بسمه تعالی

☐ دفاع از رساله دکتری ☐ سمینار عمومی (Colloquium)

■ دفاع از پایان‌نامه کارشناسی‌ارشد ☐ سمینار تخصصی (Seminar)

☐ سمینار تخصصی و مشورتی (Informal Seminar)

عنوان : یک سیاست جایگزینی نوین برای مصالحه انرژی و کارایی در حافظه‌های نهان مشترک سطح آخر محیط‌های ناهمگون

سخنران : امیررضا شمسی

چکیده:

مجتمع‌سازی هسته‌های CPU تاخیر-گرا و هسته‌های GPU ظرفیت-گذردهی-گرا بر روی یک تراشه بدل به یکی از رایج‌ترین گرایش‌های معمارانه شده است. یکی از جنبه‌های کلیدی مدیریت حافظه‌ی نهان سطح آخر (LLC) مشترک، سیاست جایگزینی آن است. وقتی توازی-سازی در سطح ریسمان (TLP) کافی وجود داشته باشد، برنامه‌های کاربردی GPU قادر به تحمل تاخیر دستیابی به حافظه می‌باشند. از آنجایی که وجود هسته‌های گوناگون بر روی یک تراشه، انرژی-کارآمدی پردازنده را تحت تاثیر قرار می‌دهد، علاوه بر کارایی کلی سیستم، ملاحظات مصرف انرژی را هم بایستی برای محیط‌های ناهمگون در نظر گرفت. این پایان‌نامه سه سیاست جایگزینی نوین برای حفظ جمعیت بسیار مفید در LLC مشترک ارائه می‌دهد. اولا، LS-LRU که یک LRU با سربار سخت‌افزاری کم و پویش-مقاوم مبتنی بر پیش‌بینی بازه‌ی مراجعه‌ی مجدد (RRIP) می‌باشد پیشنهاد شده است. در ادامه، برای درج/ارتقای تطبیقی بلوک‌ها بر مبنای درجه‌ی انجمنی، استراتژی درجه-انجمنی-آگاه درج در میانه/ارتقا به اندازه‌ی یک چهارم (HIQP) را پیشنهاد می‌نماییم. LS-LRU به منظور درج بلوک‌ها در نیمه‌ی پایینی زنجیره‌ی جایگزینی و ارتقای آنها بر اساس درجه‌ی انجمنی LLC هنگام مواجهه با اصابت، از HIQP استفاده می‌نماید. به دیگر سخن، LS-LRU توانایی آن را دارد که مکان درج و بازه‌ی ارتقا را بر اساس درجه‌ی انجمنی حافظه‌ی نهان کنترل کند. در مقام مقایسه با LRU، این سیاست زمان اقامت بلوک‌های بی‌استفاده و تعداد فقدان‌هایی را که به دلیل پویش‌ها (داده‌های غیر محلی) در حافظه‌ی نهان رخ می‌دهند را کاهش می‌دهد. LS-LRU سرعت ارتقای بلوک‌ها به مکان MRU را کاهش داده و رفتار استفاده‌ی مجدد هر کدام از بلوک‌های حافظه‌ی نهان را با دقت بیشتری از LRU آموخته و قادر است بین تاخر و تعدد بلوک‌ها مصالحه برقرار کند. برای یک حافظه‌ی نهان انجمنی-گروهی n-انتخابی، LS-LRU تعداد بیت‌های مورد نیاز به ازای هر گروه حافظه‌ی نهان را از $n \times \log_2 n$ به $4n$ کاهش می‌دهد. دوما، TLS-LRU که یک LS-LRU انحطاط-مقاوم است پیشنهاد شده است. TLS-LRU با استفاده از یک استراتژی ریزدانه، بلوک‌های بی‌استفاده را از LLC با هشت درجه‌ی گوناگون کنار می‌گذارد. به دیگر سخن، بر اساس الگوی دستیابی بار کاری در حال اجرا، درجه‌ی کنارگذاشتن از LLC به صورت تطبیقی تنظیم خواهد شد. سوما، HLS-LRU که یک سیاست جایگزینی نوین برای خطاب قرار دادن ملاحظات کارایی-انرژی در محیط‌های ناهمگون می‌باشد پیشنهاد شده است. این سیاست جایگزینی مبتنی بر LS-LRU، رقابتی که به دلیل نرخ‌های دستیابی متفاوت هسته‌های متنوع بر سر حافظه‌ی نهان ایجاد می‌شود را کاهش می‌دهد. HLS-LRU صریحا اولویت برنامه‌های کاربردی GPU در حافظه‌ی نهان را به مکان‌های دارای اولویت پایین‌تر از MRU محدود می‌نماید. این سیاست، به یک سیاق شبه-افراز، با فراهم نمودن مزیت سرقت ظرفیت از برنامه‌های کاربردی GPU حتی بیش از این به برنامه‌های کاربردی CPU اولویت می‌بخشد. HLS-LRU برای تمایز میان برنامه‌های کاربردی حساس و ناحساس به حافظه‌ی نهان GPU و بهبود تصمیمات جایگزینی از روش نمونه‌برداری از هسته استفاده می‌نماید. کنارگذاشتن برنامه‌های کاربردی ناحساس به حافظه‌ی نهان GPU از LLC، تعداد کل دستیابی‌ها به LLC را کاهش داده و بنابراین، انرژی پویایی که برای دستیابی به این بلوک‌های داده‌ی ناحساس به حافظه‌ی نهان مصرف می‌شود را حذف خواهد نمود. در مقام مقایسه با TAP-RRIP، HLS-LRU کارایی بهتری داشته و نتیجتا، مصرف انرژی ایستا را کاهش می‌دهد. نتایج تجربی نشان می‌دهند که HLS-LRU نسبت به TAP-RRIP کارایی را به میزان ٪۱۵ افزایش داده و مصرف کلی انرژی را به میزان ٪۱/۸ کاهش می‌دهد.

کلمات کلیدی: محیط‌های ناهمگون، سیاست جایگزینی حافظه نهان، حافظه نهان سطح آخر، حافظه نهان مشترک، سیاست درج، سیاست ارتقا، پویش-مقاوم، انحطاط-مقاوم، بلوک‌های بی‌استفاده

زمان برگزاری: ۹۶/۰۶/۰۴

مکان برگزاری: دانشکده مهندسی و علوم کامپیوتر