بررسی تأثیر سلسله‌مراتب حافظه نهان ناهمگن در پردازنده‌های مراکز داده

عدنان نصری، محمود فتحی و علی برومندنیا

چکیده: این مقاله به مسأله تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان برای پردازنده‌های مراکز داده در عصر سیلیکون تاریک پرداخته است. همان طور که مصرف انرژی به یکی از مباحث مهم عملیات و نگهداری مراکز داده ابری تبدیل شده است، فراهم‌کنندگان سرویس‌های ابری به شدت در این زمینه نگران شده‌اند. تکنولوژی حافظه‌های غیر فرار نوظهور جایگزینی مناسب برای حافظه‌های متداول امروزی می‌باشند. ما در این مقاله از حافظه غیر فرار STT-RAM در مقایسه با حافظه SRAM به عنوان حافظه نهان سطح آخر استفاده می‌کنیم. تراکم بالا، دسترسی خواندن سریع، توان مصرفی نشتی نزدیک به صفر و غیر فرار بودن باعث می‌شود حافظه STT-RAM یک فناوری مهم برای حافظه‌های درون تراشه باشد. در اکثر تحقیقات قبلی که از حافظه‌های غیر فرار بهره گرفته‌اند، روش‌های خاص و مبتنی بر محک‌های متعارف بررسی شده و در مورد محک‌های ابری نوظهور تحت عنوان بارهای کاری Scale-out تحلیل کاملی انجام نداده‌اند. ما در این مقاله با اجرای بارهای کاری Scale-out، تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان پردازنده‌های ابری مراکز داده را بررسی می‌کنیم. نتایج آزمایش روی محکِ CloudSuite نشان می‌دهد که استفاده از حافظه STT-RAM در مقایسه با حافظه SRAM در حافظه نهان سطح آخر، میزان انرژی مصرفی را حداکثر ۵۹٪ کاهش می‌دهد.

کلیدواژه: مرکز داده ابری، پردازنده، سلسله‌مراتب حافظه نهان، حافظه غیر فرار، محک CloudSuite.

۱- مقدمه

مراکز داده سکوهایی[1] هستند که خدمات برخط مقیاس‌پذیر ارائه می‌کنند. مایکروسافت، گوگل و فیسبوک برای ارائه بسیاری از خدمات خود از شبکه‌های مرکز داده استفاده می‌کنند. یکی از مهم‌ترین نگرانی‌های توزیع و سیستم‌های توزیع‌شده، انرژی مصرفی آنها است. مراکز داده ۱۰۰ تا ۲۰۰ برابر دفاتر استاندارد انرژی مصرف می‌کنند. برآورد شده که مراکز داده در سراسر جهان حدود ۲۶ GW انرژی مصرف می‌کنند که برابر است با ۱/۴٪ مصرف انرژی در جهان که سالانه ۱۲٪ نیز افزایش می‌یابد [۱] و [۲]. پس مراکز داده هدف اصلی طرح‌هایی هستند که بهره‌وری انرژی دارند و می‌توانند هزینه و مصرف برق را کاهش دهند. اما ماهیت حساس وظایف مراکز داده باعث می‌شود بسیاری از معیارهای طراحی غیر از بهره‌وری انرژی بهبود یابند. سرورهای درون Rackها عامل اصلی اتلاف انرژی هستند و بخش اعظم انرژی مربوط به تجهیزات کامپیوتری در یک مرکز داده معمولی را مصرف می‌کنند. سرورها بیشتر فضا را اشغال کرده و کل عملیات را اجرا می‌کنند. اخیراً در سیستم‌های خنک‌کننده پردازنده سرورها، بهبودهای گسترده‌ای ایجاد شده تا این اتلاف انرژی کاهش یابد [۳]. علاوه بر این، بیشتر برنامه‌هایی که در مراکز داده اجرا می‌شوند مانند موتور جستجو و استریمینگ رسانه[2]، از نوع بارهای کاری Scale-out هستند. چون تقاضا برای سرویس‌های ابری رو به رشد است، زیرساخت مراکز داده نمی‌تواند به خوبی پاسخگوی نیازهای این برنامه‌های نوظهور باشد. بنابراین پردازنده‌های ابری باید با نیازهای بارهای کاری نوظهور Scale-out سازگار شوند تا کارایی مراکز داده افزایش یابد. بارهای کاری Scale-out ویژگی‌هایی مانند موازی‌سازی پایین در سطح دستورالعمل، تقاضای پایین برای پهنای باند درون و بیرون از تراشه، بخش کد برنامه و مجموعه کاری بسیار بزرگ و همچنین نرخ بالای فقدان حافظه نهان دستورالعمل دارند که رفتار آنها را با بارهای کاری Parallel، Desktop و Traditioal Server متمایز می‌سازد [۴] و [۵]. می‌توانیم مصرف انرژی در مراکز داده را به دو بخش تقسیم کنیم: منابع محاسباتی و منابع فیزیکی. آمارهای ارائه‌شده نشان می‌دهد که مصرف انرژی در منابع محاسباتی حدوداً ۵۰٪ کل مصرف انرژی است. طبق شکل ۱، محاسبات در سرورها، تجهیزات ارتباطی، دستگاه‌های ذخیره‌سازی، سیستم‌های خنک‌کننده و سیستم تأمین برق به ترتیب حدوداً ۴۰، ۵، ۵، ۴۰ و ۱۰٪ از انرژی را مصرف می‌کنند. پس به این نتیجه می‌رسیم که سرورها یکی از مهم‌ترین منابع مصرف انرژی در مراکز داده هستند و بنابراین کاهش مصرف انرژی در سرورها یک مسأله مهم در مراکز داده است.

همان طور که بیان شد درون مراکز داده ابری یکی از اجزایی که بخش مهمی از انرژی مربوط به تجهیزات کامپیوتری را مصرف می‌کند، سرورها هستند. پردازنده‌ها بیشترین انرژی را در سرورهای مرکز داده مصرف می‌کنند. درون تراشه یک پردازنده، بیشترین توان در حافظه نهان سطح آخر[3] (LLC) مصرف می‌شود که توان مصرفی نشتی بخش اصلیِ این توان مصرف‌شده را تشکیل می‌دهد. در پردازنده‌های ابری بیشتر انرژی اتلاف‌شده در منابع درون تراشه مربوط به حافظه نهان سطح آخر است. توان نشتی بخش اصلی توان مصرف‌شده در حافظه نهان سطح آخر است (در حافظه نهان سطح آخر از نوع SRAM به ۸۰٪ می‌رسد [۶]) که می‌توان با استفاده از فناوری حافظه غیر فرار، آن را کاهش دهیم. حافظه‌های STT-RAM و ReRAM این ویژگی‌ها را دارند و می‌توانیم به جای حافظه نهان سطح آخر از نوع SRAM از آنها استفاده کنیم.

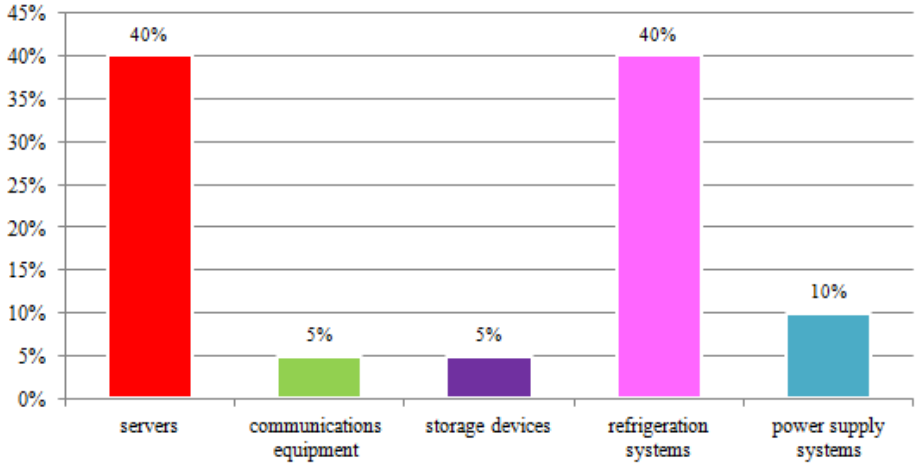
این مقاله در تاریخ ۱ اردیبهشت ماه ۱۳۹۸ دریافت و در تاریخ ۷ بهمن ماه ۱۳۹۸ بازنگری شد.

عدنان نصری، گروه مهندسی کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران، (email: adnan.nasri@gmail.com).

محمود فتحی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران، (email: mahfathy@iust.ac.ir).

علی برومندنیا، گروه مهندسی کامپیوتر، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران، (email: broumandnia@azad.ac.ir).

1. Platform
2. Media Streaming
3. Last Level Cache



شکل ۱: توزیع مصرف انرژی در مراکز داده [۳].

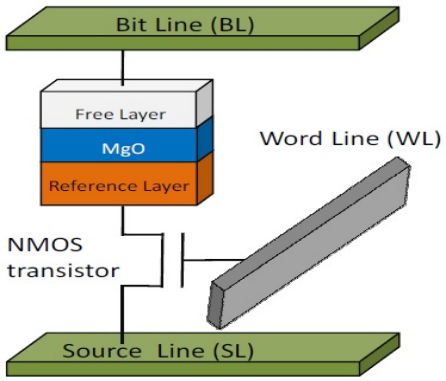
حافظه STT-RAM دارای زمان دسترسی کوتاه برای خواندن و طول عمر بالای سلول و تراکم بالا (تقریباً چهار برابر حافظه SRAM) می‌باشد. از سوی دیگر حافظه ReRAM با فناوری CMOS سازگار است اما تأخیر دسترسی آن بیشتر و همچنین طول عمر آن کمتر از حافظه STT-RAM است که باعث می‌شود برای حافظه نهان سطح آخر در سلسله‌مراتب حافظه نهان گزینه مناسبی در مقایسه با حافظه STT-RAM نباشد [۶].

هدف ما در این مقاله این است که با تغییر نوع و اندازه بانک‌های حافظه نهان سطح آخر، مصرف انرژی در پردازنده‌های مراکز داده را کاهش دهیم. ما برای این کار از حافظه غیر فرار از نوع STT-RAM در سلسله‌مراتب حافظه نهان استفاده می‌کنیم چرا که تراکم این حافظه بسیار بیشتر از حافظه متداول SRAM است و همچنین به دلیل این که غیر فرار است توان مصرفی نشتی آنها نزدیک به صفر است و در نتیجه انرژی کلی مصرفی در حافظه نهان سطح آخر را به طور قابل توجهی کاهش می‌دهد. ادامه این مقاله به صورت زیر سازمان‌دهی شده است. در بخش ۲ کارهای مرتبط پیشین مورد بررسی قرار گرفته و در بخش ۳ مقدماتی در مورد ساختار حافظه STT-RAM و فناوری مجتمع‌سازی سه‌بعدی ارائه شده است. در بخش ۴ معماری پیشنهادی و در بخش ۵ شبیه‌سازی و نتایج آن ارائه خواهد شد. در بخش ۶ نیز نتیجه‌گیری مقاله آمده است.

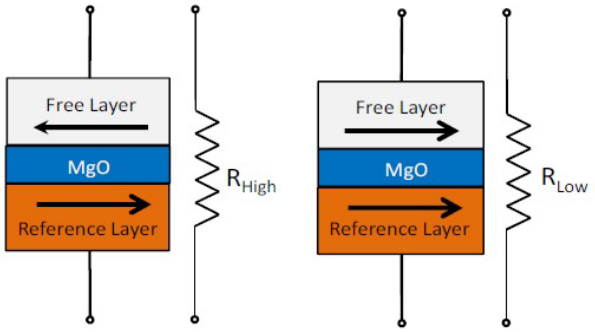
۲- کارهای پیشین

لطفی و همکاران در [۵] و [۷] سعی کرده‌اند با اصلاح معماری پردازنده، کارایی را افزایش دهند. با این که محک‌های[1] ارزیابی این تحقیقات با محک مورد ارزیابی در این مقاله یکسان می‌باشد اما در این تحقیقات میزان مصرف انرژی مورد توجه قرار نگرفته است. علاوه بر این، تفاوت دیگر این تحقیقات با این مقاله این است که تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان نیز بررسی نشده است. کاراکوستاس و همکاران در [۸] تأثیر سربار واحد مدیریت حافظه بر کارایی در زمان اجرای بارهای کاری Scale-out را بررسی کرده‌اند. در این تحقیق نیز میزان مصرف انرژی و تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان بررسی نشده است. وانگ و همکاران در [۹] هرچند که هم محک مورد ارزیابی آنها با این مقاله یکسان است و حتی میزان مصرف انرژی را نیز بررسی کرده است اما چون تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان را بررسی نکرده است با کار ارائه‌شده در این مقاله متفاوت است. پهلوان و همکاران در [۱۰] هرچند که یک محک یکسان با این مقاله را مورد ارزیابی قرار داده‌اند اما چون میزان مصرف انرژی را بررسی نکرده است و نیز تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان را بررسی نکرده‌اند با کار ارائه‌شده در این مقاله متفاوت است. در مورد [۱۱] باید به این نکته توجه داشت که از نظر معماری و تعداد هسته‌ها و پیکربندی‌های مورد ارزیابی در این مرجع، در مقایسه با این مقاله کاملاً تفاوت وجود دارد. همچنین در این مقاله از فناوری مجتمع‌سازی سه‌بعدی استفاده شده که در [۱۱] از این فناوری استفاده نگردیده است. تفاوت دیگر [۱۱] با این مقاله در محک‌های مورد ارزیابی می‌باشد. در این مقاله بارهای کاری نوظهور ابری ارزیابی شده است در حالی که در [۱۱] بارهای کاری دیگری مورد ارزیابی قرار گرفته‌اند. با این که در [۱۱] تا [۱۵] تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان بررسی شده است، اما محک مورد ارزیابی در این تحقیقات با این مقاله یکسان نیست. جوجیک و همکاران در [۱۴] و [۱۵] با استفاده از فناوری سه‌بعدی، کارایی DRAM در سلسله‌مراتب حافظه نهان را در زمان اجرای بارهای کاری Scale-out بررسی کرده‌اند اما مصرف انرژی و تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان را مد نظر قرار نداده‌اند. کیان و همکاران در [۱۶] تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان را با استفاده از فناوری سه‌بعدی بررسی کرده‌اند ولی مصرف انرژی را مد نظر قرار نداده‌اند. تفاوت دیگر این مقاله و تحقیق مذکور این است که حافظه‌های غیر فرار مورد استفاده در این مقاله از نوع STT-RAM نیستند. علاوه بر این، محک استفاده‌شده در این مقاله با محک مورد ارزیابی این مقاله تفاوت دارد. در [۱۷] تا [۲۱] با استفاده از فناوری سه‌بعدی، مصرف انرژی در زمان استفاده از حافظه‌های غیر فرار مانند STT-RAM در سلسله‌مراتب حافظه نهان بررسی شده است. تفاوت اصلی این تحقیقات با این مقاله، محک مورد ارزیابی است. هیچ کدام از این تحقیقات محک CloudSuite [۲۲] را بررسی نکرده‌اند و از محک‌های دیگری استفاده نموده‌اند. همچنین در [۵]، [۷]، [۸]، [۲۳] و [۲۴]، چند روش برای افزایش کارایی پیشنهاد شده است اما بیشتر این تحقیقات میزان مصرف انرژی را بررسی نکرده‌اند. علاوه بر این، هیچ کدام از این تحقیقات به تأثیر حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان توجه نکرده‌اند، هرچند این تحقیقات نیز مانند این مقاله از محک CloudSuite نیز به عنوان محک برای ارزیابی استفاده کرده‌اند. در [۲۵] هرچند مانند این مقاله تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان بررسی شده

1. Benchmark



شکل ۲: ساختار یک سلول حافظه STT-RAM [۲۷].



(ب) (الف)

شکل ۳: (الف) حالت صفر منطقی و (ب) حالت یک منطقی در سلول حافظه STT-RAM [۳۱].

جدول ۱: مقایسه پارامترهای حافظه‌های SRAM و STTRAM در تکنولوژی ۳۲ نانومتر [۱۹].

Technology	Area	Read (Energy)	Write (Energy)	Read (Latency)	Write (Latency)
۱ MB SRAM	۳٫۰۳ mm^2	۰٫۱۶۸ nJ	۰٫۱۶۸ nJ	۰٫۷۰۲ ns	۰٫۷۰۲ ns
۴ MB STT-RAM	۳٫۳۹ mm^2	۰٫۲۷۸ nJ	۰٫۷۶۵ nJ	۰٫۸۸۰ ns	۱۰٫۶۷ ns

است، اما محک مورد ارزیابی در این تحقیق با این مقاله کاملاً متفاوت می‌باشد. در [۲۶] تا [۲۸] علاوه بر تفاوت در محک‌های مورد ارزیابی با این مقاله، تأثیر استفاده از حافظه‌های غیر فرار در سلسله‌مراتب حافظه نهان بررسی نگردیده و از حافظه‌های غیر فرار به عنوان حافظه اصلی استفاده شده است.

۳- پیش‌زمینه

۳-۱ فناوری حافظه STT-RAM

در این بخش فناوری حافظه STT-RAM [۲۹] را که یک فناوری معروف حافظه غیر فرار است بررسی می‌کنیم. حافظه STT-RAM انقلابی در حافظه‌های مغناطیسی مانند درایوهای دیسک سخت و حافظه‌های نیمه‌رسانای جامد ایجاد کرده است و به دلیل این که در مقایسه با دیگر حافظه‌های مغناطیسی، زمان نوشتن کمتری دارد، یکی از محبوب‌ترین ساختارهای حافظه غیر فرار می‌باشد. معمولاً فناوری‌های حافظه غیر فرار در مقایسه با فناوری‌های قدیمی حافظه مانند SRAM و DRAM، ویژگی‌های مطلوب زیادی دارند. مثلاً به دلیل این که غیر فرار هستند، توان مصرفی نشتی آنها نزدیک به صفر است، تراکم سلول بالایی دارند و در برابر خطاهای نرم‌افزاری بسیار مقاوم هستند.

همان طور که در شکل ۲ نشان داده شده است، سلول اولیه حافظه STT-RAM شامل یک ترانزیستور دسترسی nMOS استاندارد و MTJ[1] است که اطلاعات را جابه‌جا می‌کند. MTJ که یکی از بخش‌های اصلی فناوری حافظه غیر فرار است از دو لایه فرومغناطیس تشکیل شده که لایه مرجع[2] و لایه آزاد[3] نامیده می‌شوند. این دو لایه با یک لایه دی‌الکتریک (MgO) از هم جدا شده‌اند. لایه اول ثابت است و لایه دوم با ارسال یک جریان بزرگ از طریق MTJ کنترل می‌شود. وقتی دو لایه جهت‌های مغناطیسی یکسانی داشته باشند، MTJ در حالت مقاومت پایین است و صفر منطقی را نشان می‌دهد و وقتی دو لایه جهت‌های مغناطیسی یکسانی نداشته باشند، MTJ در حالت مقاومت بالا است و یک منطقی را نشان می‌دهد که این حالت‌ها در شکل ۳ نشان داده شده است. برای انجام عملیات خواندن در سلول STT-RAM، nMOS باید روشن شود و یک ولتاژ کوچک بین خط- بیت[4] و خط- منبع[5] ایجاد شود. با ایجاد این ولتاژ، یک جریان از MTJ عبور می‌کند. در اینجا از یک حسگر برای سنجش جریان و مقایسه آن با یک جریان مرجع استفاده می‌شود تا مقدار منطقی سلول مشخص شود. علاوه بر این برای انجام عملیات نوشتن، جریان با توجه به مقدار سلول تغییر می‌کند. بر این اساس برای نوشتن صفر یا یک در سلول یک جریان مثبت یا منفی بین خط- بیت و خط- منبع برقرار می‌شود. طول عمر نوشتنِ سلول‌های حافظه STT-RAM محدود است و وقتی به این حد می‌رسند ممکن است سلول دیگر نتواند حالت مغناطیسی لایه آزاد را تغییر دهد که این باعث خطای داده می‌شود [۳۰].

اولین ویژگی حافظه‌های غیر فرار در مقایسه با حافظه SRAM، چگالی بیشتر و سرعت دسترسی خواندن بالای آنها می‌باشد. ویژگی دیگر حافظه‌های غیر فرار، توان مصرفی نشتی نزدیک به صفر و مقاومت در برابر خطاهای نرم این نوع از حافظه می‌باشد. با این وجود، در کنار ویژگی‌های خوبی که این نوع از حافظه‌ها ارائه می‌دهند، از معایب حافظه‌های غیر فرار می‌توان در عملیات نوشتن به تأخیر نسبتاً بالا، انرژی مصرفی زیاد و تعداد محدود دفعات نوشتن نسبت به حافظه SRAM آن اشاره کرد [۳۲]. در جدول ۱ چند پارامتر حافظه SRAM و حافظه غیر فرار STTRAM با یکدیگر مقایسه شده‌اند. باید این نکته را مورد توجه قرار داد که در این جدول انرژی خواندن و نوشتن حافظه بیان شده و این میزان انرژی با انرژی مصرفی کل که متأثر از توان نشتی و همچنین عوامل دیگر می‌باشد متفاوت است.

۳-۲ فناوری مجتمع‌سازی سه‌بعدی

معمولاً لایه‌های فعال سیلیکون در فناوری مجتمع‌سازی سه‌بعدی به

1. Magnetic Tunnel Junction
2. Reference Layer
3. Free Layer
4. Bit Line
5. Source Line

صورت عمودی بر روی یکدیگر قرار می‌گیرند. این لایه‌ها از طریق TSV[1]ها با هم ارتباط برقرار می‌کنند. فناوری مدارهای مجتمع سه‌بعدی که در آن لایه‌های سیلیکون به صورت عمودی بر روی یکدیگر هستند، روش مؤثری برای افزایش تعداد ترانزیستورهای یک تراشه است. در طراحی مدارهای مجتمع سه‌بعدی، پهنای باند بین حافظه‌ها و هسته‌های پردازنده می‌تواند افزایش یابد. علاوه بر این، مدارهای مجتمع سه‌بعدی، یکپارچه‌سازی ناهمگن و طراحی پیمانه‌ای[2] و مقیاس‌پذیر را امکان‌پذیر نموده و طول ارتباطات درون تراشه را کوتاه می‌سازند و بنابراین یکپارچه‌سازی سه‌بعدی روشی برای طراحی چندهسته‌ای است که می‌تواند مشکل دیوار حافظه[3] را حل کند [۱۷] و [۲۰].

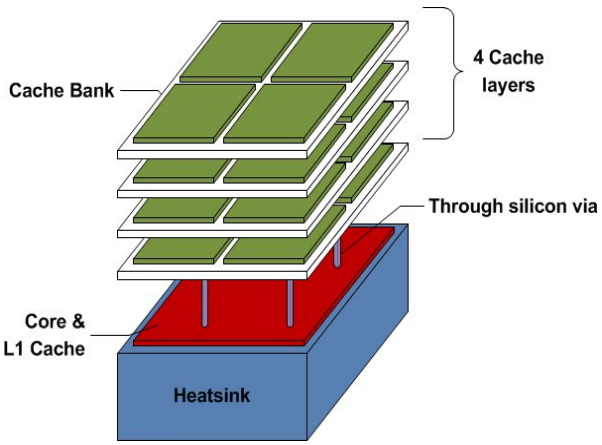
٤- معماری پیشنهادی

حافظه‌های درون تراشه با توجه به سهم عمده‌ای که در مصرف انرژی پردازنده دارند، همیشه جزء مباحث قابل توجه برای سیستم‌های تعبیه‌شده می‌باشند. استفاده از تکنولوژی‌های رایج حافظه مانند SRAM با توجه به توان نشتی بالای آنها ایجاد اشکال می‌کند، بنابراین استفاده از تکنولوژی حافظه‌های غیر فرار نظیر STT-RAM به جای حافظه SRAM به دلیل چگالی بالا، غیر فرار بودن و همچنین توان نشتی پایین می‌تواند یک گزینه کارامد و مؤثر باشد. در این مقاله ما یک سلسله‌مراتب حافظه نهان ناهمگن برای پردازنده‌های سه‌بعدی ارائه می‌کنیم که منجر به کاهش مصرف انرژی می‌گردد. همچنین بر خلاف کارهای قبلی مرتبط که تمرکز بر روی بارهای کاری برنامه‌های کاربردی کامپیوترهای رومیزی و کامپیوترهای سرور سنتی بوده است، در این مقاله تحلیل مربوط به استفاده از حافظه STT-RAM در سلسله‌مراتب حافظه نهان در پردازنده‌های مراکز داده ابری با تمرکز بر بارهای کاری نوظهور ابری، تحت عنوان بارهای کاری Scale-out ارائه گردیده است.

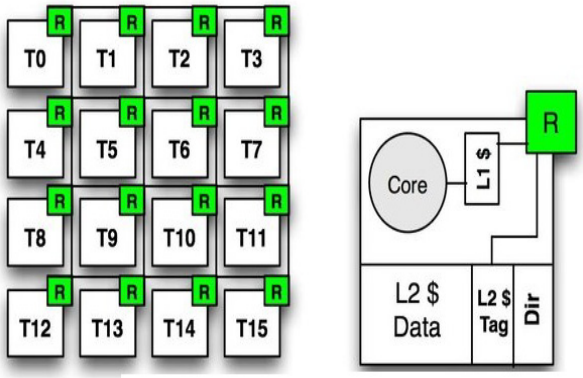
مصرف انرژی در مراکز داده ابری در سال‌های اخیر به شدت در حال افزایش است. در درون مراکز داده یکی از اجزایی که بخش مهمی از انرژی مربوط به تجهیزات کامپیوتری را مصرف می‌کند، سرورها هستند. پردازنده‌ها بیشترین انرژی را در سرورهای مرکز داده مصرف می‌کنند. درون تراشه یک پردازنده، بیشترین توان در حافظه نهان سطح آخر مصرف می‌شود که توان مصرفی نشتی بخش اصلیِ این توان مصرف‌شده را تشکیل می‌دهد. می‌توان با استفاده از فناوری حافظه‌های غیر فرار، مقدار توان مصرفی در حافظه نهان سطح آخر را کاهش داد. این کاهش مصرف انرژی با تغییر نوع و اندازه بانک‌های حافظه نهان سطح آخر و استفاده از حافظه‌های غیر فرار امکان‌پذیر خواهد بود. با توجه به ویژگی‌های حافظه‌های غیر فرار مانند زمان کوتاه دسترسی برای خواندن، طول عمر بالای سلول، توان نشتی نزدیک به صفر و تراکم بالا و همچنین با توجه به این که بیشتر دسترسی‌ها در محک CloudSuite به عنوان یک نمونه از بارهای کاری مراکز داده از نوع خواندن هستند، می‌توان گفت که به عنوان مثال حافظه غیر فرار STT-RAM می‌تواند گزینه مناسبی به جای حافظه متداول SRAM به عنوان حافظه نهان سطح آخر در پردازنده‌های ابری جهت کاهش مصرف انرژی مورد استفاده قرار گیرد [۳۰]. مراکز داده بزرگ به پردازنده‌هایی نیاز دارند که هسته‌های زیاد و مسیرهای دسترسی سریعی به آخرین سطح حافظه نهان داشته باشند تا توان عملیاتی را بیشینه کنند. اگر آخرین سطح حافظه نهان بزرگ باشد، بارهای کاری Scale-out نمی‌توانند از مزایای آن بهره ببرند (به دلیل این که در این برنامه‌ها، محلی‌بودن[4] داده‌ها کم است) و به همین دلیل پردازنده‌هایی که برای بارهای کاری Scale-out بهینه‌سازی شده‌اند باید بیشتر فضای تراشه خود را به هسته‌ها اختصاص دهند و تعداد هسته‌ها را بیشینه کنند. بارهای کاری Scale-out که در بسیاری از خدمات آنلاین امروزی وجود دارند، یک دسته از برنامه‌ها هستند که ویژگی‌های مشترکی دارند. این ویژگی‌ها آنها را از برنامه‌های desktop، پردازش رسانه و برنامه‌های کاربردی در زمینه‌های علمی متمایز می‌کند. یک بارکاری Scale-out یک سرویس جاری[5] یا جستجوی وب است. این این برنامه، جریانی از درخواست‌های مشتریان را مدیریت می‌کند که اکثراً مستقل هستند و می‌خواهند به بخشی از داده‌های یک مجموعه داده بسیار بزرگ دسترسی داشته باشند. در بارهای کاری Scale-out به دلیل پردازش درخواست‌های متنوع، فضای دستورالعمل بزرگ است. وجود ویژگی‌های مشترک (یعنی استقلال درخواست، بزرگ‌بودن فضای دستورالعمل و بزرگ‌بودن اندازه مجموعه داده) نشان می‌دهد که می‌توانیم پردازنده‌ها را برای این دسته از بارهای کاری بهینه‌سازی کنیم. به طور کلی می‌توان گفت ساختارهای پردازنده فعلی که هسته‌های زیادی دارند نیاز ما به پردازنده‌های Scale-out را رفع نمی‌کنند [۴] و [۳۴]. علاوه بر این، مصرف انرژی در مراکز داده ابری در سال‌های اخیر به شدت افزایش یافته است. پردازنده‌ها بیشترین انرژی را در مرکز داده مصرف می‌کنند. درون تراشه، بیشترین توان در حافظه نهان سطح آخر مصرف می‌شود که توان مصرفی نشتی بخش اصلیِ این توان مصرف‌شده را تشکیل می‌دهد. می‌توان با استفاده از فناوری حافظه‌های غیر فرار، مقدار توان مصرفی در حافظه نهان سطح آخر را کاهش داد. با تغییر نوع و اندازه بانک‌های حافظه نهان سطح آخر و استفاده از حافظه‌های غیر فرار، می‌توان انرژی مصرفی کل در پردازنده‌های مراکز داده را کاهش داد. با توجه به ویژگی‌های حافظه غیر فرار STT-RAM مانند زمان کوتاه دسترسی برای خواندن، طول عمر بالای سلول، توان نشتی نزدیک به صفر و تراکم بالا و همچنین با توجه به این که بیشتر دسترسی‌ها در محک CloudSuite به عنوان یک نمونه از بارهای کاری Scale-out از نوع خواندن هستند، می‌توان گفت که حافظه STT-RAM گزینه مناسبی برای ساخت حافظه نهان سطح آخر پردازنده‌های ابری می‌باشد. اگر از حافظه STT-RAM به عنوان حافظه نهان سطح آخر استفاده کنیم، با توجه به جدول ۱ عملیات نوشتن بر مصرف انرژی تأثیر منفی می‌گذارد. اما به دلیل این که بیشتر دسترسی‌های محک CloudSuite در حافظه نهان سطح آخر از نوع خواندن هستند (که ناشی از رفتار این گونه برنامه‌ها می‌باشد)، مصرف انرژی پویای[6] کل برای مجموع عملیات‌های خواندن و نوشتن در مقایسه با حالتی که در آن از حافظه SRAM به عنوان حافظه نهان سطح آخر استفاده شود، چندان افزایش نمی‌یابد. همچنین با توجه به توان نشتی نزدیک به صفر حافظه STT-RAM، میزان افزایش انرژی مصرفی مربوط به عملیات نوشتن در حافظه STT-RAM جبران خواهد شد.

به دلیل این که فناوری‌های جدید حافظه در حال تکامل می‌باشند، استفاده از آنها در سلسله‌مراتب حافظه فرصت‌های جدیدی برای طراحی حافظه ایجاد می‌کند. ما در این مقاله از فناوری مجتمع‌سازی سه‌بعدی

1. Through Silicon Via
2. Modular
3. Memory Wall Problem
4. Locality
5. Streaming Service
6. Dynamic



شکل ۴: مثالی از یک پردازنده سه‌بعدی با چند لایه حافظه نهان [۳۳].

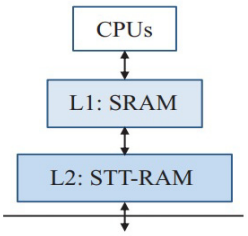


شکل ۵: شِمای کلی معماری پایه مبتنی بر حافظه نهان سطح آخر از نوع SRAM.

برای قراردادن بلوک‌های حافظه نهان سطح دوم (که در اینجا سطح آخر می‌باشد) در بالای تراشه پردازنده استفاده می‌کنیم. با این مجتمع‌سازی سه‌بعدی می‌توانیم همانند شکل ۴ سیستم‌های چندهسته‌ای و حافظه‌های نهان از نوع STT-RAM را به صورت دو لایه مجزا بسازیم که به صورت عمودی روی هم قرار می‌گیرند. هدف معماری پیشنهادی این است که با تغییر نوع و اندازه بانک‌های حافظه نهان سطح آخر، مصرف کلی انرژی در سیستم را کاهش دهیم. برای این کار به جای حافظه‌های نهان سطح دوم (که در اینجا حافظه نهان سطح آخر است) رایج مانند SRAM از حافظه STT-RAM استفاده می‌کنیم. در شکل ۵ که مربوط به معماری پایه مبتنی بر حافظه نهان از نوع SRAM می‌باشد، حافظه SRAM به عنوان حافظه نهان سطح دوم درون تراشه نشان داده شده است.

شکل ۶ نیز سلسله‌مراتب حافظه نهان ناهمگن در معماری پیشنهادی سه‌بعدی را نشان می‌دهد. معماری پیشنهادی سه‌بعدی همانند معماری نشان داده شده در شکل ۴ می‌باشد که در لایه[1] هسته ۱۶ هسته به همراه حافظه نهان سطح اول از نوع SRAM و در لایه حافظه نهان[2] ۱۶ بانک حافظه نهان از نوع STT-RAM وجود دارد. لایه‌ها از طریق TSB[3] به هم متصل شده‌اند. این گذرگاه با TSVها پیاده‌سازی شده است. حافظه STT-RAM ویژگی‌های زیادی دارد که باعث می‌شود یک حافظه مناسب برای استفاده به عنوان حافظه‌های نهان درون تراشه و حافظه‌های اصلی خارج از تراشه انتخاب گردد. اگر به جای حافظه نهان سطح آخر درون

1. Core Layer
2. Cache Layer
3. Through Silicon Bus



شکل ۶: سلسله‌مراتب حافظه نهان ناهمگن در معماری پیشنهادی سه‌بعدی.

تراشه مبتنی بر حافظه SRAM از حافظه STT-RAM استفاده کنیم، می‌توانیم مصرف انرژی را در حافظه نهان سطح آخر کاهش دهیم. توان نشتی در حافظه نهان سطح آخر مبتنی بر STT-RAM نزدیک به صفر است و بنابراین حتی اگر مقیاس‌پذیری فناوری متوقف شود و تعداد هسته‌ها ثابت بماند، تغییر نوع بانک حافظه نهان سطح آخر، انرژی مصرفی کل را که به توان نشتی در آخرین سطح حافظه نهان وابسته است کاهش می‌دهد.

۵- شبیه‌سازی و نتایج

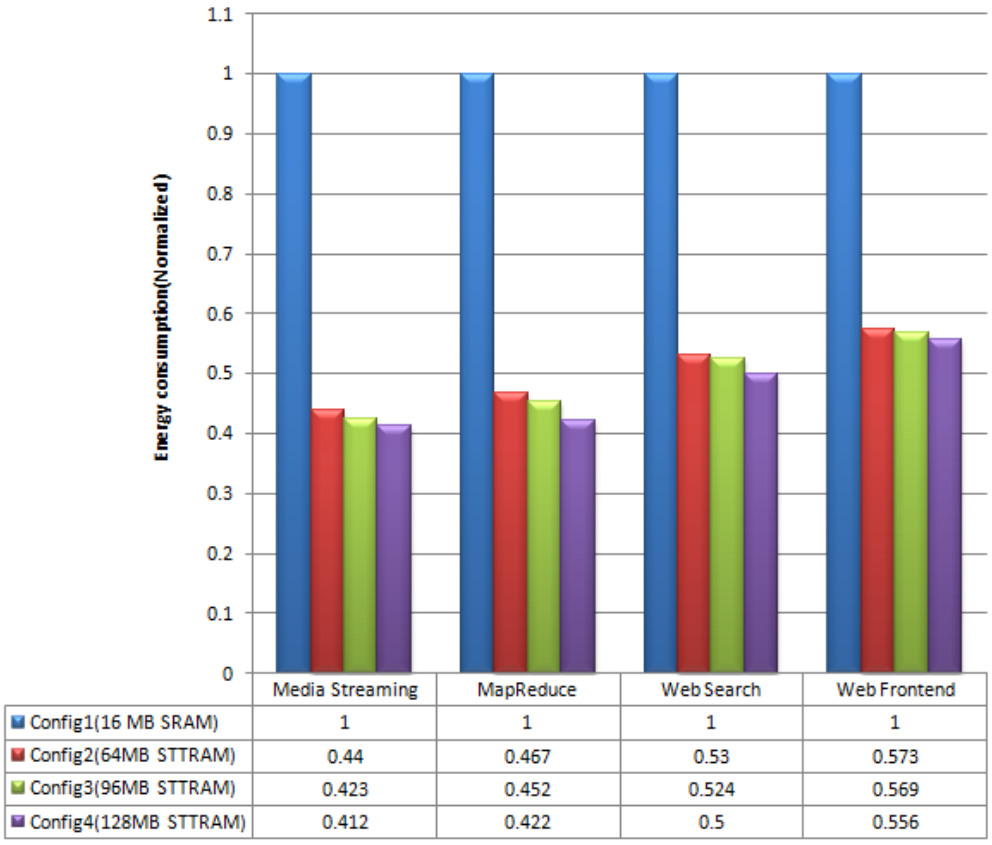
در این مقاله از شبیه‌ساز کامل Flexus [۳۵] برای پیاده‌سازی حافظه‌ها و هسته‌ها استفاده شده است. برای پیاده‌سازی دقیق رفتار طراحی CMP سه‌بعدی و معماری NoC آن، از Flexus همراه با ۳D-Noxim [۳۶] که یک شبیه‌ساز NoC مبتنی بر SystemC است، استفاده گردیده است. علاوه بر این، McPAT [۳۷] نیز همراه با موارد قبلی استفاده شده است تا مصرف انرژی محاسبه شود. از شبیه‌سازهای CACTI [۳۸] و NVSIM [۳۹] نیز به ترتیب برای محاسبه ظرفیت حافظه نهان و مصرف انرژیِ حافظه‌های SRAM و STT-RAM استفاده شده و در اینجا جزئیات SRAM و پیکربندی STT-RAM در جدول ۲ آمده است.

۵-۱ بارهای کاری محک

CloudSuite یک مجموعه محک برای سرویس‌های ابری می‌باشد. در این مقاله ۴ بارکاری scale-out از مجموعه محک CloudSuite مورد استفاده قرار گرفته است. این بارهای کاری عبارتند از MapReduce، Media Streaming، Web Frontend و Web Search.

۵-۲ نتایج

جدول ۳ میزان دسترسی خواندن و نوشتن را در حافظه نهان سطح دوم که در اینجا سطح آخر می‌باشد، برای ۴ بارکاری از محک CloudSuite نشان می‌دهد. نتایج آزمایش نشان می‌دهد که برای هر دو حالت مبتنی بر SRAM و پیکربندی‌های روش پیشنهادی مبتنی بر STT-RAM، حداکثر ۲۰/۴٪ از دسترسی‌ها برای عملیات نوشتن می‌باشد و بقیه دسترسی‌ها برای عملیات خواندن است. بنابراین اکثریت دسترسی‌ها در بارهای کاری ارزیابی‌شده در این مقاله از نوع خواندن هستند. البته در مراجع این مقاله که رفتار بارهای کاری از محک CloudSuite را تحلیل نموده‌اند، این مسأله بیان شده که نسبت درخواست‌های خواندن در تمامی بارهای کاری از این محک به مراتب بیشتر از درخواست‌های نوشتن می‌باشد. طبق جدول ۱، اگر از STT-RAM به عنوان حافظه نهان سطح آخر استفاده کنیم، عملیات نوشتن بر مصرف انرژی، کارایی و تأخیر تأثیر منفی می‌گذارد. اما بر اساس نتایج جدول ۳، چون بیشتر دسترسی‌ها در حافظه نهان سطح آخر در محک CloudSuite از نوع خواندن می‌باشد، اگر در حافظه نهان سطح دوم از حافظه STT-RAM استفاده کنیم،



	Media Streaming	MapReduce	Web Search	Web Frontend
Config1(16 MB SRAM)	1	1	1	1
Config2(64MB STTRAM)	0.44	0.467	0.53	0.573
Config3(96MB STTRAM)	0.423	0.452	0.524	0.569
Config4(128MB STTRAM)	0.412	0.422	0.5	0.556

شکل ۷: مقایسه مصرف انرژی در حافظه نهان سطح آخر در معماری مبتنی بر SRAM و پیکربندی‌های مبتنی بر STT-RAM.

جدول ۲: پارامترهای ارزیابی.

Parameter	Value
Number of Cores	۱۶, ۴×۴Mesh, ARM Cortex-A۱۵
Technology	۳۲ nm, ۰٫۹ V, ۲ GHz
L۱ Private Cache	SRAM, ۳۲ KB per Core
L۲ Cache (LLC)	Config۱ : ۱۶ MB SRAM(۱MB banks on each core) Config۲ : ۶۴ MB STT-RAM(۴MB banks on each core) Config۳ : ۹۶ MB STT-RAM(۶ MB banks on each core) Config۴ : ۱۲۸ MB STT-RAM(۸MB banks on each core)

جدول ۳: درصد دسترسی‌های خواندن و نوشتن مربوط به حافظه نهان سطح دوم در روش مبتنی بر SRAM و پیکربندی‌های مبتنی بر STT-RAM برای بارهای کاری محک CLOUDSUITE.

Workload	Config۱(۱۶ MB SRAM)		Config۲(۶۴ MB STT-RAM)		Config۳(۹۶ MB STT-RAM)		Config۴(۱۲۸ MB STT-RAM)	
	L۲_Read	L۲_Write	L۲_Read	L۲_Write	L۲_Read	L۲_Write	L۲_Read	L۲_Write
Media Streaming	٪۸۶٫۱۰	٪۱۳٫۹۰	٪۸۸٫۶۰	٪۱۱٫۴۰	٪۸۹٫۶۶	٪۱۰٫۳۴	٪۹۰٫۳۳	٪۹٫۶۷
MapReduce	٪۸۴٫۸۰	٪۱۵٫۲۰	٪۸۷٫۳۰	٪۱۲٫۷۰	٪۸۸٫۳۴	٪۱۱٫۶۶	٪۸۹٫۷۳	٪۱۰٫۲۷
Web Search	٪۸۲٫۳۰	٪۱۷٫۷۰	٪۸۳٫۸۰	٪۱۶٫۲۰	٪۸۴٫۲۸	٪۱۵٫۷۲	٪۸۵٫۷۶	٪۱۴٫۲۴
Web Frontend	٪۷۹٫۶۰	٪۲۰٫۴۰	۷۹٫۸۰	٪۲۰٫۲۰	٪۸۰٫۱۸	٪۱۹٫۸۲	٪۸۰٫۴۹	٪۱۹٫۵۱

مصرف کل پویای انرژی برای عملیات خواندن و نوشتن در حافظه نهان سطح آخر، در مقایسه با استفاده از حافظه SRAM افزایش قابل توجهی نمی‌یابد. از سوی دیگر، قابل ذکر است که توان نشتی در محاسبه کل مصرف انرژی در حافظه نهان سطح آخر نقش مهمی دارد و چون مقدار آن در حافظه STT-RAM نزدیک به صفر است، بنابراین مقدار مصرف انرژی کل در حافظه نهان سطح آخر در روش پیشنهادی که در آن از حافظه STT-RAM استفاده شده است، از حالت رایج مبتنی بر SRAM کمتر خواهد بود.

شکل ۷ مصرف انرژی در حافظه نهان سطح آخر برای ۴ بارکاری از محک CloudSuite مربوط به هر کدام از پیکربندی‌های ارائه‌شده در جدول ۲ را نشان می‌دهد. این شکل نشان می‌دهد که معماری پیشنهادی سه‌بعدی مبتنی بر حافظه STT-RAM در مقایسه با معماری مبتنی بر حافظه SRAM، مصرف انرژی در حافظه نهان سطح آخر را کاهش می‌دهد. به دلیل توان نشتی نزدیک به صفر بانک‌های حافظه STT-RAM، انرژی مصرفی حافظه نهان سطح آخر معماری پیشنهادی سه‌بعدی دارای سلسله‌مراتب حافظه نهان ناهمگن کاهش چشم‌گیری

یافته است. نتایج نشان می‌دهد که معماری پیشنهادی برای ۴ بارکاری مورد آزمایش، حداقل ۴۳٪ و حداکثر ۵۹٪ نسبت به طرح مبتنی بر حافظه SRAM در سطح آخر حافظه نهان انرژی کمتری مصرف می‌کند.

همان طور که در شکل ۷ مشاهده می‌شود، در بارهای کاری مربوط به محک CloudSuite با افزایش چند برابری اندازه حافظه نهان سطح آخر، بهبود چندانی حاصل نمی‌گردد. یعنی برخلاف بارهای کاری Desktop و بارهای کاری رایج Server، با افزایش ظرفیت حافظه نهان سطح آخر، بهبود عملکرد در بارهای کاری مربوط به محک CloudSuite افزایش چشم‌گیری پیدا نمی‌کند. دلیل این مسأله حجم بسیار زیاد مجموعه کاری این نوع از بارهای کاری نوظهور است که ناشی از پردازش تنوع بالایی از درخواست‌ها است که منجر می‌شود این بارهای کاری مجموعه کاری دستورالعمل فعال بسیار بزرگی داشته باشند. بارهای کاری نوظهور Scale-out مانند بارهای کاری ارزیابی‌شده در این مقاله که در بسیاری از خدمات آنلاین امروزی وجود دارند، یک دسته از برنامه‌ها هستند که ویژگی‌های مشترکی دارند. این ویژگی‌ها آنها را از برنامه‌های desktop، پردازش رسانه و برنامه‌های کاربردی در زمینه‌های علمی متمایز می‌کند. یک بارکاری Scale-out یک سرویس جاری یا جستجوی وب است. این برنامه، جریانی از درخواست‌های مشتریان را مدیریت می‌کند که اکثراً مستقل هستند و می‌خواهند به بخشی از داده‌های یک مجموعه داده بسیار بزرگ دسترسی داشته باشند.

٦- نتیجه‌گیری

فراهم‌کنندگان سرویس‌های ابری سعی می‌کنند بر محدودیت انرژی غلبه کنند تا بتوانند خدمات خود را گسترش دهند. با این که تحقیقات زیادی برای افزایش کارایی مراکز داده در سطح rack و معماری انجام شده است اما متأسفانه مراکز داده فعلی به دلیل موازنه ضعیف بین انرژی و کارایی، به اندازه کافی برای اجرای بارهای کاری Scale-out کارامد نیستند. بنابراین برای بهبود کارایی مراکز داده و گسترش پردازش ابری، بهره‌وری انرژی باید بهبود یابد. به همین دلیل ما در این مقاله نوع و اندازه بانک‌های حافظه نهان سطح آخر را تغییر دادیم و مصرف انرژی در پردازنده‌های مراکز داده را بررسی کردیم. ما از حافظه غیر فرار STT-RAM در حافظه نهان سطح آخر استفاده کردیم که تراکم بیشتری دارد و به دلیل غیر فرار بودن، توان نشتی در آنها نزدیک به صفر است. نتایج نشان داد که معماری پیشنهادی مبتنی بر حافظه STT-RAM در مقایسه با معماری پایه مبتنی بر حافظه SRAM، مصرف انرژی را در حافظه نهان سطح آخر به طور قابل توجهی کاهش می‌دهد. در تحقیقات آتی می‌توان از یک حافظه غیر فرار دیگر مانند SOT-RAM و یا روش‌های ترکیبی حافظه‌های با تکنولوژی متفاوت، در حافظه نهان سطح آخر برای افزایش طول عمر و کاهش مصرف انرژی استفاده نمود.

مراجع

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *J. of Internet Services and Applications*, vol. 1, no. 1, pp. 7-18, May 2010.

[2] A. Uchechukwu, K. Li, and Y. Shen, "Energy consumption in cloud computing data centers," *International J. of Cloud Computing and Services Science*, vol. 3, no. 3, pp. 31-48, Jun. 2014.

[3] H. Rong, H. Zhang, S. Xiao, C. Li, and C. Hu, "Optimizing energy consumption for data centers," *Renewable and Sustainable Energy Reviews*, vol. 58, no. 1, pp. 674-691, May 2016.

[4] M. Ferdman, *et al.*, "Clearing the clouds: a study of emerging scale-out workloads on modern hardware," in *Proc. of the 17th Conf. on Architectural Support for Programming Languages and Operating Systems, ASPLOS'12*, pp. 37-48, Mar. 2012.

[5] P. Lotfi-Kamran, B. Grot, and B. Falsafi, "NOC-Out: microarchitecting a scale-out processor," in *Proc. of the 45th Annual IEEE/ACM Int. Symp. on Microarchitecture*, pp. 177-187, Vancouver, BC, Canada, 1-5 Dec. 2012.

[6] M. R. Jokar, M. Arjomand, and H. Sarbazi-Azad, "Sequoia: a high-endurance NVM-based cache architecture," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 3, pp. 954-967, Apr. 2015.

[7] P. Lotfi-Kamran, M. Modarressi, and H. Sarbazi-Azad, "An efficient hybrid-switched network-on-chip for chip multiprocessors," *IEEE Trans. on Computers*, vol. 65, no. 5, pp. 1656-1662, Jun. 2016.

[8] V. Karakostas, O. S. Unsal, M. Nemirovsky, A. Cristal, and M. Swift, "Performance analysis of the memory management unit under scale-out workloads," in *Proc. IEEE Int. Symp. on Workload Characterization, IISWC'14*, 12 pp., Raleigh, NC, USA, 26-28 Oct. 2014.

[9] J. Wang, J. Zhang, W. Zhang, K. Qiu, T. Li, and M. Wu, "Near threshold cloud processors for dark silicon mitigation: the impact on emerging scale-out workloads," in *Proc. of the 12th ACM Int, Conf. on Computing Frontiers*, 8 pp., Ischia, Italy, 10-12 May 2015.

[10] A. Pahlevan, *et al*., "Towards near-threshold server processors," in *Proc. IEEE Design, Automation & Test in Europe Conf. & Exhibition, DATE'16*, pp. 7-12, Dresden, Germany,14-18 Mar. 2016.

[11] Z. Wang, D. A. Jimenez, C. Xu, G. Sun, and Y. Xie, "Adaptive placement and migration policy for an STT-RAM-based hybrid cache," in *Proc IEEE 20th Inte. Symp. on High Performance Computer Architecture, HPCA'14*, pp. 13-24, Orlando, FL, USA, 15-19 Feb. 2014.

[12] Y. T. Chen, J. Cong, H. Huang, B. Liu, C. Liu, M. Potkonjak, and G. Reinman, "Dynamically reconfigurable hybrid cache: an energy efficient last-level cache design," *Proc. IEEE Design, Automation & Test in Europe Conf. & Exhibition, DATE'12*, pp. 45-50, Dresden, Germany, 12-16 Mar. 2012.

[13] J. Ahn, S. Yoo, and K. Choi, "Prediction hybrid cache: an energy-efficient STT-RAM cache architecture," *IEEE Trans. on Computer*, vol. 65, no. 3, pp. 940-951, May 2015.

[14] A. Valero, J. Sahuquillo, S. Petit, P. Lopez, and J. Duato, "Design of hybrid second-level caches," *IEEE Trans. on Computers*, vol. 64, no. 7, pp. 1884-1897, Aug. 2015.

[15] Z. Zhou, L. Ju, Z. Jia, and X. Li, "Managing hybrid on-chip scratchpad and cache memories for multi-tasking embedded systems," in *Proc. 20th Asia and South Pacific Design Automation Conf., ASP-DAC'15*, pp. 423-428, Chiba, Japan, 19-22 Jan. 2015.

[16] D. Jevdjic, G. H. Loh, C. Kaynak, and B. Falsafi, "Unison cache: a scalable and effective die-stacked DRAM cache," in *Proc. 47th Annual IEEE/ACM Int. Symp. on Microarchitecture,* , pp. 25-37, Cambridge, UK, 13-17 Dec.? Dec. 2014.

[17] S. Onsori, A. Asad, K. Raahemifar, and M. Fathy, "Notice of violation of IEEE publication principles: An energy-efficient heterogeneous memory architecture for future dark silicon embedded chip-multiprocessors," *IEEE Trans. on Emerging Topics in Computing*, vol. 4, no. 2, p. 1, May 2015.

[18] A. Asad, A. Dorostkar, and F. Mohammadi., "A novel power model for future heterogeneous 3D chip-multiprocessors in the dark silicon age," *EURASIP Journal on Embedded Systems*, vol. 12, no. 1, pp. 1-16, Dec. 2016.

[19] S. Onsori, A. Asad, K. Raahemifar, and M. Fathy, "OptMem: dark-silicon aware low latency hybrid memory design," in *Proc IEEE. International Conf. on VLSI Systems, Architectures, Technology and Applications, VLSI-SATA'16,* 5 pp., Bangalore, India, 10-12 Jan. 2016.

[20] S. Onsori, A. Asad, O. Ozturk, and M. Fathy, "Hybrid stacked memory architecture for energy efficient embedded chip-multiprocessors based on compiler directed approach," in *Proc. IEEE 6th Int. Green Computing Conf. and Sustainable Computing Conf., IGSC'15,* 7 pp., Las Vegas, NV, USA, 14-16 Dec. 2015.

[21] S. Senni, L. Torres, G. Sassatelli, A. Gamatie, and B. Mussard, "Exploring MRAM technologies for energy efficient systems-on-chip," *IEEE J. on Emerging and Selected Topics in Circuits and Systems*, vol.6, no. 3pp. 279-292, Apr. 2016.

[22] -, *CloudSuite 1.0*, 2012. [Online]. Available: http://parsa.epfl.ch/cloudsuite

[23] J. S. Vetter and S. Mittal, "Opportunities for nonvolatile memory systems in extreme-scale high-performance computing," *Computing in Science & Engineering*, vol. 17, no. 2, pp. 73-82, Jan. 2015.

[24] D. Jevdjic, S. Volos, and B. Falsafi, "Die-stacked DRAM caches for servers: hit ratio, latency, or bandwidth? Have it all with footprint

cache," *ACM SIGARCH Computer Architecture News*, vol. 41, no. 3, pp. 404-415, Jun. 2013.

[25] J. Park, J. Jung, K. Yi, and C. M. Kyung, "Static energy minimization of 3D stacked L2 cache with selective cache compression," in *Proc. IFIP/IEEE 21st Int. Conf. on Very Large Scale Integration, VLSI-SoC'13*, pp. 228-233, Istanbul, Turkey, 7-9 Oct. 2013.

[26] M. H. Samavatian, H. Abbasitabar, M. Arjomand, and H. Sarbazi-Azad, "An efficient STT-RAM last level cache architecture for GPUs," in *Proc. of the 51st Annual Design Automation Conf.*, 6 pp., San Francisco, CA, USA, 1-5 Jun. 2014.

[27] M. Bakhshalipour, *et al*., "Reducing writebacks through in-cache displacement," *ACM Trans. on Design Automation of Electronic Systems*, vol. 24, no. 2, pp. 1-21, Jan. 2019.

[28] S. Rashidi, M. Jalili, and H. Sarbazi-Azad, "A survey on pcm lifetime enhancement schemes," *ACM Computing Surveys*, vol. 52, no. 4, pp. 1-38, Aug. 2019.

[29] S. Rashidi, M. Jalili, and H. Sarbazi-Azad, "Improving MLC PCM performance through relaxed write and read for intermediate resistance levels," *ACM Trans. on Architecture and Code Optimization*, vol. 15, no. 1, pp. 1-31, Mar. 2018.

[30] C. Qian, L. Huang, P. Xie, N. Xiao, and Z. Wang, "A study on non-volatile 3d stacked memory for big data applications," in *Proc. Int. Conf. on Algorithms and Architectures for Parallel Processing,* pp. 103-118, Zhangjiajie, China, 18-20 Nov. 2015.

[31] M. Hosomi, *et al.*, "A novel non-volatile memory with spin torque transfer magnetization switching: spin-ram," in *Proc. IEEE Int. Electron Devices Meeting, IEDM Technical Digest.*, pp. 459-462, Washington, DC, USA, 5-5 Dec. 2005.

[32] A. K. Mishra, X. Dong, G. Sun, Y. Xie, N. Vijaykrishnan, and C. R. Das, "Architecting on-chip interconnects for stacked 3D STT-RAM caches in CMPs," *ACM SIGARCH Computer Architecture News, ACM*, vol. 39, no. 3, pp. 69-80, Jun. 2011.

[33] Q. Wang, L. Shen, and Z. Wang, "Research on scale-out workloads and optimal design of multicore processors," in *Proc. of Int. Conf. on Soft Computing Techniques and Engineering Application*, pp. 157-166, Kunming, China, 25-27 Sept. 2013.

[34] E. Chen, D. Lottis, A. Driskill-Smith, D. Druist, V. Nikitin, S. Watts, X. Tang, and D. Apalkov, "Nonvolatile spin-transfer torque RAM (STT-RAM)," in *Proc. IEEE Device Research Conf., DRC'10*, pp. 249-252, South Bend, IN, USA, 21-23 Jun. 2010.

[35] T. Wenisch, *et al*., "SimFlex: statistical sampling of computer system simulation," *IEEE Micro*, vol. 26, no. 4, pp. 18-31, Jul./Aug. 2006.

[36] M. Palesi, S. Kumar, and D. Patti, Noxim: Network-on-Chip Simulator, http://noxim.sourceforge.net, 2010.

[37] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. Annual IEEE/ACM Int. Symp. on MICRO-42*, pp. 469-480, New York, NY, USA, 12-16 Dec. 2009.

[38] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "CACTI 6.0: a tool to model large caches," HP Laboratories, Technical Report, 2009.

[39] X. Dong, C. Xu, N. Jouppi, and Y. Xie, "NVSim: a circuit-level performance, energy, and area model for emerging non-volatile memory," *IEEE Trans, on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994-1007, Jun. 2012.

عدنان نصری در سال ۱۳۸۵ مدرک کارشناسی مهندسی کامپیوتر- سخت‌افزار خود را از دانشگاه علم و صنعت ایران و در سال ۱۳۸۷ مدرک کارشناسی ارشد مهندسی کامپیوتر- معماری کامپیوتر خود را از دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران دریافت نمود. هم‌اکنون دانشجوی دکتری مهندسی کامپیوتر- معماری کامپیوتر دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران می‌باشد و زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از معماری کامپیوتر، محاسبات ابری و شبکه‌های حسگر.

محمود فتحی در سال ۱۳۶۳ مدرک کارشناسی مهندسی برق خود را از دانشگاه علم و صنعت ایران، در سال ۱۳۶۵ مدرک کارشناسی ارشد مهندسی میکروپروسسور خود را از دانشگاه بردفورد انگلستان و در سال ۱۳۷۰ در رشته معماری کامپیوتر و پردازش تصویر مدرک دکتری خود را از دانشگاه منچستر دریافت نمود. ایشان هم‌اکنون استاد دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران می‌باشند. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از معماری کامپیوتر، پردازش تصویر، یادگیری ماشین و یادگیری عمیق.

علی برومندنیا در سال ۱۳۷۱ مدرک کارشناسی مهندسی کامپیوتر- سخت‌افزار خود را از دانشگاه صنعتی اصفهان و در سال ۱۳۷۴ مدرک کارشناسی ارشد مهندسی کامپیوتر- معماری کامپیوتر خود را از دانشگاه علم و صنعت ایران و مدرک دکتری مهندسی کامپیوتر را در سال ۱۳۸۵ از دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران دریافت نمود. هم‌اکنون ایشان عضو هیأت علمی دانشگاه آزاد اسلامی واحد تهران جنوب می‌باشد و زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از بینایی ماشین و شناسایی الگو.