从 RTL 到 GDS 的功耗优化全流程

顾东华[1],曾智勇[1],余金金[1],黄徐辉[1],朱嘉骏[2],何湘君[2],陈泽发[2]
(1.燧原科技上海有限公司,上海 200000;2.上海楷登电子科技有限公司,上海 200000)

摘　要：功耗作为大型 SoC 芯片的性能功耗面积(PPA)三要素之一,已经变得越来越重要。尤其是当主流设计平台已经发展到了 7 nm 以下。AI 芯片一般会有多个核心并行执行高性能计算任务。这种行为会产生巨大的功耗。因此在 AI 芯片的设计过程中,功耗优化变得尤为重要。利用一个典型的功耗用例波形或者一组波形,可以从 RTL 进来开始功耗优化。基本的方式是借助 Joules-replay 实现基于 RTL 波形产生相对应的网表波形。在 Genus 的 syn-gen、syn-map、syn-opt 三个综合阶段,都可以加入 Joules-replay,并且产生和综合网表相对应的波形,用于 Innovus PR 阶段进一步地进行功耗优化。在 Innovus 中实现 Place 和 Routing 也分为 3 个阶段:place_opt、cts_opt 和 route_opt。同样每一步都可以引入 Joules-replay 来生成功耗优化所需的网表波形。最终在 Tempus timing signoff 的环境中,再次引入波形进行功耗优化。基于上面的一系列各个节点的精确功耗优化该设计可以获得 10%以上的功耗节省。此时再结合 multi-bit 技术,最终可以获得 21%的功耗节省。

关键词：功耗优化设计;人工智能芯片;芯片物理设计;Joules-replay;Genus;Innovus
中图分类号：TN402　　**文献标识码**：A　　**DOI**:10.16157/j.issn.0258-7998.229807

中文引用格式：顾东华,曾智勇,余金金,等. 从 RTL 到 GDS 的功耗优化全流程[J].电子技术应用,2022,48(8):65-69.
英文引用格式：Gu Donghua,Zeng Zhiyong,Yu Jinjin,et al. Fully power optimization flow from RTL to GDS[J]. Application of Electronic Technique,2022,48(8):65-69.

Fully power optimization flow from RTL to GDS

Gu Donghua[1],Zeng Zhiyong[1],Yu Jinjin[1],Huang Xuhui[1],Zhu Jiajun[2],He Xiangjun[2],Chen Zefa[2]
(1.Enflame Technology,Shanghai 200000,China;2.Cadence Design System,Inc.,Shanghai 200000,China)

Abstract: Power as one part of PPA(Performance, Power and Area) becomes more and more important in large SoC chips, especially under 7 nm technology. AI chips schedule multi-cores in parallel for specific application scenario, which lead to very large power consumption. Power optimization for each core is highest priority for an AI chip design. With a typical power scenario or multi-scenario grouped together, we can do power optimization from RTL synthesis to GDS. The basic flow is using Joules-replay to convert RTL activity file(time-based formats-VCD/FSDB/SHM/PHY) to gate level activity file. Synthesis with Genus has 3 steps: syn-gen, syn-map and syn-opt, Joules-replay is added after each step, and the replayed activity file will be used in power optimization in next step, which increase power estimation accuracy. Innovus place and route also has 3 main steps: place-opt, CTS-opt and route-opt, same flow with Joules-replay can be involved after each step, and it generates stimulus activity for next step. At final timing signoff stage, we use post-sim activity for power opt in Tempus. With this full flow power optimization flow, we can achieve more than 10% power reduction, combined with MBFF(Multi-Bit Flip-Flop) optimization, we can get 21% power reduction finally.
Key words: power optimization;AI chip design;SoC physical design;Joules-replay;Genus;Innovus

0 引言

芯片设计一直在追求最好的 PPA,在28 nm 之前的技术节点上,很多时候更多地优先考虑性能和面积。随着技术节点向 7 nm 进化,标准单元的密度不断提升,随之而来的功耗密度也越来越大。因此作为 PPA 之一的功耗在设计中变得尤为重要。设计芯片需要在流程的各个节点尽量对功耗进行精确评估并进行优化,否则最终芯片的性能很可能由于功耗过大而无法充分发挥。

1 芯片功耗

首先来看下从原理上芯片的功耗的计算方式。集成电路的功耗一般分为静态功耗和动态功耗。如图 1 所示,静态功耗又称为泄露功耗(leakage power),是指电路处于等待或不激活状态时泄漏电流所产生的功耗。

图 1 中箭头表明了在通电状态下 PMOS 内主要的泄漏电流及其走向,意即:

泄漏电流(Leakage Current)=漏极->N-Well+Gate->N-Well+源极->漏极

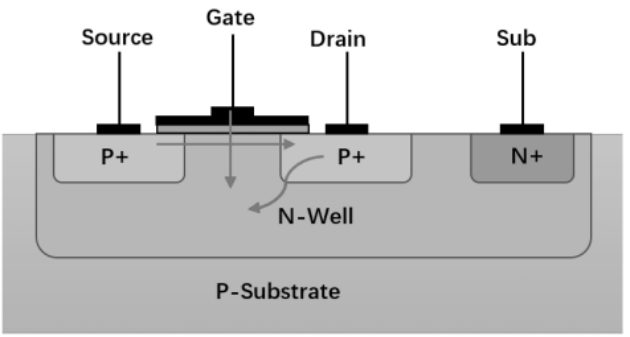

图 1　静态功耗示意图

泄漏电流存在的原因在于，MOS 管中的多种掺杂区形成导电区域，同时这些区域会组成多个 PN 节，从而在通电后形成一系列微小的电流。

尽管在现今芯片的工作电压已经很低的前提下每个 MOS 管的漏电流很小，但由于每颗芯片中集成的晶体管多至几亿甚至几百亿，积少成多，导致芯片的整体泄露功耗变得越来越恐怖。

在后端设计中，由于每个标准单元(standard cell)的 leakage 都集成在其 liberty 库文件(.lib)中，因此计算 leakage power 只需在制定条件下将 design 中所有的标准单元(包括各种 Macro)的 leakage 值相加即可。

目前所有的主流 PR 工具对此都有支持。需要指出的是，由于一个标准单元的 leakage power 和其面积成正比，因此在实际后端设计的各个阶段，尤其是 low power 设计中，一般会重点关注芯片中逻辑门的面积变化并以此快速推断 design 的 leakage 功耗变化。

另一部分称为动态功耗，是指芯片在工作过程中晶体管状态跳变所产生的功耗。当把反相器简化成一个简单的 RC 电路时，就可以清晰地看出充放电时的电流走向。当芯片处于工作状态时，每一个工作中的标准单元都会随着时钟以及数据的翻转而不断重复上述过程，从而产生大量的动态功耗。

在实际后端设计时，动态功耗由于和芯片的功能息息相关，因此在计算的时候会引入翻转率(toggle rate)的概念。翻转率是指单位时间内标准单元上信号翻转的次数，翻转率的高低直接影响到标准单元上的动态功耗开销。

在实际计算动态功耗的时候，又会分成两个部分。一部分为标准单元内部的动态功耗，一般叫做短路功耗，又可以称为内部功耗(Internal Power)，如图 2 所示，这部分的计算是嵌入 liberty 库文件内部，通过标准单元的 input transition 和 output load 来查表得到的。

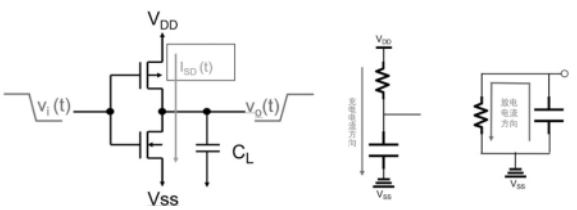

图 2　内部功耗示意图

另一部分为互连线(net)上的动态功耗，也称为翻转功耗(Switching Power)，这部分的计算通过将所有 net 上每个翻转周期的功耗乘以其翻转率并相加得到。

翻转率通过某种固定格式的文件传入 EDA 工具，比较常用的格式有 SAIF (Switching Activity Interchange Format)、VCD (Value Change Dump)以及 FSDB (Fast Signal Database)文件。

目前主流的 PR 工具均支持此类用法，但是签核 signoff 时仍然需要比较专业的 power 计算工具如 Synopsys PrimeTime PX 或 Cadence Palladium 等。

至此，我们基本了解了一颗芯片整体功耗的计算方法。而在现今十分重要的低功耗设计中，所有的手法都是从降低以上两个方面(Static，Dynamic)的功耗着手的：比如应用多个 power domain 以便在芯片的某一部分功能不用的时候将其断电关闭；或者通过升级更先进的工艺来降低每个晶体管的尺寸从而降低整体面积；抑或通过改善时钟树综合手段来降低芯片中占比很大的时钟网络功耗。

传统的功耗优化方案一般会采用减少 ULVT cell 的使用率来优化静态功耗，另外引入无向量模式(vector-less)设置一个大概的 switching activity 如 15%，然后进行动态功耗优化。但是这样的优化就要一定的随机性因此目标不明确，效果不明显。

在 7 nm 的 AI 类芯片中，动态功耗占据了主体，仅靠对于静态功耗的优化，无法满足功耗优化的目的。因此带入能表征芯片实际工作的工况波形，再进行精确的动态功耗优化更具有决定作用。

2 工况波形

一般模块设计者会对模块进行功能验证，某些工况下该模块的功耗会达到峰值。此时通过验证工具可以给出峰值功耗波形。该波形会记录该模块所有信号的翻转信息。这里的验证工具可以是 Cadence 的 Palladium(如图3 所示)。当然有些模块的峰值功耗可能有多个情况，并且会涉及不一样的逻辑空间，那么就需要一组波形来表征该模块的功耗行为。但是一般验证工具是基于 RTL 设计给出来的 RTL 波形，它虽然可以直接用于网表的优化，但由于它只能一定程度地映射网表中的寄存器，而无法精确匹配寄存器中间的大量组合逻辑。因此仅仅依赖 RTL 波型进行优化不能达到最优效果。

Cadence 在其功耗计算工具 Joules 中集成了 Joules-replay，该功能可以将 RTL 波形转换到与之相对应的门级网表(Gate Level Netlist)并且进行仿真，产生所有组合逻辑的详细波形。有了这一功能就可以在优化过程中使用更为精确的网表波形。

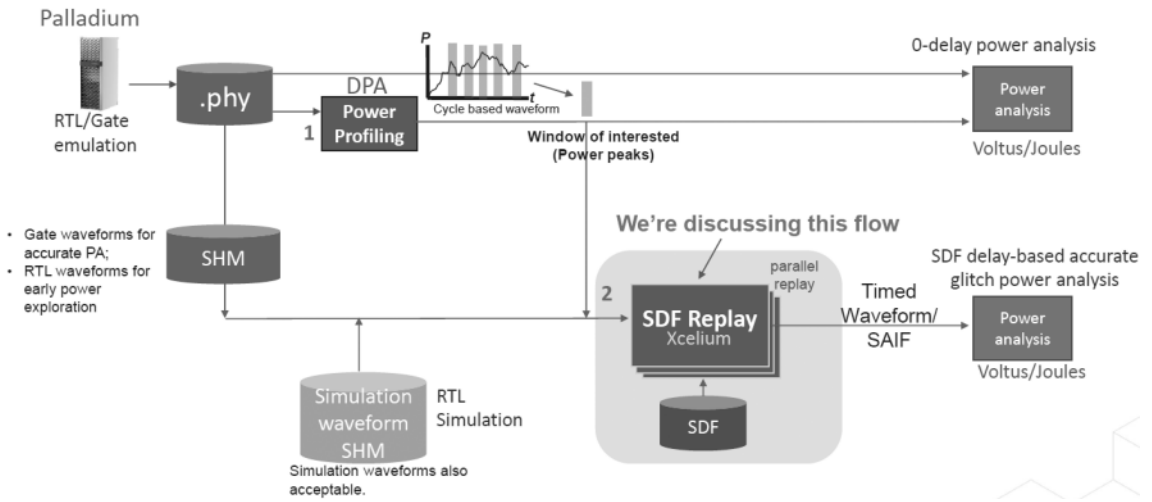

图 3 Palladium 示意图

3 功耗优化全流程

3.1 综合流程

当拿到 RTL 设计，并利用 Genus 进行综合时，可以利用对应的 RTL 波形开始进行功耗优化。Genus 综合可以分为 3 个步骤：syn_gen、syn_map 和 syn_opt。因此，可以形成如图 4 所示 Genus 结合 FSDB 的综合流程图。

· Genus：

– Syn_gen：read_stimulus –file operator_power_test.fsdb –dut * –start 27972.777ns–end 28047.837ns

– Syn_opt：read_stimulus –file syn_map_xreplay.vcd –start 27972.777ns–end 28047.837ns

3.2 PR 流程以及 signoff 阶段

如图 5 所示，在 Innovus 中实现 Place 和 Routing 也分为 3 个阶段：place_opt、cts_opt 和 route_opt。同样每一步都加入 Joules–replay 来实现优化所需的网表波形。同时可以通过设定 weight 值来实现引入多个波形同时优化。

· Innovus：

– Place_opt：

read_activity_file ../genus/syn_opt_xreplay.vcd –format VCD –scope /joules_testbench/joules_top_inst –start 27972.777ns–end 28047.837ns

read_activity_file –format TCF file1 –weight 0.3

read_activity_file –format TCF file2 –weight 0.2

read_activity_file –format TCF file2 –weight 0.5

– Cts_opt：

read_activity_file –reset

read_activity_file ../X–replay/placed_xreplay.vcd –format VCD –scope/joules_testbench/joules_top_inst –start 27972.777ns –end 28047.837ns

3.3 Signoff 流程

如图 6 所示，在 Signoff 阶段，通过 IPA 工具进行 power 的 signoff，可以了解到最终 power 的静态功耗和动态功耗的大小。引入 IPA 产生的 FSDB 至 IR 仿真中，得到 vector 的 IR 结果。

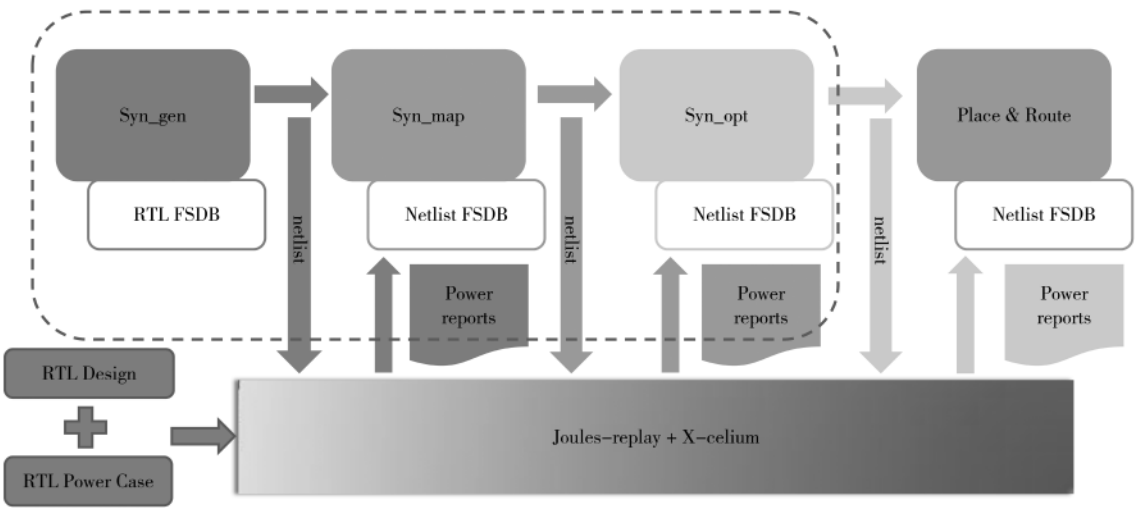

图 4 Genus 结合 FSDB 综合流程

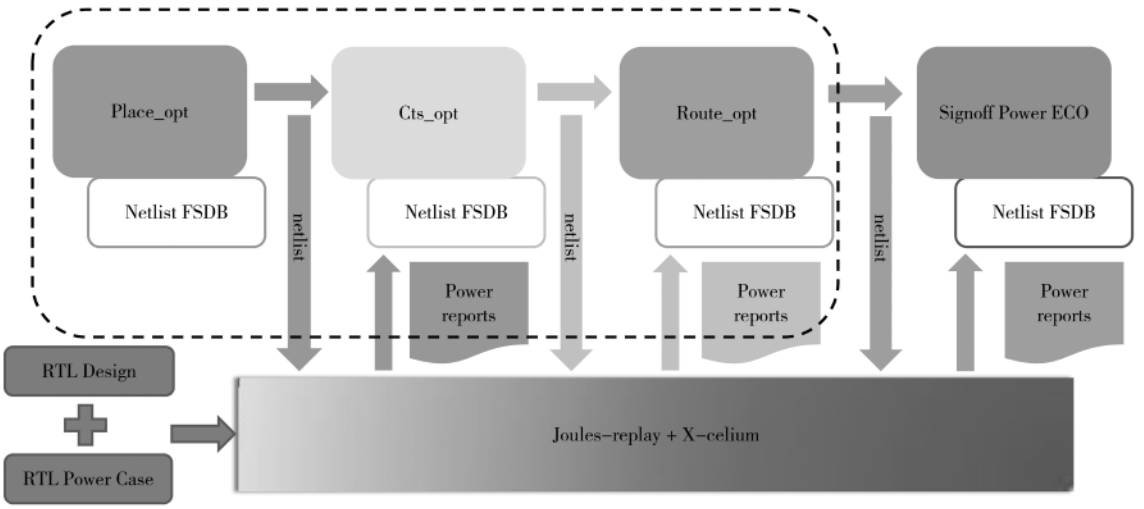

图 5 PR 各阶段结合 FSDB 功耗优化流程

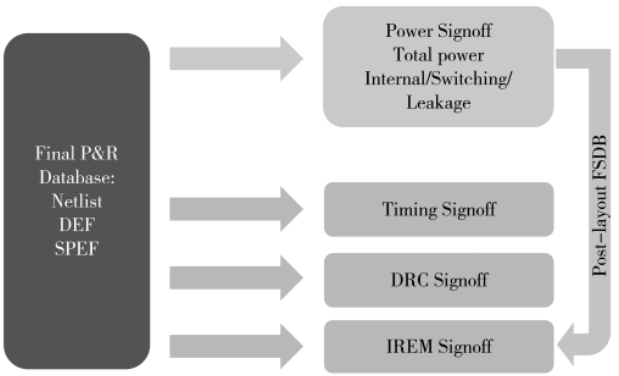

图 6 sigoff 流程图

4 功耗优化实验和结果

选取了三个 case 作为比较对象，如表 1 所示，case 0 为不引入 FSDB 的 base 情况，case 1 为由 Genus 综合开始至 Innovus PR 各阶段均引入 FSDB 的全流程功耗优化，case 2 为在 case 1 的基础上同时使用 MBB cell。

表 1 实验设定

	Flow Type	
	Genus	Innovus
Case 0	No FSDB	No FSDB
Case 1	FSDB	FSDB
Case 2	FSDB+MBB	FSDB+MBB

4.1 Signoff QOR 结果分析

如表 2 所示，对比 case0、case1 和 case2 的 QOR 结果可以发现：(1)三个 case 时序结果都比较一致，且在安全范围；(2)case1 和 case 2 的 DRC 数量有增多，但增加的 DRC 主要集中在 IO ports 附近，核心区域的 DRC 没有明显变化；(3)case 1 和 case 2 的 IR 结果变好，这是引入 FSDB 的 case，平均功耗更优化。总的来看，引入 FSDB 全流程功耗优化对 PR 的 QOR 影响不是很大，而且对 IR 还有提升。

表 2 实验 QOR 结果对比

	Case 0	Case 1	Case2
WNS/ns	-0.015	-0.029	-0.032
TNS/ns	-0.3	1.2	-0.2
NVP	56	139	20
Shorts	425	557	542
Total DRC	4 655	7 208	7 486
AVG-IR/%	4.23	3.79	3.39
IR VIO	6 323	2 292	2 081

4.2 各阶段功耗优化效果

Joules-replay 可以具体表征每个阶段功耗的数值，为此总结了引入 FSDB 后各阶段功耗优化的比例，如图 7 所示，在综合阶段有 7%的优化效果，当到达 signoff 时有 10.5%的优化，此外，当 FSDB 结合 MBFF cell 使用时，功耗优化效果会进一步提升，在签核阶段达到 19.8%的比例。

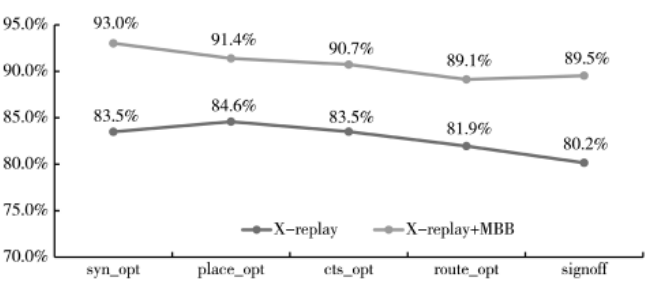

图 7 各阶段功耗优化的比例

表 3 列举了签核 signoff 阶段，这三个 case 的具体功耗组成，结果显示，case1 和 case 2 的功耗优化主要是由于内部功耗和开关功耗大幅度降低。

表 3 各 case 功耗组成占比

	Case 0	Case 1	Case2
总功耗/mW	1 117.2	1 000.8	895.5
比例/%	100	89.6	80.2
泄露功耗/mW	17.4	15.4	14.6
内部功耗/mW	519	476.5	414.3
开关功耗/mW	580.9	508.9	466.7

4.3 签核流程中功耗优化

为进一步优化功耗，在 Tempus timing signoff 的流程中进行 PowerECO，如图 8 所示，PowerECO 能够进一步优化 1.8%的功耗，最终全流程功耗优化能够达到 21%的优化效果。

5 结论

综上所述，本文使用了带 FSDB Genus 综合流程，带 FSDB Innovus PR 实现流程，以及 Tempus Power ECO 签核优化流程，并在整个实现与优化流程中结合 MBFF 技术，可以实现从 RTL 到 GDS 的 21%的功耗优化，这为大芯片的功耗优化带来全新的选择，为芯片的 PPA 的提升提供了一种全新的方法。

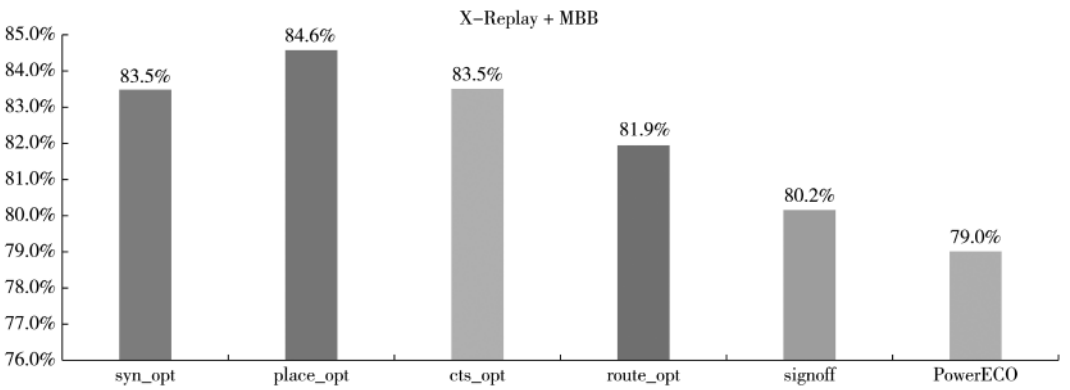

图 8 Joules-replay+MBB 各阶段功耗优化的比例

参考文献

[1] Cadence Innovus user guide[EB/OL].[2019-05-11].http://www.cadence.com.

[2] Cadence Genus user guide[EB/OL].[2021-10].http://www.cadence.com.

(收稿日期：2022-06-20)

作者简介：

顾东华(1982-)，男，硕士，硬件总监，主要研究方向：先进工艺下物理设计实现与优化、2.5D 等先进封装设计。

曾智勇(1991-)，男，硕士，芯片物理设计主管，主要研究方向：先进工艺下物理设计实现与优化、低功耗设计与优化。

余金金(1987-)，男，硕士，资深芯片物理设计主管，主要研究方向：先进工艺下物理设计实现与优化、2.5D 等先进封装设计。

扫码下载电子文档

(上接第 64 页)

[4] CUO X，STAN M R.Circadian rhythms for future resilient electronic systems: accelerated active self-healing for integrated circuits[M].Switzerland：Springer，2020.

[5] CACHO F，MORA P，ARFAOUI W，et al.HCI/BTI coupled model：the path for accurate and predictive reliability simulations[C]//2014 IEEE International Reliability Physics Symposium(IRPS).IEEE，2014.

[6] HUANG K，ZHANG X Q，KARIMI N.Real-time prediction for ic aging based on machine learning[J].IEEE Transactions on Instrumentation and Measurement，2019，68(12)：4756-4764.

[7] MISHRA S，AMROUCH H，JOE J，et al.A simulation study of NBTI impact on 14-nm Node FinFET technology for logic applications: device degradation to circuit-level interaction[J].IEEE Transactions on Electron Devices，2019，66(1)：271-278.

[8] KARAPETYAN S，SCHLICHTMANN U.Integrating aging aware timing analysis into a commercial STA tool[C]//VLSI Design，Automation and Test，2015.

[9] JAIN P，CANO F，PUDI B，et al.Asymmetric aging：introduction and solution for power-managed mixed-signal SoCs[J].IEEE Transactions on Very Large Scale Integration(VLSI) Systems，2014，22(3)：691-695.

(收稿日期：2022-06-20)

作者简介：

欧阳可青(1981-)，男，硕士，工程师，主要研究方向：先进工艺芯片物理实现方法学。

王彬(1996-)，男，硕士，工程师，主要研究方向：数字集成电路物理实现。

魏琦(1990-)，男，博士，工程师，主要研究方向：数字集成电路物理实现。

扫码下载电子文档

版权声明