Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 12, Number 4, Winter 2023, pp. 52-66
Received: 10/23/2023, Revised: 05/14/2024, Accepted: 07/08/2024

An Enhanced Replacement Policy to Mitigate Coherence-Induced Bias Temperature Instability in Cache Memories

Mohammad Maghsoudloo[1*]
1*- Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran.
[1*]mo.maghsoudloo@gu.ac.ir

Corresponding author's address: Mohammad Maghsoudloo, Faculty of Engineering, Golestan University, Gorgan, Iran.

Abstract- This paper proposes an enhanced replacement policy to mitigate the coherence-induced vulnerability of cache memories by static noise margin degradation prevention. The enhancement is conducted based on the outcomes of a comprehensive study intended to investigate the causes of static noise margin degradation in the SRAM cells. The empirical analysis demonstrates that the unbalanced distribution of data blocks associated with different coherency states over lines of a cache set can also be interpreted as a cause of the static noise margin degradation. Based on the findings, an aging-aware cache replacement policy is presented to balance the distribution of dirty/clean data blocks over the cache lines. To this intent, the decision tree of the Pseudo-LRU is revisited concerning the coherency state of cache lines and the type of address conflict miss. Using the enhanced cache, the hold and read static noise margin degradation are improved by about 9.9% and 11.5%, with less than 1.0% reduction in the cache hit ratio and negligible area and energy overheads.

Keywords- Cache memory, Bias temperature instability, Static noise margin, Hard errors, Cache replacement policy, Cache coherence protocol.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۲، شماره ۴، زمستان ۱۴۰۲ ، صفحه ۵۲-۶۶
دریافت: ۱۴۰۲/۰۸/۰۱، بازنگری: ۱۴۰۳/۰۲/۲۵، پذیرش: ۱۴۰۳/۰۴/۱۸

یک سیاست جایگزینی بهبودیافته جهت پیشگیری از ناپایداری‌های بایاس-دما در حافظه نهان ناشی از پروتکل‌های هماهنگی

محمد مقصودلو[1]*
*۱- گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی گرگان، دانشگاه گلستان.
[1*]mo.maghsoudloo@gu.ac.ir

* نشانی نویسنده مسئول: محمد مقصودلو، گرگان، خیابان شهید بهشتی، دانشگاه گلستان، دانشکده فنی و مهندسی گرگان.

چکیده- در این طرح، ساختاری بهبود یافته برای سیاست جایگزینی در حافظه نهان به منظور پیشگیری از وقوع آسیب‌پذیری‌های سالمندی ناشی از پروتکل‌های هماهنگی ارائه شده است. روش پیشنهادی بر اساس خروجی‌های یک مطالعه جامع بر روی دلایل وقوع کاهش حاشیه نویز ایستا سلول‌های حافظه ایستا با دسترسی تصادفی ارائه شده است. مطالعه جامع مذکور نشان می‌دهد که توزیع نامتوازن بلوک‌های داده با حالات هماهنگی متفاوت در بین خطوط حافظه نهان به عنوان یکی از دلایل وقوع کاهش حاشیه نویز در سلول‌های حافظه نهان به حساب می‌آید. بر اساس یافته‌ها، یک سیاست جایگزینی بهبود یافته برای حافظه نهان به منظور ایجاد توازن در توزیع بلوک‌های داده از نوع تمیز و کثیف بین خطوط مختلف حافظه ارائه شده است. در این راستا، سیاست جایگزینی شبه اخیراً کمتر استفاده شده بازطراحی شده است، به نحوی که در زمان خروج بلوک‌های داده، حالت هماهنگی بلوک‌ها و نوع بروز عدم برخورد ناشی از تناقض آدرس نیز لحاظ گردد. نتایج روش پیشنهادی نشان‌دهنده بهبودی حدود ۱۰ و ۱۱ درصدی میزان تنزل در روند کاهشی حاشیه نویز ایستا در زمان ذخیره‌سازی و در زمان خواندن در سلول‌های حافظه نهان می‌باشد. روش پیشنهادی منجر به سربار ناچیز مساحت و انرژی به همراه کمتر از یک درصد کاهش در نرخ برخورد حافظه نهان می‌گردد.

واژه‌های کلیدی: حافظه نهان، ناپایداری بایاس-دما، حاشیه نویز ایستا، خطاهای سخت، سیاست جایگزینی حافظه نهان، پروتکل هماهنگی حافظه نهان.

۱- مقدمه

با افزایش مقیاس‌پذیری فناوری نیمه‌هادی در ساخت ترانزیستور، نگرانی‌های مربوط به قابلیت اطمینان برای طراحی مدارهای مجتمع به‌طور فزاینده‌ای شدیدتر شده است [۱]. سالمندی ترانزیستور یک چالش قابلیت اطمینان رو به رشد در این فناوری در مقیاس نانو است که باعث خطاهای سخت‌افزاری دائمی (خطاهای سخت[1]) می‌شود [۲]. ناپایداری‌های بایاس-دما[2] پدیده سالمندی غالبی است که منجر به افزایش ولتاژ آستانه ترانزیستورها و کاهش تحرک حامل‌های بار می‌شود [۳،۴]. عوارض جانبی ناپایداری‌های بایاس-دما، که در بالا ذکر شد، می تواند تاخیر انتشار مدارهای ترکیبی را افزایش دهد [۵،۶] و حاشیه نویز ایستا[3] عناصر ترتیبی مانند حافظه با دسترسی تصادفی ایستا[4] را کاهش دهد [۷،۸]. شرایط عملیاتی مدار (شامل دما، بایاس ولتاژ و چگالی جریان) می‌تواند اثرات حاشیه نویز ایستا ناشی از ناپایداری‌های بایاس-دما را تشدید کند و منجر به آسیب‌پذیری‌های زودرس در مدار شود [۹].

حاشیه نویز ایستا معیار پایداری و قابلیت اطمینان برای استحکام سلول‌های حافظه نهان در برابر منابع خطاهای سخت افزاری است [۱۰]. در واقع، حاشیه نویز ایستا نشان‌دهنده حداکثر مقدار نویز ولتاژی است که می‌تواند توسط یک سلول حافظه بدون تغییر محتوای آن تحمل شود [۱۱-۹]. نگرانی‌های مربوط به قابلیت اطمینان سلول‌های حافظه نهان در میان مسائل مربوط به قابلیت اطمینان اجزای سخت افزاری مختلف اولویت بالایی درند [۱۴-۱۲].

حافظه‌های نهان روی تراشه[۵]، ساخته شده از سلول‌های حافظه نهان، بیش از ۶۰ درصد از مساحت تراشه را اشغال می‌کنند و حاوی بیشترین ترانزیستورهای یک ریزپردازنده مدرن در میان عناصر ترکیبی و ترتیبی مختلف مدار هستند [۱۵]. آنها می توانند به راحتی در معرض انواع خطاهای سخت و نرم قرار گیرند، خصوصا زمانی که کاهش حاشیه نویز ایستا پایداری سلول‌های حافظه نهان را تحت تاثیر قرار می‌دهد [۱۶]. کاهش حاشیه نویز ایستا ناشی از ناپایداری‌های بایاس-دما زمانی به حداکثر می‌رسد که یک مقدار ثابت در یک سلول حافظه نهان برای مدت زمان طولانی ذخیره شود [۱۹-۱۷]. یک سلول حافظه نهان معمولی شامل شش ترانزیستور MOSFET به شکل دو معکوس‌کننده جفت متقابل و دو ترانزیستور دسترسی است. در مورد ذخیره سازی "۰" یا "۱" به طور مداوم، یک PMOS و یک NMOS مستعد تحت تاثیر قرار گرفتن تغییرات آستانه ولتاژ ناشی از ناپایداری‌های بایاس-دما هستند [۱۴]. سنجش سالمندی ناشی از ناپایداری‌های بایاس-دما و پیشگیری از سالمندی ناشی از ناپایداری‌های بایاس-دما دو دسته اصلی روش‌های پیشین هستند، که به منظور کاهش اثرات جانبی کاهش حاشیه نویز ایستا ناشی از ناپایداری‌های بایاس-دما بر قابلیت اطمینان سلول‌های حافظه نهان ارائه شده‌اند.

روش‌های سنجش سالمندی ناشی از ناپایداری‌های بایاس-دما مبتنی بر قرار دادن حسگرها در آرایه‌های حافظه نهان برای نظارت بر رفتار زمان اجرای سیستم حافظه و اجزای آن برای یافتن نشانه‌های سالمندی هستند [۲۲-۲۰]. در سطح مدار، پارامترهای الکتریکی مختلف ترانزیستورها در یک سلول حافظه نهان به عنوان نشانگر استفاده شده است [۲۱،۲۰]. اما، دقت حسگرها به دلیل تغییرات فرآیند و سایر تغییرات پارامترهای محیطی به مرور زمان کاهش می‌یابد [۲۴،۲۳]. دیگر اشکال روش‌های سنجش سالمندی، پیچیدگی اصلاح سطح مدار سلول‌های حافظه نهان است [۲۵].

جلوگیری از ذخیره‌سازی مقدار ثابت در سلول‌های حافظه نهان برای مدت زمان طولانی به عنوان هدف اصلی غالب روش‌های طبقه‌بندی شده در دسته پیشگیری از سالمندی ناشی از ناپایداری‌های بایاس-دما لحاظ شده است [۲۸-۲۶]. در تمامی این روش‌ها، این هدف با متعادل کردن ضریب احتمال سیگنال[۶] مقادیر ذخیره شده در سلول‌های حافظه نهان و نزدیک کردن آن به ۵۰٪ به دست آمده است [۲۶]. نوشتن الگوهای بیتی خاص در قسمت‌های بلااستفاده حافظه نهان، معکوس کردن محتوای مقادیر ذخیره شده، تغییر طرح‌های آدرس‌دهی حافظه نهان و تغییر الگوی دسترسی حافظه نهان از نوآوری‌های کلیدی روش‌های طبقه‌بندی شده در این دسته می‌باشد [۲۹-۲۶]. در یک حالت ایده‌آل برای جلوگیری از کاهش حاشیه نویز ایستا، مقدار بیت ذخیره‌شده در سلول حافظه نهان باید یک متغیر تصادفی مستقل با شانسی برابر برای ذخیره‌سازی "۰" و "۱" باشد. در این حالت، توزیع احتمال ضریب احتمال سیگنال برای همه بیت‌های ذخیره شده در حافظه نهان در یک زمان مشخص، از توزیع نرمال با میانگین ۵۰٪ پیروی می کند [۲۹].

در این مقاله، در ابتدا یک تحلیل جامع برای بررسی توزیع احتمال ضریب احتمال سیگنال برای بیت‌های خطوط حافظه نهان ارائه شده است. از زیرساخت شبیه‌سازی عملکردی با کمک نسخه توسعه‌یافته ابزار gem5 [۳۱،۳۰] به عنوان محیط شبیه‌سازی و تحلیلگر عملکرد سیستم بهره گرفته شده است. همچنین، رفتار سه برنامه کرنلی (FFT، LU، و Cholesky) و سه برنامه کاربردی (Barnes، Ocean، و Raytrace) از مجموعه برنامه‌های بارکاری SPLASH-3 [۳۲] جهت ارزیابی روش‌های طراحی، مورد بررسی قرار گرفته است. تجزیه و تحلیل قابلیت اطمینان نیمه هادی اکسید فلز مبتنی بر HSPICE [۳۳] برای تجزیه و تحلیل سالمندی سلول‌های حافظه نهان ناشی از ناپایداری‌های بایاس-دما استفاده شده است. شبیه ساز CACTI 6.5 [۳۴] نیز برای تخمین مصرف انرژی بانک‌های حافظه نهان استفاده شده است. از روش مدل‌سازی سطح انتقال ثبات حافظه نهان و کنترل‌کننده آن نیز برای تجزیه و تحلیل مساحت پیاده‌سازی مدار استفاده شده است و از ابزار سنتز Synopsys Design Compiler بهره گرفته شده است [۳۵]. نتایج تجربی نشان می‌دهد که میانگین درصد فاصله بین توزیع مشاهده شده ضریب احتمال سیگنال و توزیع نرمال حدود ۸۸.۳ درصد است. علاوه بر دلایل کشف شده پیشین، این مقاله نشان می‌دهد که توزیع نامتوازن حالت‌های پروتکل‌های هماهنگی[۷] بر روی خطوط حافظه نهان نیز می‌تواند به عنوان دلیلی برای انحراف ذکر شده تفسیر شود.

بر اساس یافته‌های مرحله اول، این مقاله یک سیاست جایگزینی حافظه نهان[۸] آگاه از ناپایداری‌های بایاس-دما را جهت متوازن کردن شانس قرارگیری بلوک‌های داده با حالت‌های هماهنگی متفاوت بر روی خطوط حافظه نهان ارائه می‌دهد. برای این منظور، مدارات کنترل‌کننده حافظه نهان، به ویژه ساختار سیاست جایگزینی حافظه نهان، برای متعادل کردن مجموع فواصل زمانی بین دو نوشتن متوالی در خطوط مختلف حافظه نهان یک مجموعه، مورد بازبینی قرار گرفته است. حالت هماهنگی خطوط حافظه نهان و نوع عدم برخورد ناشی از تضاد آدرس[۹] دو عنصر کلیدی هستند که در روال تصمیم‌گیری توسط سیاست جایگزینی پیشنهادی استفاده شده‌اند. با استفاده از حافظه نهان پیشرفته، میزان کاهش حاشیه نویز ایستا در زمان ذخیره‌سازی و در زمان خواندن[۱۰] مقادیر حدود ۹/۹٪ و ۱۱٪/۵ بهبود می‌یابند. این بهبود با حدود ۰/۹۸٪ کاهش در نرخ

برخورد حافظه نهان، ۳/۳٪ افزایش در تعداد عملیات‌های بازنویسی[۱۱]، ۳/۸٪ افزایش مصرف انرژی پویا، و حدود ۱/۱٪ و ۱/۶٪ سربار مساحت از نظر تعداد سلول‌ها و اتصالات، همراه بوده است.

ادامه مقاله به شرح زیر سازماندهی شده است: پیشینه، و کارهای مرتبط در بخش‌های ۲ و ۳ مورد بحث قرار گرفته‌اند. انگیزه و ساختار سیاست جایگزینی پیشنهادی در بخش‌های ۴ و ۵ بیان شده است. در نهایت، بخش ۶ و بخش ۷ خلاصه نتایج و جمع‌بندی مقاله را ارائه خواهد کرد.

۲- پیشینه

اولین اثرات جانبی ناپایداری ولتاژ آستانه در ترانزیستورهای MOSFET در سال ۱۹۶۰ دیده شد، و تاثیرات جانبی این نوع ناپایداری به مرور و با پیشرفت فناوری ساخت ترانزیستور تشدید شد [۱]. این تشدید به دلیل افزایش دمای کاری تراشه و کاهش ابعاد ساخت ترانزیستور به وجود آمد [۲]. ناپایداری بایاس دما را می توان به دو دسته کلی تقسیم کرد: ناپایداری بایاس دما منفی[۱۲] که ترانزیستورهای جنس P را تحت تاثیر قرار می‌دهد و ناپایداری بایاس دما مثبت[۱۳] که ترانزیستورهای جنس N را متاثر می‌کند [۳].

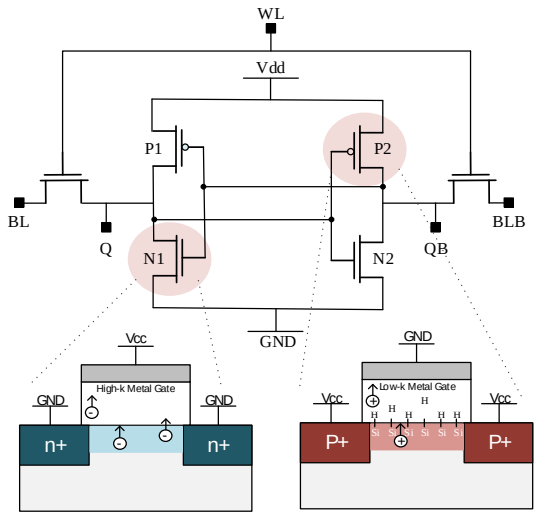
ناپایداری‌های بایاس-دما منفی نتیجه تولید تله پیوسته در رابط Si-SiO2 ترانزیستورهای MOSFET نوع P است [۲]. در ساختار MOSFET نوع P، پیوندهای نامتصل Si، به دلیل عدم تطابق ساختاری در رابط Si-SiO2 وجود دارند. این پیوندهای نامتصل به عنوان تله‌های سطحی باردار عمل می‌کنند. به طور معمول، غیرفعال‌سازی هیدروژن روی سطح Si جهت پیشگیری از مورد فوق اعمال می‌شود تا اتم‌های نامتصل Si را به پیوندهای Si-H تبدیل کند [۳]. با این حال، با گذشت زمان، این پیوندهای Si-H می‌توانند در حین کار شکسته شوند. پیوند شکسته شده به عنوان یک تله سطحی عمل می‌کند و ولتاژ آستانه ترانزیستورها را افزایش می‌دهد، و در نتیجه بر عملکرد ترانزیستور تأثیر می‌گذارد. اثرات ناپایداری‌های بایاس-دما منفی در فناوری‌های گذشته در مقایسه با ناپایداری‌های بایاس-دما مثبت توجه بیشتری را به خود جلب کرده است [۲]. امروزه، با معرفی فناوری دروازه فلزی، پدیده احتمالا مرتبط ناپایداری‌های بایاس-دما مثبت نیز به عنوان یک مسئله حیاتی برای پایداری MOSFET نوع N ظاهر شده است. ناپایداری‌های بایاس-دما مثبت می تواند به طور قابل توجهی ولتاژ آستانه MOSFETهای نوع N را تغییر دهد [۴].

حافظه‌های نهان روی تراشه، ساخته شده از سلول های حافظه نهان، درصد زیادی از مساحت تراشه را تشکیل می‌دهند و بیشترین ترانزیستورهای یک ریزپردازنده مدرن را دربردارند. پایداری سلول‌های حافظه نهان می‌تواند به راحتی توسط ناپایداری‌های بایاس-دما منفی و ناپایداری‌های بایاس-دما مثبت تحت تاثیر قرار گیرد [۵]. شکل ۱ معماری یک سلول حافظه نهان نمونه شش ترانزیستوری و شرایط استرس ناپایداری‌های بایاس-دما در MOSFETهای نوع n و نوع p را نشان می‌دهد. این معماری شامل معکوس‌کننده‌های CMOS متقاطع P1-N1 و P2-N2 است که برای ذخیره بیت، و دو ترانزیستور دسترسی برای انجام عملیات خواندن و نوشتن استفاده می‌شود [۶]. ترانزیستورهای دسترسی با استفاده از خط کلمه[۱۴] فعال/غیرفعال می‌شوند. خط بیت در حین انجام عملیات نوشتن به عنوان ورودی عمل می‌کند. با توجه به شکل ۱، یک MOSFET نوع n و یک MOSFET نوع p در یک سلول حافظه نهان همیشه در طول حالت عملیات ذخیره‌سازی مقادیر صفر و یک تحت استرس ناپایداری‌های بایاس-دما هستند [۶].

کارهای پیشین در راستای پیشگیری از وقوع پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما را می‌توان به دو دسته کلی تقسیم کرد: روش‌های مبتنی بر تشخیص پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما، و روش‌های مبتنی بر پیشگیری پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما.

۲-۱- روش‌های مبتنی بر تشخیص پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما

روش‌های مبتنی بر تشخیص پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما برای جلوگیری از خرابی فاجعه‌بار سایر بخش‌های



شکل ۱. معماری یک سلول حافظه نهان شامل ۶ ترانزیستور و شرایط وقوع استرس سالمندی ناشی از ناپایداری‌های بایاس-دما

سیستم پیشنهاد شده‌اند [۷-۱۲]. در این روش‌ها، از حسگرهای تشخیص‌دهنده پارامترهای الکتریکی استفاده می‌شود. یک حسگر می‌تواند برای تشخیص ناپایداری‌های بایاس-دما منفی، ناپایداری‌های بایاس-دما مثبت یا ترکیبی طراحی شود. یک حسگر مانند چراغ چک موتور در خودروها عمل می‌کند و توسط علائم حیاتی آسیب، فعال می‌شود [۷]. حسگرها اغلب محتویات حافظه نهان را پس از آزمایش تغییر نمی‌دهند. حسگرها برای نظارت بر رفتار زمان اجرا ساخته شده‌اند. در سطح مدار، پارامترهای الکتریکی مختلف MOSFETها در سلول‌های حافظه نهان به عنوان نشانگر استفاده شده است، مانند جریان نشتی زیرآستانه، ولتاژ آستانه، جریان تخلیه، ترانس رسانایی، سرعت افزایش و کاهش، تاخیرهای حاشیه، ولتاژ مرجع، و جریان خط-بیت. علاوه بر این، کدهای تشخیص/تصحیح خطا به عنوان یک حسگر برای تشخیص خطاهای نرم و سخت در سطح معماری مورد استفاده قرار گرفته‌اند [۹]. این روش‌ها با بررسی مجدد محتوای سلول‌های حافظه نهان، خطاهای سخت را از خطاهای نرم متمایز می‌کنند. اشکال اصلی روش‌های فوق، سربار قابل‌توجه مساحت و توان مصرفی و پیچیدگی تغییر ساختار سلول‌های حافظه نهان در سطح مدار است [۹].

۲-۲- روش‌های مبتنی بر پیشگیری پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما

این روش‌ها شرایط ایجاد ناپایداری‌های بایاس-دما قابل پیش‌بینی را در سلول‌های حافظه نهان برای کاهش وقوع پدیده سالمندی حذف می‌کنند [۱۳-۲۰]. جلوگیری از ذخیره‌سازی یک مقدار ثابت در سلول‌های حافظه نهان برای مدت طولانی، به عنوان هدف اصلی اکثر روش‌های طبقه‌بندی‌شده در این دسته در نظر گرفته شده است. قرار دادن بخش‌های غیرفعال حافظه نهان در حالت خواب‌آلود، خاموش کردن برخی سلول‌های خاص، روش‌هایی هستند که برای کاهش استرس سالمندی ناشی از ناپایداری‌های بایاس-دما بر روی سلول‌های حافظه نهان در نظر گرفته شده‌اند [۱۱]. روش‌هایی برای متعادل کردن ضریب احتمال سیگنال سلول‌های حافظه نهان تا ٪۵۰ تعیین شده‌اند، به طوری که حاشیه نویز ایستا به حداقل مقدار قابل قبول برسد. نوشتن الگوهای بیتی خاص بر روی سلول‌های بلااستفاده، معکوس کردن محتوای مقادیر ذخیره شده، تغییر طرح‌های آدرس‌دهی حافظه نهان، و تغییر الگوی دسترسی حافظه، نوآوری‌های کلیدی روش‌های ذکر شده برای تنظیم توزیع احتمال مقادیر ذخیره شده است [۱۱-۱۴]. این دسته از روش‌ها می‌توانند هزینه‌های سربار مساحت و توان مصرفی قابل توجه روش‌های تشخیص پدیده سالمندی را تعدیل کنند [۱۵].

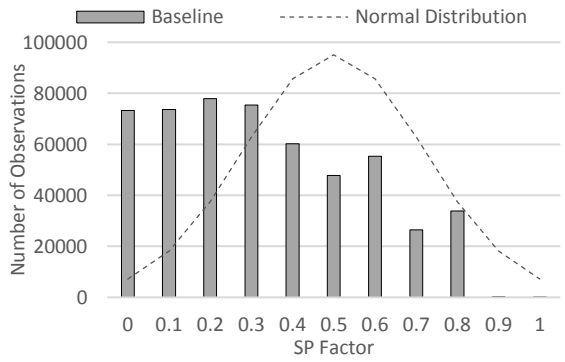
۳- روش پیشنهادی

۳-۱- انگیزش

در یک حالت ایده‌آل برای جلوگیری از کاهش حاشیه نویز ایستا، مقدار بیت ذخیره شده در یک سلول حافظه نهان باید یک متغیر تصادفی مستقل با شانس برابر برای ذخیره‌سازی صفر یا یک باشد [۲۱]. در این حالت، توزیع احتمال ضریب احتمال سیگنال برای تمام بیت‌های ذخیره شده در حافظه نهان در یک زمان مشخص، از توزیع نرمال با میانگین ٪۵۰ پیروی خواهد کرد [۲۲]. در تئوری احتمال، قضیه حد مرکزی بیان می‌کند که در بسیاری از موقعیت‌ها، زمانی که متغیرهای تصادفی مستقل جمع می‌شوند، مجموع نرمال شده مناسب آنها به سمت توزیع نرمال میل می‌کند، حتی اگر خود متغیرهای اصلی به طور معمول توزیع نشده باشند.

در ادامه، یک تحلیل جامع برای بررسی توزیع احتمال ضریب احتمال سیگنال برای بیت‌های حافظه نهان، انجام می‌شود. شکل ۲ توزیع احتمال ضریب احتمال سیگنال را برای ۵۲۴۲۸۸ بیت (۶۴ کیلوبایت) حافظه نهان نشان می‌دهد. لازم به ذکر است که این منحنی‌ها حاوی نتایج میانگین از همه بارهای کاری هستند و مقادیر با روش گرد کردن به پایین گرد شده‌اند. عنوان مشخص شده با برچسب "Baseline" در نمودارها به توزیع عملی ضریب احتمال سیگنال برای بیت‌های حافظه نهان تحت مدیریت مشخصات پایه ذکر شده در جدول ۱ اشاره دارد.

همانطور که در شکل ۲ نشان داده شده است، منحنی مبنا از توزیع نرمال با میانگین ۰/۵ پیروی نمی‌کند. میانگین درصد فاصله بین توزیع مشاهده شده و نرمال حدود ۸۸/۳ درصد است. ضریب احتمال سیگنال بیش از ۶۸ درصد بیت های حافظه نهان کمتر از ۰/۵ است. این مفهوم را می‌رساند که بیت‌ها به احتمال زیاد در طول دوره‌های



شکل ۲. توزیع احتمال ضریب احتمال سیگنال برای بیت‌های حافظه نهان ۶۴ کیلوبایتی

جدول ۱. مشخصات سیستم شبیه‌سازی شده

Processor	Cache Hierarchy
Frequency: 5 GHz Branch Predictor: 16K entry Bimodal and gshare Fetch/Issue/Commit: 4/4/4 Branch Penalty: 17 Cycles RAS: 32 entries BTB: 2K entries, 4way	IL1, DL1: 64KB, 4way Line size: 64B Coherency Protocol: MESI Replacement Policy: Tree-based P-LRU Write-back/Inclusive L2: 1MB, 4-Way Line size: 32B

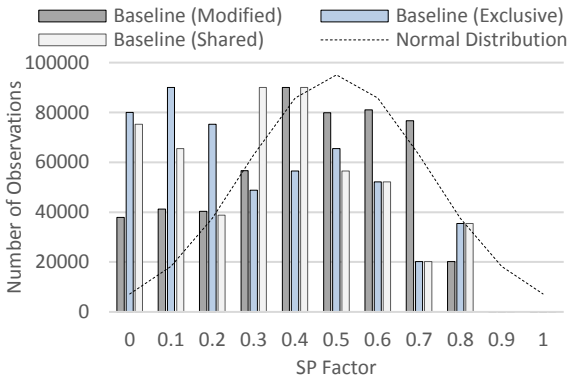
زمانی معتبر خود در حافظه نهان با مقدار صفر ثبت می‌شوند. دو دلیل برای این مشاهدات می‌توان بیان کرد: اصول محلیت داده و مقادیر با عرض کم .

بر اساس اصول محلیت، یک پردازنده تمایل دارد به مجموعه‌ای از مکان‌های حافظه به طور مکرر در یک دوره کوتاه دسترسی پیدا کند [۲۳]. برای عملکرد بهتر حافظه نهان، نگه داشتن بلوک‌های داده اخیر استفاده شده در طول زمان مفید است، اما این مساله کاهش حاشیه نویز ایستا را تسریع می‌بخشد. در نتیجه، مهم‌ترین بیت‌های هر کلمه، تنها پسوند بیت علامت یا صفر را برای بخش زیادی از عمر خود نگه می‌دارند [۲۴].

پس از بررسی بیشتر، متوجه شدیم که توزیع نامتعادل بلوک‌های داده مرتبط با حالت‌های مختلف پروتکل هماهنگی حافظه نهان نیز می‌تواند باعث شود که توزیع ضریب احتمال سیگنال بیت‌های حافظه نهان از توزیع نرمال منحرف شود. پروتکل هماهنگی حافظه نهان، پروتکلی است که هماهنگی و سازگاری را بین تمام حافظه‌های نهان سطح یکسان در یک سیستم حافظه مشترک توزیع شده حفظ می‌کند [۲۵].

شکل ۳ توزیع‌های عملی ضریب احتمال سیگنال برای بیت‌های خطوط حافظه نهان در هر یک از حالت هماهنگی مشخص را نشان می‌دهد: Exclusive، Shared و Modified. همانطور که در شکل ۳ مشخص شده است، مقادیر ضریب احتمال سیگنال برای خطوط حافظه نهان Modified از توزیع نرمال با درصد خطای کمتری نسبت به خطوط حافظه نهان Exclusive و Shared پیروی می‌کند. میانگین درصد خطا بین توزیع‌های مشاهده شده و نرمال ضریب احتمال سیگنال حدود ۲۸/۹ درصد برای بیت‌های خطوط حافظه نهان Modified است. در مقایسه، این مقدار برای بیت های خطوط حافظه نهان Exclusive و Shared به ترتیب تقریباً ۸۳/۳ درصد و ۹۹/۲ درصد است.

توالی عملیات نوشتن روی خطوط حافظه نهان در حالت Modified

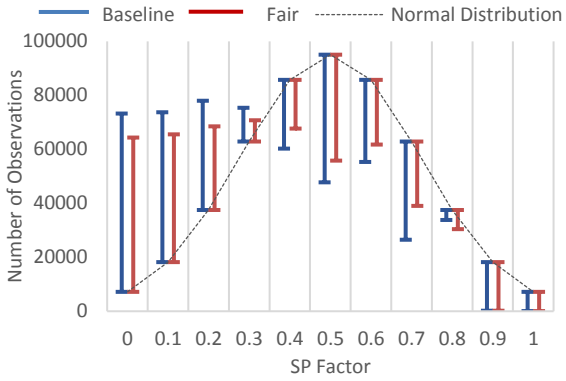


شکل ۳. توزیع احتمال ضریب احتمال سیگنال برای بیت‌های حافظه نهان ۶۴ کیلوبایتی با در نظر گرفتن حالات مختلف پروتکل هماهنگی

منجر به افزایش احتمال تغییرات در مقادیر بیت می‌شود. ضریب احتمال سیگنال بیش از ۸۵/۵ درصد از بیت‌های خطوط حافظه نهان در حالت Modified در محدوده [۰/۳، ۰/۶] است. در نتیجه، خطوط حافظه نهان Modified نسبت به خطوط حافظه نهان Exclusive و Shared کمتر در معرض خطاهای سخت ناشی از ناپایداری‌های بایاس-دما قرار دارند. علاوه بر این، نتایج تجربی نشان می‌دهد که بلوک‌های داده به‌طور متوسط حدود ۳۶/۶، ۳۳/۰ درصد و ۳۱/۳ درصد از طول عمر خود را در در حالت‌های Modified، Exclusive و Shared سپری می‌کنند. علاوه بر این، بیش از ۴۷/۸ درصد از بلوک های داده Modified به طور غیرمنصفانه در یک جهت مشخصی از خطوط حافظه نهان ذخیره می‌شوند.

ما آزمایش‌هایی را برای بررسی اثرات انواع مختلف روش‌های توزیع داده در میان خطوط حافظه نهان بر روی ضریب احتمال سیگنال طراحی کردیم. بر اساس مشاهدات ما، توزیع عادلانه بلوک‌های داده مرتبط با حالت Modified بر روی جهت‌های یک مجموعه حافظه نهان می‌تواند زمانی را که مقادیر ثابت در خطوط حافظه ذخیره می‌شوند، کاهش دهد و از کاهش حاشیه نویز ایستا جلوگیری کند. در نتیجه، با نادیده گرفتن سربار کارآیی، سیاست جایگزینی حافظه نهان را تغییر دادیم تا به هر خط از مجموعه شانس یکسانی برای ذخیره بلوک‌های داده‌ای که به دلیل عدم برخورد به منظور نوشتن وارد حافظه نهان می‌شوند، بدهیم. سپس، نتایج را بررسی کردیم.

شکل ۴ درصد خطا بین توزیع احتمال ضریب احتمال سیگنال برای همه بیت‌های حافظه نهان بعد و قبل از اصلاحات بیان شده در سیاست جایگزینی را نشان می‌دهد. عناوین با برچسب "Fair" به مقادیر خطای بین توزیع ضریب احتمال سیگنال برای بیت‌های حافظه نهان اشاره دارد که از توزیع عادلانه بلوک‌های داده مرتبط با حالت‌های مختلف پروتکل هماهنگی و توزیع نرمال پیروی می‌کنند.



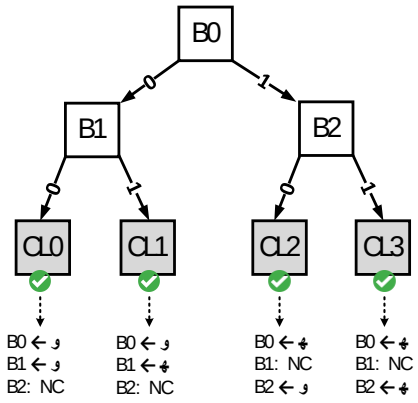
شکل ۴. درصد خطای بین توزیع ضریب احتمال سیگنال بیت‌های حافظه نهان در زمان استفاده از روش Fair و Baseline و توزیع نرمال

همچنین، عناوین با برچسب "Baseline" به مقادیر خطا بین توزیع‌های پایه و نرمال اشاره دارد. با در نظر گرفتن تفاوت بین توزیع ضریب احتمال سیگنال بیت‌های حافظه نهان قبل و بعد از اصلاح (نشان داده شده در شکل ۲ و شکل ۴)، میانگین درصد خطا بین توزیع‌های مشاهده شده و نرمال ضریب احتمال سیگنال حدود ۲۳/۷ درصد بهبود یافته است. این بهبود به ما انگیزه داد تا ساختار سیاست جایگزینی حافظه نهان را برای دستیابی به توزیع عادلانه‌تری از بلوک‌های داده مرتبط با حالت‌های مختلف پروتکل هماهنگی بر روی خطوط یک مجموعه حافظه نهان بازطراحی کنیم.

۳-۲- روش پیشنهادی

طراحان حافظه عموماً از سازماندهی حافظه نهان با بهره‌گیری از مدل مجموعه انجمنی[15] استفاده می‌کنند، زیرا این مدل توانایی ایجاد تعادل مناسبی بین نرخ برخورد و هزینه‌های پیاده‌سازی را دارد [۲۶]. با این حال، هزینه اجرای سیاست جایگزینی بهینه بسیار زیاد است. امروزه، اکثر حافظه‌های نهان از الگوریتم PLRU[16] استفاده می‌کنند تا از مضرات طراحی سخت‌افزاری پیچیده LRU اصلی جلوگیری کنند [۲۷]. به عنوان مثال، PLRU مبتنی بر درخت، یک روش جایگزین ترکیبی است که قدیمی‌ترین بلوک داده قابل دسترسی از یک مجموعه حافظه نهان را که احتمالاً اخیراً کمتر استفاده شده است (بر اساس اصل محلیت زمانی حافظه نهان) جایگزین می‌کند [۲۸].

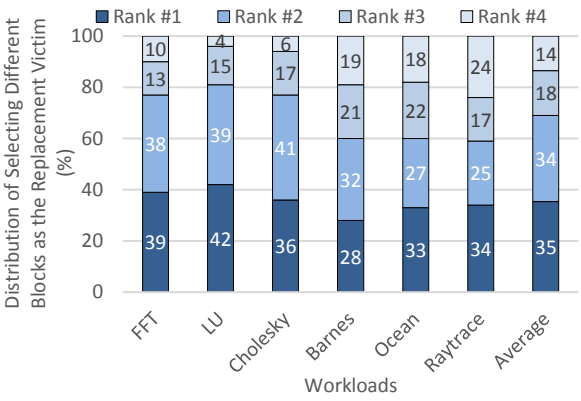
شکل ۵ ساختار PLRU مبتنی بر درخت را برای حافظه نهان ۴ جهته نشان می‌دهد. این شکل نشان می‌دهد که چگونه بیت‌های تصمیم (B0، B1 و B2) به جایگزینی بلوک‌هایی که اخیراً کمتر استفاده شده‌اند، کمک می‌کنند؛ و چگونه بیت‌های تصمیم‌گیری باید پس از بازدید از حاظفه نهان به دلیل خواندن/نوشتن در بلوک‌های حافظه، به‌روزرسانی شوند. PLRU مبتنی بر درخت شامل ذخیره



شکل ۵. روند تصمیم‌گیری تحت سیاست جایگزینی PLRU برای یک حافظه نهان چهارجهته

وضعیت حافظه در یک درخت باینری است. این درخت نشان می‌دهد که کدام خط از مجموعه حافظه نهان، بلوک داده‌ای است که باید در صورت عدم برخورد به دلیل تداخل آدرس جایگزین شود. بیت‌ها را می توان به صورت اشاره‌گرهایی به چپ یا راست در گره‌های فرزند تفسیر کرد [۲۲]. هر موقعیت درخت حاوی یک اشاره‌گر است که نشان می‌دهد کدام بلوک در صورت خطای خواندن/نوشتن هدف قرار خواهد گرفت. در صورت رخداد برخورد خواندن/نوشتن در حافظه نهان، بیت‌های موجود در مسیر از برگ مربوطه به ریشه به‌گونه‌ای مقداردهی می‌شوند که نشانگر از خطی که اخیراً به آن دسترسی شده است، فاصله بگیرد [۲۹]. همانطور که در شکل ۵ نشان داده شده است، سایر بیت‌های درخت بدون تغییر باقی مانده‌اند. پس از جایگزینی بلوک، الگوریتم روند تغییر همه گره ها از برگ تا ریشه درخت را تکرار می‌کند تا به بلوک جایگزین شده، اشاره نکند.

به منظور تعیین نزدیکی تصمیمات PLRU به LRU، آزمایش‌ها را گسترش دادیم تا شانس هر بلوک حافظه نهان در یک مجموعه برای انتخاب به عنوان کاندید جایگزینی توسط PLRU و LRU را با در نظر گرفتن سابقه دسترسی آنها ارزیابی کنیم. ما از بیت‌های تاریخچه برای پیگیری سفارش‌های دسترسی گذشته به خطوط حافظه نهان تحت مدیریت سیاست جایگزینی PLRU مبتنی بر درخت استفاده کردیم. سپس، تاریخچه دسترسی هر بلوک حذف شده را پس از اجرای بارهای کاری مختلف تجزیه و تحلیل کردیم. شکل ۶ توزیع انتخاب بلوک به عنوان قربانی جایگزینی در یک مجموعه نهان ۴ جهته را نشان می‌دهد. بلوک‌ها بر اساس سابقه دسترسی‌شان رتبه‌بندی شده‌اند و بلوکی که اولین/آخرین رتبه را دارد، دورترین/نزدیک‌ترین بلوک مورد دسترسی است. همانطور که در شکل ۶ نشان داده شده است، PLRU بلوک‌های دارای رتبه های

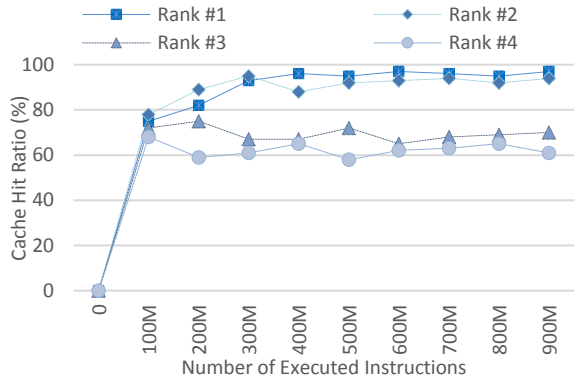


شکل ۶. نحوه توزیع بلوک‌ها با رتبه‌بندی‌های مختلف حذف شده با

اول، دوم، سوم و چهارم را به ترتیب به طور متوسط در حدود ۳۴/۹، ۳۴/۲، ۱۸/۴ و ۱۳/۸ درصد حذف می‌کند. بر اساس نتایج تجربی، بلوکی که رتبه دوم را دارد تقریباً شانس مشابه بلوکی که اخیراً کمتر استفاده شده است، برای قربانی شدن توسط PLRU را دارد. به طور کلی، نتایج نشان می‌دهد که بلوک داده جایگزین شده با PLRU کمترین استفاده اخیر را ندارد (بیش از ۶۵/۵ درصد موارد).

به منظور بررسی اینکه چقدر تصمیم‌های سیاست‌های جایگزینی مختلف می‌تواند بر نرخ برخورد حافظه نهان تأثیر بگذارد، آزمایشی انجام شد به طوری که بلوک‌های مختلف به اجبار به عنوان کاندید جایگزینی انتخاب شدند. شکل ۷ نرخ برخورد حافظه نهان شبیه‌سازی شده، پس از اجرای تعداد مشخصی از دستورالعمل‌ها تحت شرایط قربانی کردن بلوک‌های مختلف با رتبه‌های متفاوت در یک مجموعه حافظه نهان را نشان می‌دهد.

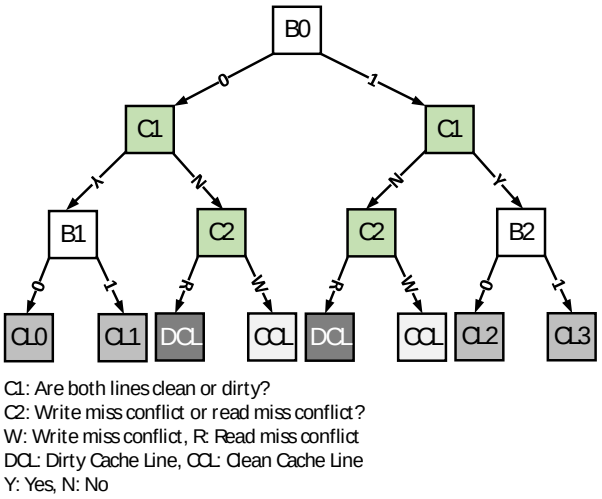
نتایج میانگین نرخ برخورد حافظه نهان برای همه بارهای کاری است. همانطور که در شکل ۷ نشان داده شده است، قربانی کردن بلوک‌های دارای رتبه‌های سوم و چهارم منجر به کاهش ۲۴٪ و ۳۱٪



شکل ۷. نرخ برخورد حافظه نهان پس از جایگزینی بلوک‌ها با رتبه‌بندی‌های مختلف

در نرخ برخورد حافظه نهان می‌شود. با این حال، زمانی که بلوک‌های دارای رتبه دوم به جای بلوک‌های رتبه اول جایگزین می‌شوند، نرخ برخورد حافظه نهان به طور قابل توجهی کاهش نمی‌یابد (کمتر از یک درصد). با توجه به شکل ۵، هر دو بلوک داده که رتبه‌های اول و دوم را دارند در انشعاب چپ یا راست درخت برای یک حافظه نهان چهار جهته قرار دارند. بنابراین، انتخاب یکی از برگ‌ها در انشعاب چپ (راست) درخت، اگر B0 برابر با ۰ (۱) باشد، تقریباً به نتایج مشابهی منجر می‌شود. در نتیجه، بلوک داده‌ای که اخیراً کمتر استفاده شده است لزوماً همیشه بهترین گزینه جایگزینی در میان بلوک‌های داده یک مجموعه حافظه نهان برای همه بارهای کاری نیست.

با بهره‌گیری از آزادی نسبی برای قربانی کردن بلوک‌های حافظه نهان که اخیراً استفاده شده‌اند، ساختار درخت تصمیم اصلاح شده است تا توزیع بلوک‌های حافظه نهان Modified را بر روی جهت‌های مختلف یک مجموعه حافظه نهان را متعادل کند. به عبارت دیگر، تمرکز اصلی سیاست جایگزینی بهبودیافته بر این اصل استوار است که دادن شانس تقریباً مشابهی به خطوط حافظه نهان برای ذخیره داده‌های کثیف/تمیز می‌تواند باعث کاهش زمان ذخیره مقادیر ثابت خطوط حافظه نهان و جلوگیری از کاهش حاشیه نویز ایستا شود. شکل ۸ ساختار بهبود یافته درخت تصمیم برای انتخاب نامزد جایگزینی در یک نمونه حافظه نهان چهار جهته را نشان می‌دهد. هدف اصلی اصلاح درخت این است که از ذخیره بلوک‌های داده با پتانسیل بالا برای اصلاح در یک مجموعه محدود یا ثابت از خطوط حافظه نهان جلوگیری شود. این هدف با قرار دادن دو گره به عنوان شرایط جدید (C1 و C2) در درخت تصمیم جایگزینی به دست می‌آید.



شکل ۸. درخت تصمیم در سیاست جایگزینی پیشنهادی

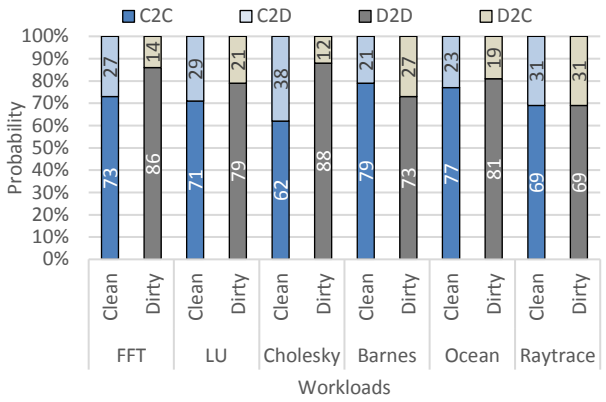
با توجه به شکل ۵، بلوک‌های داده‌ای که رتبه‌های اول و دوم را دارند، در برگ‌های انشعاب‌های چپ (راست) درخت تصمیم قرار دارند. در ساختار درخت تصمیم پیشنهادی (شکل ۸)، بیت‌های کثیف خطوط حافظه نهان ذکر شده قبل از بررسی مقدار بیت‌های B1 یا B2 بررسی می‌شوند. هنگامی که خط حافظه نهان مرتبط با آن پس از رخداد عدم برخورد پر شود، کنترل کننده حافظه نهان یک بیت کثیف را روی "۰" تنظیم می‌کند. اگر عملیات نوشتن بعدی، داده ها را در یک خط حافظه نهان تغییر دهد، بیت کثیف روی "۱" تنظیم می‌کند، که نشان می‌دهد مقدار موجود در حافظه نهان اکنون با مقدار حافظه اصلی متفاوت است.

پس از اینکه به دلیل تضاد آدرس، عدم برخورد خواندن/نوشتن حافظه نهان رخ داد، مقادیر بیت‌های کثیف خطوط حافظه نهان با رتبه اول و دوم، که با برگ‌های انشعاب‌های چپ یا راست (مقدار B0) درخت تصمیم به آن اشاره می‌کند، بررسی می‌شوند. در حین بررسی بیت‌های کثیف (C1 در شکل ۸)، وقوع دو پدیده ممکن است:

- هر دو خط حافظه نهان تمیز یا کثیف هستند: نمی توان استنباط کرد که یکی از خطوط حافظه نهان شانس بیشتری/کمتری برای ذخیره‌سازی داده‌ها با پتانسیل بالا برای اصلاح دارد. به منظور جلوگیری از کاهش حاشیه نویز ایستا، هدف مناسب‌تر جایگزینی را نمی‌توان با رعایت مقادیر فعلی بیت‌های کثیف نتیجه گرفت. متعاقباً، روال عادی PLRU برای کاهش اثرات جانبی منفی روش پیشنهادی بر نرخ برخورد حافظه نهان انجام می‌شود.

- یکی از خطوط حافظه نهان کثیف و دیگری تمیز است: می‌توان اطمینان داشت که خط حافظه نهان کثیف، بلوک داده را با پتانسیل بیشتری برای اصلاح در این زمان ذخیره کرده است. در این مورد، مقادیر بیت‌های کثیف و انواع عدم برخورد می‌توانند به تعیین کاندید جایگزینی مناسب‌تر برای متعادل کردن توزیع بلوک‌های داده کثیف بر روی خطوط و پیشگیری از کاهش حاشیه نویز ایستا کمک کنند. با توجه به شکل ۸، شرط C2 به جای بررسی مقدار B1 یا B2 برای انتخاب نامزد جایگزین در نظر گرفته می‌شود.

منطق استفاده از شرط C2 در درخت تصمیم بر اساس مشاهدات تجربی است که نشان می‌دهد بلوک‌های داده تمیز/کثیف تمایل دارند در طول عمرشان در حافظه نهان تمیز/کثیف بمانند. یک رخداد برخورد خواندن محلی یا از راه دور در یک بلوک حافظه نهان انحصاری یا مشترک (تمیز) بر نوع بلوک داده تأثیری نخواهد داشت. با این وجود، یک رخداد برخورد نوشتن محلی در یک بلوک داده مشترک یا انحصاری (تمیز) باعث انتقال به حالت Modified و تغییر



شکل ۹. احتمال تغییر مقادیر بیت‌های نشان‌دهنده حالت بلوک‌ها در حین اجرای بارهای کاری مختلف

مقدار بیت‌های کثیف می‌شود. علاوه بر این، رخداد عدم برخورد محلی خواندن/نوشتن در یک بلوک داده اصلاح شده منجر به حفظ مقادیر بیت‌های کثیف قبلی می‌شود. از سوی دیگر، خواندن از راه دور یک بلوک حافظه نهان اصلاح شده باعث دو عمل متوالی می شود: بازگرداندن داده‌ها به سطح پایین سلسله مراتب حافظه و ارسال داده‌های به روز شده به درخواست کننده. پس از آن، بیت کثیف تغییر می‌کند به نحوی که نشان دهد بلوک داده تمیز است.

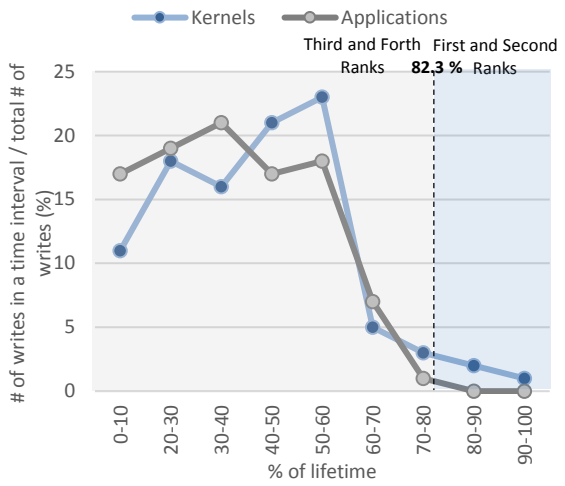
شکل ۹ احتمال انتقال بین انواع مختلف حالات بلوک حافظه نهان (تمیز و کثیف) را در طول اجرای هر بار کاری نشان می‌دهد. نتایج با نظارت بر تغییرات مقدار در فهرست پروتکل هماهنگی حافظه نهان و استخراج احتمال انتقال بین حالات هماهنگی مختلف مرتبط با خطوط حافظه نهان به دست آمده است. همانطور که توسط نتایج مشخص شد، یک بلوک داده، که به دلیل عدم برخورد تضاد آدرس، در هر خط حافظه نهان بارگذاری شده است، به طور متوسط در بیش از ۷۱/۸٪ موارد، تا زمانی که جایگزین شود، تمیز باقی می‌ماند. علاوه بر این، یک خط حافظه نهان پس از بازیابی بلوک داده از حافظه سطح پایین‌تر در اثر عدم برخورد تضاد آدرس، به‌روزرسانی می‌شود و به طور متوسط در بیش از ۷۹/۴٪ موارد، تا زمانی که به سطح پایین بازگردانده شود، کثیف باقی می‌ماند.

بر اساس یافته‌های فوق، شرط C2 برای متعادل کردن توزیع بلوک‌های داده تمیز/کثیف در خطوط مختلف یک مجموعه حافظه نهان در نظر گرفته شده است. برای این منظور، نامزد جایگزینی با در نظر گرفتن بیت‌های کثیف بلوک‌های دارای رتبه‌های اول و دوم انتخاب می‌شود و نوع پدیده عدم برخورد به دلیل تضاد آدرس، نیاز به جایگزینی را افزایش می‌دهد. در این مرحله دو مورد به شرح زیر قابل مشاهده است:

- مورد ۱: مشاهده عدم برخورد نوشتن به دلیل تضاد آدرس: در این حالت، خط حافظه نهان که یک بلوک داده تمیز را ذخیره می‌کند، انتخاب می‌شود تا بلوک داده جدید را با پتانسیل بالا برای اصلاح ذخیره کند.
- مورد ۲: مشاهده عدم برخورد خواندن به دلیل تضاد آدرس: در این حالت، خط حافظه نهان که یک بلوک داده کثیف را ذخیره می‌کند، انتخاب می‌شود تا بلوک داده جدید را با پتانسیل کم برای اصلاح ذخیره کند.

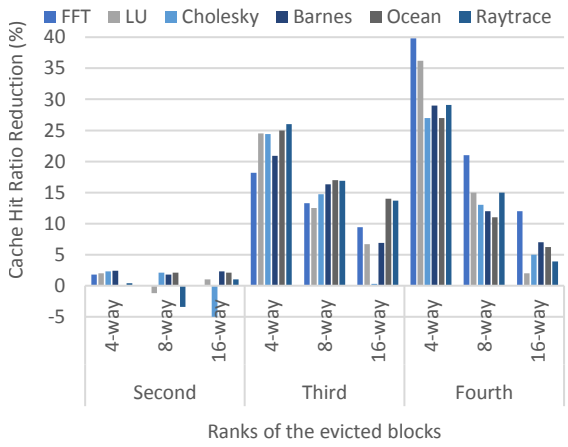
اقدام پیش‌بینی‌شده‌ای پس از وقوع مورد دوم برای افزایش شانس ذخیره بلوک‌های داده آینده با پتانسیل بالا برای اصلاح در خطوط حافظه نهان که قبلاً حاوی داده‌های تمیز بوده‌اند، در نظر گرفته شده است. علاوه بر این، نتایج تجربی نشان می‌دهد که بیشتر تغییرات در بلوک‌های حافظه نهان کثیف در ۵۰٪ اول عمر آنها در حافظه نهان ایجاد می‌شود. شکل ۱۰ میانگین درصد تعداد نوشتن‌های انجام شده بر روی بلوک‌های حافظه نهان اصلاح شده در بازه‌های زمانی مختلف طول عمر آنها را نشان می‌دهد. لازم به ذکر است که این منحنی‌ها حاوی نتایج متوسط از دو مجموعه بار کاری هستند: کرنل‌ها (FFT، LU، و Cholesky) و برنامه‌های کاربردی (Barnes، Ocean و Raytrace). به طور متوسط، حدود ۶۶/۸٪ و ۷۴/۷ ٪ از به‌روزرسانی‌ها در بلوک های حافظه نهان کثیف به ترتیب در ۵۰٪ اول طول عمر آنها در بارهای کاری هسته‌ها و برنامه‌ها انجام می‌شود. همچنین، بلوک‌های کثیف پس از ۸۲/۳٪ از عمر خود از رتبه‌های سوم/چهارم به رتبه‌های اول/دوم در رتبه‌بندی جایگزینی صعود می‌کنند (به طور متوسط برای بارهای کاری کرنل و برنامه). پس از این مرحله، تعداد دسترسی‌ها به بلوک‌های داده اصلاح شده با هدف نوشتن، به طور متوسط به ۳/۱٪ کاهش می‌یابد. این مفهوم را می‌رساند که اکثر بلوک‌های داده کثیف که رتبه‌های اول و دوم را در اختیار دارند تا زمان جایگزینی بدون تغییر باقی مانده‌اند. بنابراین، حذف زودهنگام بلوک‌های داده کثیف می‌تواند به جلوگیری از کاهش حاشیه نویز ایستا نیز کمک کند. حذف زودهنگام بلوک‌های داده می‌تواند منجر به کاهش کارآیی شود (زمانی که خطوط حافظه نهان نامعتبر بعداً توسط CPU مجددا مورد دسترسی قرار گیرند). با این حال، نتایج تجربی نشان می‌دهد که سربار کارآیی را می‌توان در بیشتر موارد تعدیل کرد زیرا خطوط حافظه نهان پس از عدم تغییر نامعتبر می‌شوند.

روش پیشنهادی به عنوان جایگزینی برای روش PLRU مبتنی بر درخت معرفی شده است، تا در یک حافظه نهان چهار جهته استفاده شود. با این حال، می توان آن را برای حافظه نهان در سطوح مختلف



شکل ۱۰. میانگین درصد تعداد عملیات نوشتن برروی بلوک‌های داده از جنس Modified

استفاده کرد. شکل ۱۱ کاهش نرخ برخورد حافظه نهان تحت شرایط حذف بلوک‌ها با رتبه‌های دوم تا چهارم را در یک حافظه نهان با خصوصیات متفاوت (۴ جهته، ۸ جهته، و ۱۶ جهته)، مقایسه می‌کند. در این آزمایش، بلوک‌های رتبه دوم، سوم و چهارم به اجبار به عنوان کاندیدای جایگزینی انتخاب شدند. نتایج ارائه شده در شکل ۱۱، میانگین کاهش نرخ برخورد حافظه نهان در مقایسه با حالتی است که بلوک رتبه اول خارج شود. با توجه به شکل ۱۱، نرخ برخورد حافظه نهان کمتر تحت تأثیر حذف بلوک‌های رتبه‌های دیگر با افزایش درجه انجمنی قرار می‌گیرد. با استفاده از حافظه نهان ۴ جهته، قربانی کردن بلوک‌های دارای رتبه‌های دوم، سوم و چهارم (به جای رتبه اول) به ترتیب منجر به کاهش ۱/۴، ۲۳/۵ و ۳۱/۳ درصدی در نرخ برخورد حافظه نهان می‌شود. این مقادیر به طور متوسط حدود ۰/۲، ۱۵/۱ و ۱۴/۵درصد برای حافظه نهان ۸ جهته



شکل ۱۱. مقایسه کاهش نرخ برخود در شرایط حذف بلوک‌ها در رتبه‌های دوم تا چهارم در حافظه‌های نهان با درجه انجمنی متفاوت

و ۰/۲، ۸/۵ و ۶/۰ درصد برای حافظه نهان ۱۶ جهته است. با افزایش درجه انجمنی حافظه نهان، انتخاب‌های بیشتری به عنوان جایگزینی برای بلوک داده‌ای که اخیراً استفاده شده است، وجود دارد تا با هدف جلوگیری از کاهش حاشیه نویز ایستا، جایگزین شوند. شرایط در نظر گرفته شده در ساختار درخت تصمیم پیشنهادی را می‌توان به طور مشابه برای مدیریت شانس ذخیره‌سازی بلوک های داده با پتانسیل بالاتر اصلاح در میان خطوط حافظه نهان مجموعه‌ای با درجه بالاتر انجمنی گسترش داد. در این مورد، خطوط حافظه نهان با رتبه پایین را نیز می‌توان به همراه خطوط اول و دوم در نظر گرفت.

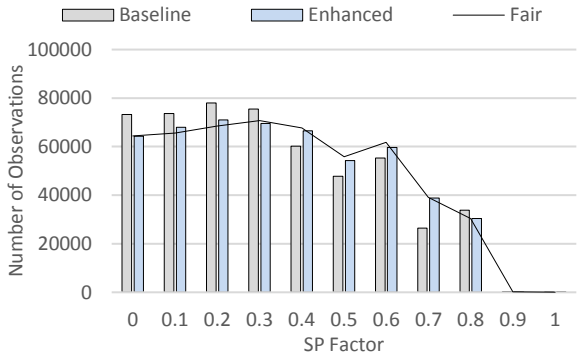
۴- نتایج

به منظور ارزیابی تصمیم‌های طراحی قبلی و جدید، یک زیرساخت شبیه‌سازی عملکردی با نسخه توسعه‌یافته از gem5 استفاده شده است [۳۱،۳۰]. سیستم شبیه‌سازی شده یک پردازنده چندهسته‌ای (۱۶ هسته) است که بر اساس مدل CPU RAW سازماندهی شده است. RAW یک معماری با ساختار موازی را پیاده‌سازی می‌کند و اتصال، ورودی/خروجی، حافظه و عناصر محاسباتی آن را در دسترس کامپایلر قرار می‌دهد. همچنین، رفتارهای سه برنامه کرنلی FFT، LU و Cholesky و سه برنامه کاربردی Barnes، Ocean و Raytrace از مجموعه برنامه‌های کاربردی SPLASH-3 در محیط شبیه‌سازی مورد مطالعه قرار گرفته است [۳۲]. SPLASH-3 مجموعه‌ای از برنامه‌های موازی بر اساس نیاز ماشین‌ها با فضای آدرس اشتراکی حافظه نهان، نوشته شده است [۳۲]. تصمیمات طراحی از نظر چهار پارامتر مورد بررسی قرار می‌گیرند: پیشگیری از وقوع پدیده سالمندی، سربار کارآیی، مصرف انرژی و مساحت مصرفی. همه نتایج با توجه به نتایج معماری پایه که در فصل سوم و جدول ۱ توضیح داده شده است، نرمال شده‌اند.

۴-۱- پیشگیری از وقوع پدیده سالمندی

آنالیز قابلیت اطمینان نیمه هادی اکسید فلز[۱۷] برای تجزیه و تحلیل وقوع پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما سلول‌های حافظه نهان نانومتری قبل و بعد از اعمال تغییرات در مدارهای کنترل‌کننده حافظه نهان استفاده شده است. دو فاز شبیه‌سازی مجزا، پیش استرس[۱۸] و پس استرس[۱۹]، برای تکمیل تحلیل وقوع پدیده سالمندی با استفاده از MOSRA ضروری هستند، که می‌تواند در اجرای شبیه‌سازی یکسان یا مستقل انجام شود [۳۳]. نتایج شبیه‌سازی پیش استرس مقادیر یک دستگاه قبل از نمایان شدن علائم سالمندی را نشان می‌دهد. مقدار استرس محاسبه شده ابتدا برای یک زمان مشخص توسط کاربر ادغام می‌شود و سپس برای رسیدن به استرس کل در پایان زمان مشخص شده برون‌یابی می‌شود. نتایج شبیه‌سازی پس از استرس، مقادیر دستگاه در معرض پدیده سالمندی، از جمله مقادیر کاهش‌یافته برای هر مرحله زمانی مشخص را نشان می‌دهد. بر اساس مقدار کل استرس محاسبه شده توسط فاز شبیه‌سازی پیش استرس، HSPICE استرس را به صورت تغییر پارامترهای مدار مدل می‌کند. مقدار تغییرات ولتاژ آستانه ترانزیستورهای سلول‌های حافظه نهان را می‌توان در دماهای مختلف با استفاده از مدل وقوع پدیده سالمندی داخلی MOSRA تخمین زد. چرخه کاری[۲۰] ترانزیستورهای سلولی حافظه نهان به عنوان یک عامل وابسته به بار کاری در مدل‌سازی وقوع پدیده سالمندی فوق مورد نیاز است. به عنوان تابعی از ضریب احتمال سیگنال سلول حافظه نهان، چرخه کاری، زمان استرس و زمان بازیابی ترانزیستورهای سلول حافظه نهان با کمک مدل MOSRA محاسبه می‌شود که مقادیر پارامترهای مورد نیاز آن را می توان از زیرساخت شبیه‌سازی عملکردی (gem5) استخراج کرد.

به عنوان اولین گام، توزیع احتمال ضریب احتمال سیگنال برای بیت‌های حافظه نهان پس از استفاده از سیاست جایگزینی بهبودیافته بررسی می‌شود، و میانگین درصد خطا بین توزیع‌های مشاهده شده و عادی با نتایج حافظه نهان پایه مقایسه می‌شود. شکل ۱۲ توزیع احتمالی ضریب احتمال سیگنال بیت‌های حافظه نهان را در زمان استفاده از معماری‌های پایه و بهبودیافته نشان می‌دهد. منحنی توزیع احتمالی منصفانه در نمودار گنجانده شده است تا مقایسه واضحی از نظر میانگین درصد خطا بین نتایج به‌دست‌آمده و توزیع منصفانه ارائه دهد. با توجه به شکل ۱۲، میانگین درصد خطا بین نتایج حافظه نهان بهبودیافته و توزیع‌های نرمال حدود ۷۰/۸٪ است که نشان دهنده بهبود ۱۹/۸٪ در مقایسه با میانگین درصد خطا بین توزیع‌های پایه و عادی است. علاوه بر

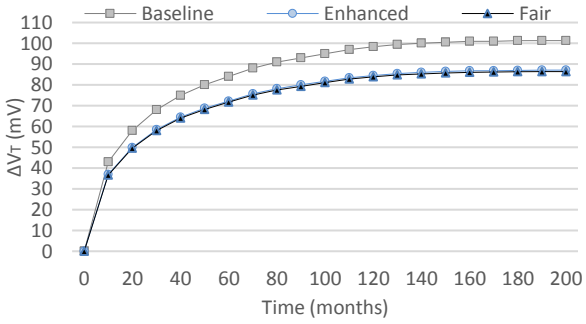


شکل ۱۲. توزیع احتمالی ضریب احتمال سیگنال بیت‌های مختلف حافظه نهان بعد از اعمال روش پیشنهادی در مقایسه با روش پایه

این، میانگین درصد خطا بین نتایج حافظه نهان بهبودیافته و توزیع عادلانه حدود ۱۶/۴ ٪ است که نشان‌دهنده اثربخشی سیاست جایگزینی پیشنهادی در دستیابی به توزیع تقریباً منصفانه بلوک‌های داده با پتانسیل بالا برای اصلاح در بین خطوط حافظه نهان است.

به عنوان گام دوم، مقدار تغییر ولتاژ آستانه (ΔVT) ناشی از ناپایداری‌های بایاس-دما با توجه به مدل وقوع پدیده سالمندی داخلی MOSRA مورد مطالعه قرار می‌گیرد. این مدل برای تجزیه و تحلیل پیری ناشی از ناپایداری‌های بایاس-دما در سلول‌های ۳۲ نانومتری حافظه نهان قبل و بعد از اعمال تغییرات به مدارهای کنترلی حافظه استفاده شده است. دو فاز جداگانه شبیه‌سازی، پیش از تنش و پس از تنش، ضروری است تا تجزیه و تحلیل پیری با استفاده از MOSRA کامل شود که می‌تواند همزمان یا جداگانه اجرا شوند. نتایج شبیه‌سازی پیش از تنش مقادیر یک دستگاه جوان را نشان می‌دهد؛ و نتایج شبیه‌سازی پس از تنش، مقادیر یک دستگاه تحت تاثیر سالمندی را نشان می‌دهد. براساس مقدار کل تنش محاسبه شده توسط فاز شبیه‌سازی پیش از تنش، HSPICE تنش را به تنزل پارامترهای مدار ترجمه می‌کند. مقدار تغییر ولتاژ آستانه ترانزیستورهای سلول حافظه را می‌توان با استفاده از مدل پیری داخلی MOSRA در دماهای مختلف برآورد کرد. چرخه کاری ترانزیستورها به عنوان یک آرگومان وابسته به بارکاری توسط مدل پیری فوق استفاده می‌شود. به عنوان یک تابع از ضریب احتمال سیگنال سلول حافظه، این مدل چرخه کاری را در زمان‌های تنش و بازیابی ترانزیستورها در نظر می‌گیرد، که می‌توان از زیرساخت شبیه‌سازی عملکردی استفاده شده (gem5) استخراج کرد.

شکل ۱۳ میانگین مقدار ΔVT برای کل ترانزیستورهای سلول‌های حافظه نهان یک خط حافظه نهان با فناوری ۳۲ نانومتری به مدت ۲۰۰ ماه را نشان می‌دهد. نتایج بر اساس چرخه کاری ترانزیستورهای سلول‌های حافظه نهان استخراج شده از gem5، به دست آمده است. با توجه به شکل ۱۳، تغییرات ولتاژ آستانه



شکل ۱۳. مقادیر تغییرات ولتاژ آستانه ترانزیستورهای سلول‌های حافظه نهان طی ۲۰۰ سال

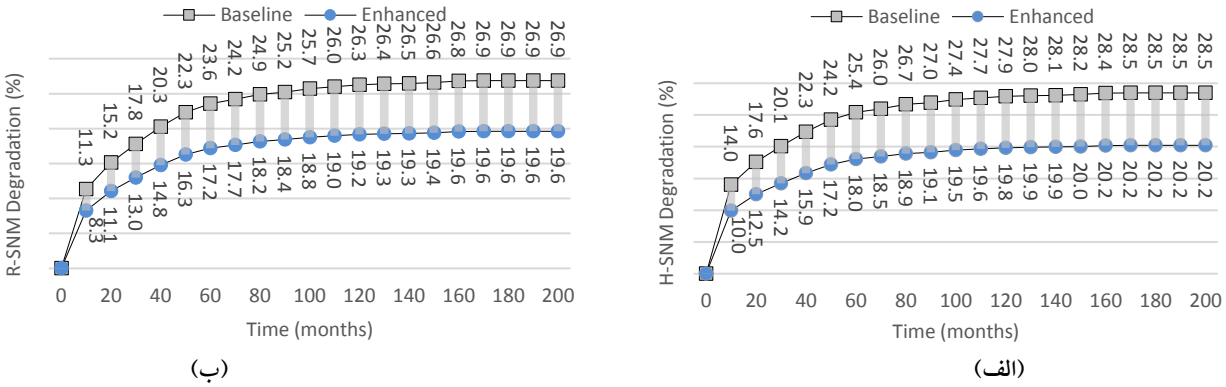
ترانزیستورهای سلول‌های حافظه نهان در حافظه نهان پایه و بهبودیافته را می توان در طول زمان تا ۱۰۱/۳۳ میلی‌ولت و ۸۷/۱۲ میلی‌ولت افزایش داد که منجر به کاهش حدود ۱۴/۰ درصدی کاهش ولتاژ آستانه در زمان استفاده از معماری پیشنهادی می‌شود. علاوه بر این، میانگین درصد خطا بین نتایج حافظه‌های نهان بهبودیافته و منصفانه حدود ۰/۸۲ درصد است که در مقایسه با درصد خطای ۱۷/۲۳ درصدی بین نتایج حافظه‌های نهان پایه و منصفانه، بسیار ناچیز است.

پایداری سلول‌های حافظه نهان به حاشیه نویز ایستا بستگی دارد و ΔVT نامتوازن ترانزیستورهای سلول حافظه نهان می‌تواند حاشیه نویز ایستا سلول را با عدم تعادل عملکردی مواجه کند. رویکرد "آزمایش و خطا" برای اندازه‌گیری حاشیه نویز ایستا سلول‌های حافظه نهان استفاده شده است. حاشیه نویز ایستا یک سلول حافظه نهان را می‌توان بر حسب دو معیار اندازه‌گیری کرد: H-SNM و R-SNM. H-SNM و R-SNM حداقل ولتاژ مورد نیاز برای تغییر مقدار ذخیره شده سلول به ترتیب در طول زمان ذخیره‌سازی و خواندن مقادیر را نشان می‌دهند. یک منبع ولتاژ کنترل‌پذیر با ولتاژ برای ساده کردن طراحی استفاده شده است. این بدان معنی است که ولتاژ در یک منبع ولتاژ ایده‌آل توسط ولتاژ دیگری در مدار تعیین می‌شود. به این ترتیب، اطمینان حاصل می‌کنیم که ولتاژ نویز وارد شده در هر دو سلول یکسان است.

میزان کاهش H-SNM و R-SNM سلول‌های حافظه نهان مورد استفاده در حافظه نهان پایه و بهبودیافته در شکل ۱۴ نشان داده شده است. H-SNM و R-SNM پس از ۱۸۰ ماه عملیات حافظه نهان به حدود ۷۷/۷ میلی‌ولت و ۷۴/۷ میلی‌ولت کاهش می‌یابد که به ترتیب منجر به ۲۸/۵٪ و ۲۶/۹٪ کاهش حاشیه نویز ایستا می‌شود. با این حال، حاشیه نویز ایستای سلول‌های حافظه نهان تحت مدیریت حافظه نهان بهبودیافته به میزان کمتری کاهش می‌یابد. در این مورد، H-SNM و R-SNM حدود ۲۰/۲٪ و ۱۹/۶٪ پس از ۱۸۰ ماه کاهش می‌یابند، که به بهبود ۹/۹٪ و ۱۱/۵٪ در کاهش H-SNM و R-SNM منجر می‌شود.

۴-۲- سربار کارآیی

سیاست‌های جایگزینی بر عملکرد کلی حافظه نهان از نظر نرخ برخورد در دسترسی به حافظه نهان و استفاده از پهنای باند شبکه میان‌ارتباطی تأثیر می‌گذارند. یک سیاست جایگزینی ضعیف می‌تواند تعداد موارد رخداد عدم برخورد حافظه نهان را افزایش دهد. علاوه بر این، از آنجایی که تمرکز روش پیشنهادی بر تنظیم توزیع داده‌های کثیف بر روی خطوط حافظه نهان می‌باشد، لذا می‌تواند بر



شکل ۱۴. مقایسه میزان کاهش مقادیر R-SNM (الف) و H-SNM (ب) سلول‌های حافظه نهان در زمان استفاده از روش پیشنهادی و روش پایه

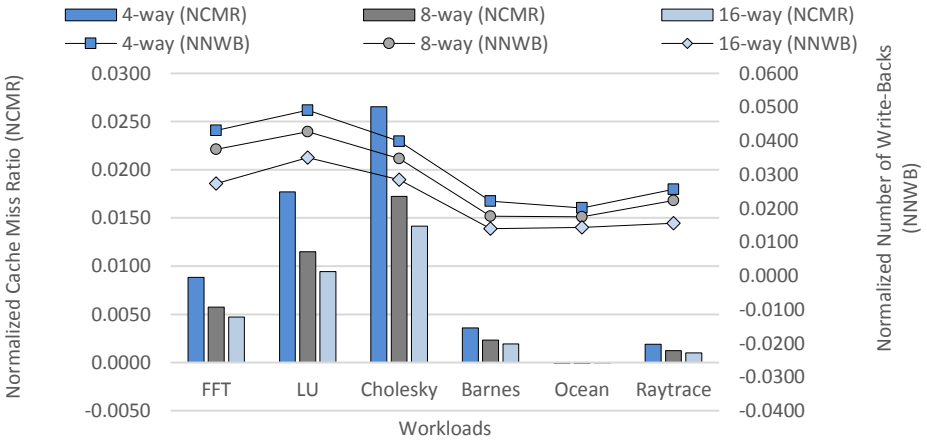
تعداد بازگردانی‌های نوشتن[۲۱] به سطح پایین‌تر حافظه (حافظه نهان سطح دوم) نیز تأثیر بگذارد. بنابراین، دو معیار برای مقایسه اثرات تصمیم‌های طراحی بر کارآیی حافظه نهان در نظر گرفته شده است: نرخ برخورد حافظه نهان و تعداد بازگردانی‌ها. شکل ۱۵ اثرات حافظه نهان بهبودیافته بر معیارهای بیان شده را نشان می‌دهد.

با توجه به شکل ۱۵، حافظه نهان بهبودیافته نرخ عدم برخورد حافظه نهان را در مقایسه با حافظه نهان پایه به طور متوسط حدود ۰/۹۸٪، ۰/۶۳٪ و ۰/۵۲٪ در صورت استفاده از مجموعه‌های ۴ جهته، ۸ جهته و ۱۶ جهته افزایش می‌دهد. نتایج تجربی نشان می‌دهد که حذف زودهنگام بلوک‌های داده انتخابی در روش پیشنهادی تاثیر اندکی بر نرخ برخورد دارد، زیرا روش پیشنهادی غالبا نامزدهای جایگزین را پس از عدم دسترسی طولانی از حافظه نهان خارج می‌کند. افزایش درجه انجمنی باعث کاهش تعداد عدم برخورد می‌شود. بنابراین، کاهش نرخ برخورد، مشاهده شده به دلیل روش پیشنهادی، با افزایش درجه انجمنی قابل تعدیل است.

علاوه بر این، روش پیشنهادی تعداد بازنویسی‌ها را در مقایسه با حافظه نهان پایه به طور متوسط حدود ۳/۳۴٪، ۲/۸۷٪ و ۲/۲۵٪ در حافظه نهان ۴، ۸ و ۱۶ جهته افزایش می‌دهد. با افزایش درجه انجمنی حافظه نهان، فاصله بین نتایج عملی انتخاب دو بلوک داده متوالی در رتبه‌بندی جایگزینی کاهش می‌یابد. بنابراین، خروج نامزد جایگزینی رتبه دوم، سربار کارآیی کمتری را به سیستم تحمیل می‌کند (در زمانی که درجه انجمنی حافظه از ۴ جهته به ۱۶ جهته افزایش می‌یابد).

۴-۳- سربارهای انرژی و مساحت

ضریب فعالیت[۲۲] سلول‌های حافظه نهان، که عمدتاً برای جلوگیری از کاهش حاشیه نویز ایستا توسط روش‌های مدیریت حافظه نهان آگاه از سالمندی مورد تمرکز قرار می‌گیرند، می‌تواند مستقیماً بر مصرف انرژی پویا حافظه نهان تأثیر بگذارد. شبیه ساز CACTI 6.5 برای تخمین مصرف انرژی پویای حافظه نهان پایه و بهبودیافته با استفاده از فناوری ۳۲ نانومتری استفاده شده است [۳۴]. جدول ۲ اثرات حافظه نهان بهبودیافته بر مصرف انرژی پویا را نشان می‌دهد. همانطور که در جدول ۲ نشان داده شده است، مصرف انرژی پویا در نتیجه افزایش فعالیت‌های سوئیچینگ سلول‌های حافظه نهان بهبودیافته افزایش می‌یابد. متوازن کردن توزیع احتمال ضریب



شکل ۱۵. تاثیرات اعمال روش پیشنهادی بر پارامترهای کارآیی حافظه نهان

جدول ۲. تاثیرات اعمال روش پیشنهادی بر روی انرژی مصرفی پویا حافظه نهان

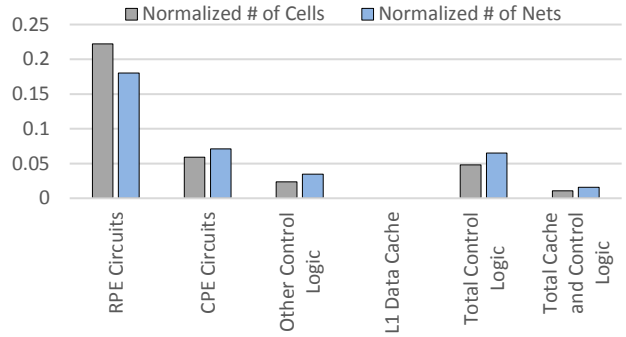
Workloads	FFT	LU	Cholesky	Barnes	Ocean	Raytrace	Avg.
Normalized Dynamic Energy	0.047	0.052	0.044	0.029	0.027	0.032	0.038

احتمال سیگنال بیت‌های خطوط حافظه نهان با استفاده از حذف زودهنگام انتخابی بلوک‌های داده تمیز/کثیف استفاده نشده منجر به افزایش نرخ عدم برخورد حافظه نهان و ضریب فعالیت می‌شود. با توجه به جدول ۲، حافظه نهان ۴ جهته بهبودیافته مصرف انرژی حافظه نهان پویا را در مقایسه با حافظه نهان ۴ جهته پایه به طور متوسط حدود ۳/۸۳٪ افزایش می‌دهد.

علاوه بر این، مدارات اضافی در کنترل‌کننده حافظه نهان که برای اجرای سیاست جایگزینی بهبودیافته درج شده است، نیاز به تجزیه و تحلیل مساحت مصرفی را ایجاب می‌کند. با توجه به محدودیت‌های ابزار CACTI برای مدل‌سازی مساحت مصرفی مدارهای کنترل‌کننده حافظه نهان، که برای اجرای سیاست جایگزینی و پروتکل هماهنگی تعبیه شده‌اند، ما از مدل‌سازی و سنتز در سطح انتقال ثبات[۲۳] برای ارزیابی مساحت مصرفی استفاده کردیم. برای این منظور، یک مدل RTL از منطق کنترل‌کننده حافظه نهان پیاده سازی شده است. علاوه بر این، از Synopsys Design Compiler برای انجام سنتز سخت‌افزار بهره گرفته شده است [۳۵]. شکل ۱۶ نتایج سنتز سخت‌افزار مدل‌های توصیف شده تصمیمات طراحی حافظه نهان را نشان می‌دهد. اثرات حافظه نهان بهبودیافته بر روی مساحت بخش‌های مختلف حافظه نهان داده و کنترل‌کننده آن، از جمله مدارهای اجرای سیاست جایگزینی، مدارهای اجرای پروتکل هماهنگی و سایر اجزای کنترل‌کننده حافظه نهان گزارش شده است. با جمع‌بندی نتایج آنالیز مساحت، تعداد سلول‌ها و مدارات کنترل‌کننده حافظه نهان بهبودیافته حدود ۴/۷۸ و ۶/۵۱ درصد رشد می‌کند. با در نظر گرفتن کل مدارات حافظه نهان، استفاده از روش پیشنهادی به ترتیب منجر به حدود ۱/۰۸ و ۱/۵۶ درصد سربار مساحت می‌شود.

۴-۴- جمع‌بندی نتایج

ساختار حافظه نهان پیشنهادی با حذف یکی از دلایل احتمال وقوع



شکل ۱۶. تاثیرات اعمال روش پیشنهادی بر روی مساحت مصرفی حافظه نهان

ضریب احتمال سیگنال نامتعادل سلول‌های حافظه نهان، که مورد تمرکز کارهای پیشین نبوده است، از وقوع پدیده سالمندی ناشی از ناپایداری‌های بایاس-دما جلوگیری می‌کند. بنابراین، روش پیشنهادی به عنوان جایگزینی برای کارهای مرتبط قبلی ارائه نشده است و می‌توان آن را همراه با تغییرات در بخش‌های مختلف حافظه نهان و مدارهای کنترلی آن که برای جلوگیری از کاهش حاشیه نویز ایستا توسط سایر دلایل اعمال شده است، استفاده کرد. با توجه به این که ساختار حافظه نهان پیشنهادی و روش‌های قبلی بر حذف علل مختلف ضریب احتمال سیگنال نامتعادل سلول‌های حافظه نهان تمرکز کرده‌اند، در نتیجه، نتایج آن‌ها قابل مقایسه نیستند. در شرایط ایده‌آل، توزیع عادلانه بلوک‌های داده مرتبط با حالت‌های مختلف پروتکل هماهنگی می‌تواند کاهش حاشیه نویز ایستا را تا ۱۳/۶ درصد، همراه با حدود ۷/۳ درصد کاهش نرخ برخورد حافظه نهان و حدود ۹/۲ درصد سربار انرژی مصرفی پویا، بهبود بخشد. همانطور که توسط نتایج تجربی نشان داده شد، ساختار حافظه نهان پیشنهادی می‌تواند به طور موثر کارآیی و سربار انرژی توزیع ایده‌آل را تعدیل کند و در عین حال بهبود کاهش حاشیه نویز ایستا را تا بیش از ۱۰ درصد حفظ کند.

۵- جمع‌بندی

این مقاله یک سیاست جایگزینی بهبود یافته برای بهبود طول عمر سلول‌های حافظه نهان تحت استرس پدیده سالمندی پیشنهاد می‌کند. کاهش حاشیه نویز استاتیک سلول‌های حافظه نهان یکی از پیامدهای مضر سالمندی است که حافظه نهان را آسیب‌پذیرترین جزء روی تراشه در برابر خطاهای سخت ناشی از ناپایداری‌های بایاس-دما کرده است. هدف از بهبود ارائه شده توسط این طرح کاهش خطاهای سخت در سلول‌های حافظه نهان ناشی از کاهش حاشیه نویز ایستا است. نتایج تجربی نشان می‌دهد که توزیع نامتعادل بلوک‌های داده مرتبط با حالت‌های مختلف پروتکل

هماهنگی بر روی خطوط یک مجموعه حافظه نهان می‌تواند کاهش حاشیه نویز ایستا سلول‌های حافظه نهان را تشدید کند. بر اساس یافته‌ها، یک سیاست جایگزینی حافظه نهان آگاه از ناپایداری‌های بایاس-دما، برای متعادل کردن شانس قرارگیری بلوک‌های داده مرتبط با حالت‌های مختلف پروتکل هماهنگی ارائه شده است. برای این منظور، درخت تصمیم PLRU با در نظر گرفتن حالات پروتکل هماهنگی خطوط حافظه نهان و نوع عدم برخورد مورد بازدید و اصلاح قرار گرفت. با استفاده از حافظه نهان بهبود یافته، کاهش حاشیه نویز ایستای ذخیره‌سازی و خواندن حدود ۹/۹ درصد و ۱۱/۵ درصد، با کمتر از ۱/۰ درصد کاهش در نرخ برخورد حافظه نهان و سربار ناچیز انرژی و مساحت، بهبود می‌یابد.

مراجع

[1] A. Agarwal, B. C. Paul, H. Mahmoodi, A. Datta and K. Roy, "A process-tolerant cache architecture for improved yield in nanoscale technologies," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 1, pp. 27-38, Jan. 2005, doi: 10.1109/TVLSI.2004.840407.

[2] T. Grasser, "Bias Temperature Instability for Devices and Circuits," Springer New York, 2014. doi: 10.1007/978-1-4614-7909-3.

[3] E. Mintarno, V. Chandra, D. Pietromonaco, R. Aitken and R. W. Dutton, "Workload dependent NBTI and PBTI analysis for a sub-45nm commercial microprocessor," *2013 IEEE International Reliability Physics Symposium (IRPS)*, 2013, pp. 3A.1.1-3A.1.6, doi: 10.1109/IRPS.2013.6531971.

[4] F. Crupi, C. Pace, G. Cocorullo, G. Groeseneken, M. Aoulaiche, and M. Houssa, "Positive bias temperature instability in nMOSFETs with ultra-thin Hf-silicate gate dielectrics," in *Microelectronics Engineering*, vol. 80, pp. 130–133, Jun. 2005, doi: 10.1016/j.mee.2005.04.055.

[5] J. Keane, X. Wang, D. Persaud and C. H. Kim, "An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI, and TDDB," in *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 817-829, April 2010, doi: 10.1109/JSSC.2010.2040125.

[6] A. T. Krishnan, V. Reddy, S. Chakravarthi, J. Rodriguez, S. John and S. Krishnan, "NBTI impact on transistor and circuit: models, mechanisms and scaling effects [MOSFETs]," *IEEE International Electron Devices Meeting 2003*, 2003, pp. 14.5.1-14.5.4, doi: 10.1109/IEDM.2003.1269296.

[7] A. Bansal, R. Rao, J. Kim, S. Zafar, J. H. Stathis amd C. Chuang, "Impacts of nbti and pbti on sram static/dynamic noise margins and cell failure probability," in *Microelectronics Reliability*, vol. 49, no. 6, pp. 642–649, 2009, doi: 10.1016/j.microrel.2009.03.016.

[8] M. Bagatin, S. Gerardin, A. Paccagnella and F. Faccio, "Impact of nbti aging on the single-event upset of sram cells," in *IEEE Transactions on Nuclear Science*, vol. 57, no. 6, pp. 3245–3250, Dec. 2010, doi: 10.1109/TNS.2010.2084100.

[9] A. T. Krishnan, V. Reddy, D. Aldrich, J. Raval, K. Christensen, J. Rosal, C. O'Brien, R. Khamankar, A. Marshall, W-K. Loh; R. McKee and S. Krishnan, "SRAM Cell Static Noise Margin and VMIN Sensitivity to Transistor Degradation," *2006 International Electron Devices Meeting*, 2006, pp. 1-4, doi: 10.1109/IEDM.2006.346778.

[10] Y. Kunitake, T. Sato and H. Yasuura, "A case study of Short Term Cell-Flipping technique for mitigating NBTI degradation on cache," *2nd Asia Symposium on Quality Electronic Design (ASQED)*, 2010, pp. 301-307, doi: 10.1109/ASQED.2010.5548256.

[11] A. Jafari, M. Raji and B. Ghavami, "BTI-Aware Timing Reliability Improvement of Pulsed Flip-Flops in Nano-Scale CMOS Technology," in *IEEE Transactions on Device and Materials Reliability*, vol. 21, no. 3, pp. 379-388, Sept. 2021, doi: 10.1109/TDMR.2021.3102521.

[12] S. Mukhopadhyay, H. Mahmoodi and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 12, pp. 1859-1880, Dec. 2005, doi: 10.1109/TCAD.2005.852295.

[13] Q. Chen, H. Mahmoodi, S. Bhunia and K. Roy, "Modeling and testing of SRAM for new failure mechanisms due to process variations in nanoscale CMOS," *23rd IEEE VLSI Test Symposium (VTS'05)*, 2005, pp. 292-297, doi: 10.1109/VTS.2005.58.

[14] F. Firouzi, S. Kiamehr and M. B. Tahoori, "Statistical analysis of BTI in the presence of process-induced voltage and temperature variations," *2013 18th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2013, pp. 594-600, doi: 10.1109/ASPDAC.2013.6509663.

[15] M. Maghsoudloo and H. R. Zarandi, "Design space exploration of non-uniform cache access for soft-error vulnerability mitigation", in *Microelectronics Reliability*, vol. 55, no. 11, 2015, pp. 2439-2452, doi: 10.1016/j.microrel.2015.07.049.

[16] J. Franco, B. Kaczer and G. Groeseneken, "Reliability of High Mobility SiGe Channel MOSFETs for Future CMOS Applications," Springer Dordrecht, 2014, doi: 10.1007/978-94-007-7663-0.

[17] H. Amrouch, T. Ebi and J. Henkel, "Stress balancing to mitigate NBTI effects in register files," *2013 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2013, pp. 1-10, doi: 10.1109/DSN.2013.6575315.

[18] Y. Kunitake, T. Sato and H. Yasuura, "Signal probability control for relieving NBTI in SRAM cells," *2010 11th International Symposium on Quality Electronic Design (ISQED)*, 2010, pp. 660-666, doi: 10.1109/ISQED.2010.5450504.

[19] A. Gebregiorgis, M. Ebrahimi, S. Kiamehr, F. Oboril, S. Hamdioui and M. B. Tahoori, "Aging mitigation in memory arrays using self-controlled bit-flipping technique," *The 20th Asia and South Pacific Design Automation Conference*, 2015, pp. 231-236, doi: 10.1109/ASPDAC.2015.7059010.

[20] M. Karimi, N. Rohbani and S. Miremadi, "A Low Area Overhead NBTI/PBTI Sensor for SRAM Memories," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 11, pp. 3138-3151, Nov. 2017, doi: 10.1109/TVLSI.2017.2734839.

[21] B. Liu and C. -H. Chen, "Testing, diagnosis and repair methods for NBTI-induced SRAM faults," *2014 IEEE International Conference on IC Design & Technology*, 2014, pp. 1-4, doi: 10.1109/ICICDT.2014.6838608.

[22] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B. C. Paul, W. Wang, B. Yang, Y. Cao and S. Mitra, "Optimized Circuit Failure Prediction for Aging: Practicality and Promise," *2008 IEEE International Test Conference*, 2008, pp. 1-10, doi: 10.1109/TEST.2008.4700619.

[23] K. Kang, M. A. Alam and K. Roy, "Characterization of NBTI induced temporal performance degradation in nano-scale SRAM array using IDDQ," *2007 IEEE International Test Conference*, 2007, pp. 1-10, doi: 10.1109/TEST.2007.4437590.

[24] N. Rohbani, H. Gau, S. Mohammadinejad, T. Kumar Maiti, D. Navarro, M. Miura-Mattausch, H. J. Mattausch and H. Takatsuka, "Power Reduction and BTI Mitigation of Data-Cache Memory Based on the Storage Management of Narrow-Width Values," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 7, pp. 1675-1684, July 2019, doi: 10.1109/TVLSI.2019.2909488

[25] J. Hu, S. Wang and S. G. Ziavras, "In-Register Duplication: Exploiting Narrow-Width Value for Improving Register File

Reliability," *International Conference on Dependable Systems and Networks (DSN'06)*, 2006, pp. 281-290, doi: 10.1109/DSN.2006.43.

[26] M. Imani, S. Patil and T. Rosing, "DCC: Double capacity Cache architecture for narrow-width values," *2016 International Great Lakes Symposium on VLSI (GLSVLSI)*, 2016, pp. 113-116, doi: 10.1145/2902961.2902990.

[27] K. Flautner, Nam Sung Kim, S. Martin, D. Blaauw and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power," *Proceedings 29th Annual International Symposium on Computer Architecture*, 2002, pp. 148-157, doi: 10.1109/ISCA.2002.1003572.

[28] S. Kiamehr, F. Firouzi, M. Ebrahimi and M. B. Tahoori, "Aging-aware standard cell library design," *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2014, pp. 1-4, doi: 10.7873/DATE.2014.274.

[29] N. Rohbani, M. Ebrahimi, S. Miremadi and M. B. Tahoori, "Bias Temperature Instability Mitigation via Adaptive Cache Size Management," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 3, pp. 1012-1022, March 2017, doi: 10.1109/TVLSI.2016.2606579.

[30] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, B. Werner, "Simics: A full system simulation platform," in *Computer*, vol. 35, no. 2, pp. 50-58, Feb. 2002, doi: 10.1109/2.982916.

[31] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill and D. A. Wood, "The gem5 simulator," *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, May 2011, doi: 10.1145/2024716.2024718.

[32] C. Sakalis, C. Leonardsson, S. Kaxiras and A. Ros, "Splash-3: A properly synchronized benchmark suite for contemporary research," *2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2016, pp. 101-111, doi: 10.1109/ISPASS.2016.7482078.

[33] B. Tudor, J. Wang, C. Sun, Z. Chen, Z. Liao, R. Tan, W. Liu and F. Lee, "MOSRA: An efficient and versatile MOS aging modeling and reliability analysis solution for 45nm and below," *2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, 2010, pp. 1645-1647, doi: 10.1109/ICSICT.2010.5667399.

[34] CACTI: An integrated cache and memory access time, cycle time, area, leakage, and dynamic power model, https://www.hpl.hp.com/research/cacti.

[35] SYNOPSYS Design Compiler, https://www.synopsys.com.

پاورقی‌ها:

[1] Hard errors

[2] Bias temperature instability (BTI)

[3] Static noise margin (SNM)

[4] Static random-access memory (SRAM)

[5] On-chip cache memories

[6] Signal probability (SP) factor

[7] Cache coherence protocols

[8] Cache replacement policy

[9] Address-conflict miss

[10] Hold SNM and Read SNM

[11] Write-back

[12] Negative Bias-Temperature Instability (NBTI)

[13] Positive Bias-Temperature Instability (PBTI)

[14] Word-line

[15] Set associative

[16] Pseudo-Least Recently Used

[17] Metal-oxide semiconductor reliability analysis (MOSRA)

[18] Pre-stress

[19] Post-stress

[20] Duty cycle

[21] Write-backs

[22] Activity factor

[23] Register-transfer level (RTL)