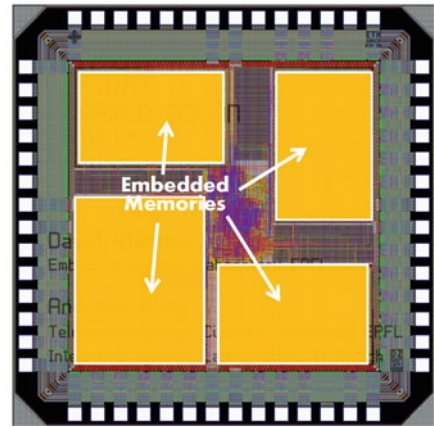
(a) Layout picture of 45 nm Intel® Core™ i7 processor (Nehalem) [4].



(b) Layout picture of 22 nm Intel® processor (a multi-CPU and GPU SoC) codenamed Ivy Bridge [9].



(c) Chip microphotograph of 4-stream 802.11n baseband transceiver [6].



(d) Layout picture of an ultra-low power, sub- V_T microprocessor for biomedical applications.

Fig. 1.2 Layout pictures and/or chip microphotographs of high-end microprocessors (a–b), a baseband transceiver (c), and a low-power processor for biomedical signals (d). All these VLSI SoCs require a significant amount of embedded memories, which are visible as regular tiles in the layout

of the chip. Furthermore, the 4-stream 802.11n baseband transceiver [4], whose chip microphotograph is shown in Fig. 1.2c, contains a large number of SRAM macrocells which are visible as dark areas.

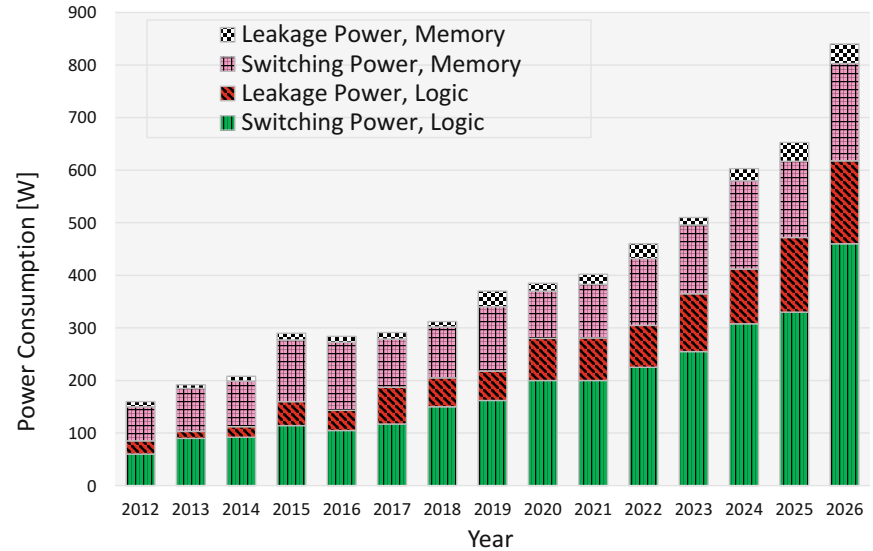
In addition to the large area share, embedded memories are also responsible for a large percentage of the power consumption of most VLSI SoCs. For example, the embedded memories of TamarISC-CS, a ULP application-specific processor for compressed sensing [6], consume 70–95% of the total power, depending on the mode of operation. As a further example, in a configurable high-throughput decoder for quasi-cyclic low-density parity-check (LDPC) codes [23], the embedded memories are responsible for 68% of the total power consumption. Furthermore, as

of today, VLSI SoCs for stationary applications typically have a total power consumption of up to 100 W, corresponding to the total of dynamic and static power consumptions of logic blocks and embedded memories [10], as shown in Fig. 1.3a. As opposed to this, Fig. 1.3b shows that VLSI SoC processors for portable applications have a considerably lower total power budget of 0.5 W, as per a requirement established by the ITRS in 2009. Only consumer processors for tablet computers may have a total power consumption as high as 2 W, given the physical product dimensions and advanced power management techniques [10]. For portable applications, the power consumption of embedded memories is expected to increase further and become almost 50% of the total power budget of processors in the next 15 years (see Fig. 1.3b). Reducing the power consumption of embedded memories is of utmost importance for all VLSI SoC application fields, for a number of quite different reasons. For example, low-power embedded memories and VLSI SoCs are essential to ensure runtimes of several years for ULP systems, such as implanted biomedical devices, to continue ensuring runtimes of ideally 1 day for portable computing devices of ever-increasing complexity (such as smartphones), or to reduce cooling costs for servers in data centers [8].

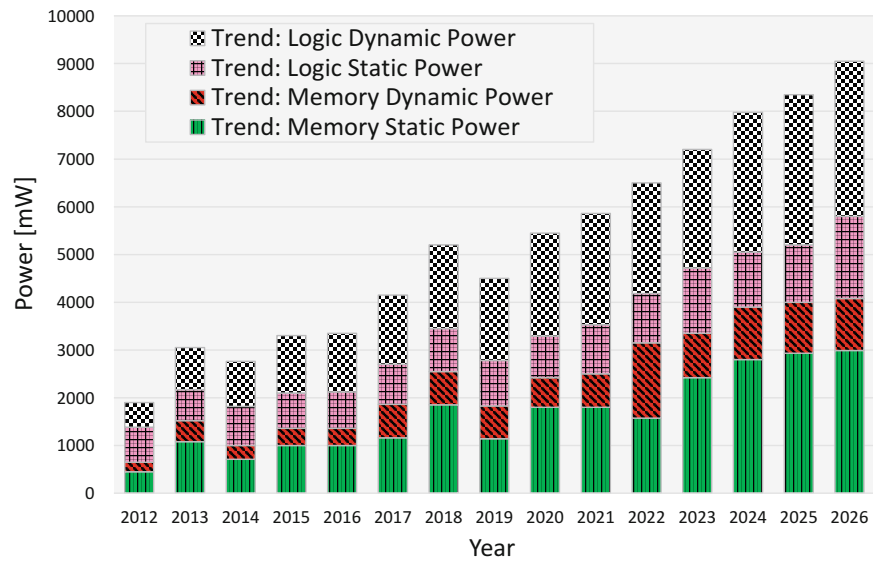
In addition to consuming dominant area and power percentages of VLSI SoCs, embedded memories are normally the first point of failure under voltage and technology down-scaling, due to the extremely high replication count of the basic bitcell, which, in most cases, is the 6-transistor (6T) SRAM bitcell. For example, if the supply voltage (V_{DD}) is scaled from its nominal value to the near-threshold (near- V_T) domain, the functional failure rate of embedded memories was shown to increase by five orders of magnitude [8]. As a consequence, under voltage and technology scaling, embedded memories typically limit the overall manufacturing yield of VLSI SoCs, whereas the complementary-metal-oxide-semiconductor (CMOS) based logic counterpart works more robustly.

1.2 Memory Requirements of Various Low-Power VLSI SoCs

Conventional personal computers and servers exhibit a deep memory hierarchy, ranging from on-chip, ultra-high speed, low storage capacity register files and cache memories, through fast, off-chip, higher capacity random-access memory (RAM), to slower, off-chip, high capacity, nonvolatile data storage. Traversing this memory hierarchy, the predominant, mainstream memory technologies are: (1) distributed or arrayed flip-flops or latches, (2) 6T-bitcell SRAM, (3) external, conventional 1-transistor-1-capacitor (1T-1C) dynamic random-access memory (DRAM), (4) Flash memory using a floating-gate transistor as a bitcell, and (5) mechanical hard disk drives, which are nowadays often replaced with solid-state drives. Note that only the register files and cache memories are embedded within the microprocessor chip, while the remaining parts of the computer memory hierarchy is off-chip.



(a) Power breakdown of stationary consumer SoCs [2].



(b) Power breakdown of portable consumer SoCs [2].

Fig. 1.3 Predicted power breakdowns of VLSI SoCs for (a) stationary, and (b) portable consumer electronics [10]

Beside servers, personal computers, and laptop computers, battery-powered mobile computing devices such as smartphones and tablet computers impose extremely challenging requirements on embedded memory solutions due to the increasing power awareness—required to extend the runtime on a single battery charge—accompanied by an ever increasing demand for higher integration density and higher speed performance.

Beyond microprocessors for computers, a large number of target applications in the broad field of VLSI SoCs often have diametrically opposite requirements on embedded memories. A comparison of such target applications is provided in Table 1.1. On the one hand, embedded memories in ULP VLSI SoCs for biomedical or remote sensing applications (such as [27, 28]) require ultra-low leakage power and access energy and entail significant engineering effort to ensure high robustness, while area and speed are secondary concerns. Therefore, such ULP VLSI systems, including their embedded memories, are often operated at ultra-low voltages (ULV), typically residing in the sub- V_T domain. On the other hand, power-aware high-performance VLSI SoCs, often used in wireless communications

Table 1.1 Memory requirements of different classes of VLSI SoCs, from ultra-low power to power-aware, high-performance systems

	Ultra-low power	Low-power, medium-performance	Power-aware, high-performance
Application fields	Biomedical implants, remote sensors	Near-threshold computing, complex sensor nodes, simple handheld devices	Wireless communications, tablet computers, smartphones
Robustness	Robust	Potentially unreliable (detect + correct, or error-resilient)	
Area priority	Secondary		High
Supply voltage V_{DD}	Subthreshold (sub- V_T), e.g., 400 mV	Slightly scaled, near-threshold (near- V_T), e.g., 600 mV	Nominal, e.g., 1 V
Power	Ultra low, fW–pW		High, mW–W
Speed	Very slow, kHz–MHz		Fast, 100 MHz–GHz
State of the art	Bistables (latches, flip-flops), pipeline registers		
	8T, 10T, . . . , 14T-bitcell SRAM, write and read assist	6T-bitcell SRAM, compilers 1T-1C eDRAM: special technology, extra cost Gain-cells: logic-compatible	
This book	Gain-cell eDRAMs (GC-eDRAMs)		
	2T sub- V_T	2T near- V_T	
	4T internal feedback	3T transmission-gate (TG)	
	4T redundant	Assist techniques: replica, body biasing	

(e.g., channel decoders) or in smartphones, require high-capacity, high-density, high-speed embedded memories operated at nominal supply voltages. In this case, rather than using robust, upsized SRAM bitcells, one-time programmable address decoders, if desired in combination with spare rows or columns to maintain storage capacity, are commonly used to cope with manufacturing defects (such as shorts and opens) [5]. Moreover, to cope with soft errors, caused by radiation occurring in the natural environment (e.g., alpha-particle impacts), redundant memory cells in conjunction with error detection and correction codes are often employed. A prominent example of such codes is the single-error-correction-double-error-detection (SECDED) code [11, 15]. Furthermore, as a new research direction, scientists and engineers have recently started to argue that the memory reliability can even be deliberately relaxed for VLSI systems which are inherently resilient to a small number of hardware defects. Examples of such inherently error-resilient systems include high-speed packet access (HSPA) systems [14] and wireless body sensor network (WBSN) nodes [20].

In addition to the above, an increasing number of VLSI systems feature dynamic voltage and frequency scaling (DVFS), in order to support different operating modes (such as high performance or low power modes), and/or reduce voltage and frequency guardbands for improved energy-efficiency and speed performance, respectively. Systems employing DVFS ideally contain embedded memories that are fully functional over the same voltage and frequency ranges as the logic. Besides the well-known Razor technique [9], as a further prominent example in the category of power-aware, high-performance VLSI SoCs supporting DVFS, Intel has presented an experimental, fully functional, error-resilient processor (codenamed Palisades) which has built-in mechanisms to detect and correct timing errors, allowing higher performance (by means of over-clocking) or better energy-efficiency (by means of voltage scaling) than a traditional processor with frequency and voltage guardbands [3].

In between the two extreme categories of ultra-low power VLSI SoCs operating in the sub- V_T domain and high-performance, power-aware, potentially error-resilient VLSI SoCs operating at nominal voltage, there is a third class corresponding to low-power, medium-performance SoCs (see Table 1.1). These SoCs and their embedded memories are typically operated at near- V_T supply voltages. Near-threshold computing (NTC) retains much of the energy savings of sub- V_T operation but has much better speed performance and suffers less from parametric variability [8]. An experimental, near-threshold voltage IA-32 microprocessor is able to successfully boot Windows XP™ while being supplied from a small solar panel providing only 10–20 mW of power [1, 18]. As a further example of NTC SoCs, Diet SODA [21] is a power-efficient processor for digital cameras relying on near-threshold circuit operation.

1.3 Brief Review of the State of the Art

Broadly speaking, embedded memories can be divided into two main categories: (1) SRAM and (2) embedded DRAM (eDRAM). SRAM uses a cross-coupled inverter pair to retain the stored data statically as long as a power supply voltage is provided. The eDRAM technology stores data in the form of electric charge on a capacitor; unfortunately, the stored data is compromised due to leakage currents, which results in a requirement for a periodic refresh operation.

As shown in Table 1.1, latches and flip-flops (mostly implemented as static storage cells) are commonly used as pipeline registers or also in the form of small, synthesized storage arrays distributed within datapaths [26]. Static latches and flip-flops operate reliably at a large range of supply voltages, including sub- V_T voltages [16]. Memory macrocells based on the conventional 6T SRAM bitcell can be used for all applications running at nominal or slightly scaled supply voltages. In fact, almost invariably, SRAM has been the mainstream solution for on-chip embedded memories for virtually all VLSI SoC target applications for the last few decades [12]. This unquestioned dominance of SRAM technology for on-chip storage mostly arises from their fast write and read accesses and their robust operation in mature CMOS nodes and at nominal supply voltages. Also, for most process nodes, SRAM memory compilers are readily available, facilitating their wide deployment. However, the footprint of the 6T SRAM bitcell is relatively large, since six transistors need to be accommodated. In order to increase the storage density, eDRAM macrocells are an interesting alternative to SRAM macrocells. We distinguish between two types of eDRAM: (1) conventional, one-transistor, one-capacitor (1T-1C) eDRAMs, whose basic bitcell is built from a special, high-density, 3D capacitor and a single access transistor, and (2) gain-cell eDRAMs (GC-eDRAMs) (e.g., [22]), whose basic bitcell is built from 2 to 4 MOS transistors [25]. Conventional 1T-1C eDRAMs typically require special process options to build high-density stacked or trench capacitors [13] and are therefore not compatible with the widespread standard digital CMOS technology. Such process options are only available at an extra manufacturing cost and are not readily available for all technology processes. As opposed to this, GC-eDRAMs are fully compatible with baseline digital CMOS technologies and can easily be integrated into any SoC at no extra cost. The main drawback of gain-cells is the small storage node capacitor (compared to the dedicated DRAM capacitors) and the resulting low retention time. From a functional perspective, all types of dynamic memories usually require data refresh cycles which are costly in terms of power and have a small access bandwidth penalty.

6T-bitcell SRAM fails to operate reliably at aggressively scaled supply voltages [24]. As shown in Table 1.1, alternative SRAM bitcells consisting of 8, 10, or even up to 14 transistors are required to ensure reliable sub- V_T operation [19]. In addition to large, alternative SRAM bitcells, various low-voltage write and read assist techniques have been proposed.