عنوان مقاله:

زمانبندی توزیع شده وظایف در سیستم های سرویس دهی مبتنی بر GPU بر حسب تقاضا

محل انتشار:

فصلنامه مهندسی برق دانشگاه تبریز, دوره 54, شماره 2 (سال: 1403)

تعداد صفحات اصل مقاله: 10

نویسندگان:

آرزو جهانی – استادیار، دانشکده مهندسی برق، دانشگاه صنعتی سهند، تبریز، ایران

لیلا سادات مومنی – دانشجوی کارشناسی ارشد، دانشکده مهندسی برق، دانشگاه صنعتی سهند، تبریز، ایران

خلاصه مقاله:

زمانبندی بهینه منابع بر روی سرورهای مبتنی بر GPU که برای وظایف موازی مناسب هستند، بسیار ضروری است. این منابع معمولا دارای سرعت بالایی بوده و بنابراین هزینه بالایی نیز دارند. جهت استفاده بهینه از این منابع، مراکز ارائه دهنده خدمات، باید بتوانند به ازای هر درخواست، بهترین نوع ماشین مجازی، بهترین نوع پردازندهGPU و همچنین بهترین تعداد این نوع پردازنده را انتخاب نمایند. چنین مسئلهای، یک مسئله بهینهسازی نامیده میشود. مقاله حاضر، ضمن مدلسازی مسئله تخصیص منابع به عنوان یک مسئله بهینهسازی خطی، روش جدیدی را برای توزیع درخواستها ارایه میدهد. روش پیشنهادی از یک صف مرکزی استفاده نموده و سپس درخواستها را با استفاده از یک روش نوین توزیع درخواست، بین چندین صف محلی توزیع میکند. سپس وظایف موجود در هر صف محلی را به صورت موازی زمانبندی و اجرا میکند. زمانبندی در هر صف محلی، تعیین میکند که به ازای هر درخواست: (۱) بهترین نوع ماشین مجازی (۲) بهترین نوع پردازنده GPU و (۳) بهترین تعداد پردازندههای GPU کدام است. مقایسه روش پیشنهادی با آخرین روشهای موجود، نشانگر کاهش زمان اجرا، کاهش زمان پاسخ و همچنین کاهش چشمگیر هزینه استفاده از منابع در روش پیشنهادی است.

کلمات کلیدی:

زمانبندی وظایف, سرورهای مبتنی بر GPU, توزیع درخواستها, صف محلی