Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 11, Number 1, Spring 2022, pp. 49-59
Received: 08/26/2021, Revised: 01/11/2022, Accepted: 04/04/2022

Scheduling and Resource Allocation based on Priority and SLA in Cloud Computing

Shiva Razzaghzadeh[1*], Farzaneh seyed soleymani[2] and Parisa Norouzi Kivi[3]
1*- Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran
2- Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran.
3- Young Researchers Club, Ardabil Branch, Islamic Azad University of Ardabil, Ardabil, Iran.
[1*]shiva.razzaghzadeh@srbiau.ac.ir, [2]f.soleymani681212@gmail.com, and [3]p.norouzi_2014@yahoo.com

Corresponding author's address: Shiva Razzaghzadeh, Faculty of Electrical and Computer Engineering, Islamic Azad University of Ardabil, Ardabil, Iran.

Abstract- In recent decades, with the popularity and application of cloud computing, significant changes have taken place in communications and technologies. Most service-level service providers are offering applications and developing their own hardware infrastructure as service improvements. Since cloud computing provides different resources to these providers, of course, the cost of access, speed, and other important parameters have led to a significant response, but an important point has given the increase and volume of requests from stakeholders. It has led to challenges in providing services at the service level. Therefore, scheduling and allocating resources to requests made with low-cost horizons and completion time has become a necessity, and service providers and stakeholders seek to receive the best possible service with high efficiency, and this has led to extensive research in this area. In this research, a model for scheduling and resource allocation considering priority and SLA in cloud computing is presented. In fact, the proposed model of several different levels of access has been developed to achieve the main goal of the research, which is the optimal scheduling and allocation of resources to the requests made. In the proposed method, using the RR algorithm and the technique of weighting requests and online review of virtual machines, the best possible source for proposal allocation and scheduling has been identified. The method presented in the form of a dynamic and executable model with the help of the Cloudsim simulation tool and in terms of Makespan, cost, and speedprocess have been compared and analyzed with several similar methods. The results obtained from the simulation performed by applying different scenarios indicate average processing speed around 2.15, and average Makespan is reduced at 8.68s by new method than similarity methods. Also, the rang of cost has not big change.
Keywords- Cloud Computing, Resource Allocation, Scheduling, Service Level Agreement.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۱، شماره ۱، بهار ۱۴۰۱ ، صفحه ۴۹-۵۹
دریافت: ۱۴۰۰/۰۶/۰۲، بازنگری: ۱۴۰۰/۱۱/۱۷، پذیرش: ۱۴۰۱/۰۱/۱۵

زمانبندی و تخصیص منبع با در نظر گرفتن اولویت و SLAدر رایانش ابری

شیوارزاق‌زاده*[۱]،فرزانه سید سلیمانی[۲]،پریسا نوروزی کیوی[۳]
*۱- گروه مهندسی کامپیوتر، واحد اردبیل، دانشگاه آزاد اسلامی، اردبیل، ایران.
۲- گروه مهندسی کامپیوتر، واحد اردبیل، دانشگاه آزاد اسلامی، اردبیل، ایران.
۳- باشگاه پژوهشگران جوان و نخبگان، واحد اردبیل، دانشگاه آزاد اسلامی، اردبیل، ایران.
[1*]shiva.razzaghzadeh@srbiau.ac.ir, [2]f.soleymani681212@gmail.com, [3]p.norouzi_2014@yahoo.com

* نشانی نویسنده مسئول: شیوا رزاق زاده، اردبیل، میدان بسیج، دانشگاه آزاد اسلامی واحد اردبیل، دانشکده مهندسی کامپیوتر.

چکیده- در چند دهه‌ی اخیر با محبوبیت و کاربردی شدن رایانش ابری، تغییرات قابل توجهی در ارتباطات و فناوری‌ها شکل گرفته است. به صورتی که بیشتر سرویس‌دهندگان خدمات در سطح سرویس، در حال ارائه برنامه‌ها و توسعه زیرساخت‌های سخت‌افزاری خود بعنوان بهبود خدمات هستند. از آنجایی که رایانش ابری منابع مختلف را در اختیار این سرویس‌دهندگان قرار می‌دهد طبعا هزینه دسترسی، سرعت و دیگر پارامترهای مهم منجر به استقبال چشم‌گیر شده است; اما نکته‌ی مهم با توجه به ازدیاد و حجم درخواست‌های ذینفعان منجر به بروز چالش‌هایی در ارائه خدمات در سطح سرویس شده است. لذا زمانبندی و تخصیص منابع به درخواست‌های انجام شده با افق هزینه و زمان اتمام کار پایین، یک ضرورت شده است و سرویس‌دهندگان به دنبال دریافت بهترین سرویس ممکن با کارایی بالا هستند. این امر منجر به تحقیقات گسترده در این حوزه شده است. روش پیشنهادی مدلی برای زمانبندی و تخصیص منابع با در نظر گرفتن اولویت و توافق نامه سطح سرویس در رایانش ابری ارائه شده است. در حقیقت مدل ارائه شده در چند سطح مختلف دسترسی برای محقق کردن هدف اصلی، که زمانبندی و تخصیص بهینه منابع به درخواست‌ها می‌باشد، توسعه داده شده است. در روش ارائه شده به کمک الگوریتم نوبت گردشی و تکنیک وزن‌دهی به درخواست‌ها و نیز بررسی برخط ماشین‌های مجازی بهترین منبع ممکن برای تخصیص پیشنهاد و زمانبندی شده است. روش ارائه شده در قالب یک مدل پویا و قابل اجرا به کمک ابزار کلودسیم شبیه‌سازی و پارامترهای طول زمانبندی، هزینه و سرعت پردازش با چند روش مشابه مقایسه و آنالیز شده است. نتایج بدست آمده از سناریوهای مختلف بیانگر کاهش ۸.۶۸ ثانیه در متوسط طول زمانبندی و افزایش۲.۱۵ در متوسط سرعت پردازش است. همچنین هزینه پردازش تغییر قابل توجهی نداشته است.

واژه‌های کلیدی:رایانش ابری، نخصیص منابع، زمانبندی، توافق‌نامه سطح سرویس.

۱- مقدمه

با پیشرفت فناوری اطلاعات، نیاز به دسترسی به اطلاعات و انجام کارهای محاسباتی در هرمکان و زمان افزایش یافته است. یکی از جدیدترین تغییرات در نحوه کارکرد اینترنت، با معرفی رایانش ابری[۱] صورت پذیرفته است. این فناوری جدید به دلیل ویژگی‌هایش به سرعت محبوب شده است؛ چرا که در رایانش ابری همه نوع امکانات به کاربران، بعنوان یک سرویس ارائه می‌شود[1] . رایانش ابری یک فناوری شتاب‌دهنده در زمینه محاسبات توزیعی است که می‌توان در برنامه‌های ذخیره‌سازی داده‌ها، تجزیه و تحلیل داده‌ها و اینترنت اشیا استفاده کرد. این فناوری روش‌های سنتی استقرار خدمات توسط شرکت‌ها یا افراد را تغییر داده است. انواع مختلفی از خدمات را بعنوان

سرویس‌های وب به کاربران ارائه می‌دهد تا آنها نیازی به سرمایه‌گذاری در زیرساخت‌های محاسباتی نداشته باشند. رایانش ابری خدماتی مانند زیرساخت به عنوان سرویس[۲]، بستر به عنوان سرویس[۳] و نرم‌افزار به عنوان سرویس[۴] را ارائه می‌دهد [2].

با افزایش محبوبیت رایانش ابری بهره‌برداران شروع به برون‌سپاری خواسته‌های محاسباتی خود به سرویس‌های ابری کرده‌اند. ابرها به طور معمول مراکز داده مجازی[۵] در مقیاس بسیار بزرگ هستند که تعداد زیادی سرور را میزبانی می‌کنند. برای اینکه این محیط‌ها بتوانند به موفقیت‌های تجاری دست یابند، نیاز است تا تضمین‌های کیفیت سرویس[۶] بهتر و دقیق‌تری را فراهم کنند، این تضمین‌ها که به عنوان ضمانت نامه سطح خدمت، مستند می‌شوند بسیار حیاتی هستند، چرا که تنها در این صورت کاربران به ابرها اطمینان خواهند کرد. بنابراین رعایت اهداف SLA[۷](توافق نامه سطح سرویس) به منظور کسب رضایت مشتری و نیز افزایش سود فراهم‌کننده از موضوعات مهم این محیط‌ها محسوب می‌شوند. همچنین تفاوت‌های ماهیتی بین بارهای کاری متفاوت باعث می‌شود تهیه منبع کاری دشوار باشد. برای تهیه منبع روش‌های مختلفی وجود دارد، از جمله روش‌های مبتنی بر ضمانت نامه سطح خدمت، روش‌های بازارگرا [۸]، روش‌های مبتنی بر قانون و روش‌های مبتنی بر کاهش مصرف انرژی. هر یک از روش‌های مذکور روی موضوع خاصی تمرکز دارند[3] . در روش پیشنهادی، تخصیص منبع با روش مبتنی بر SLA مورد بررسی قرار گرفته است.

بطور کلی، SLA از دهه ۱۹۸۰ در حوزه‌های مختلفی استفاده شده است و بیشتر تعاریف در دسترس، مفهومی بوده و در هر حوزه متفاوت با دیگری است. موضوعات اساسی در مدیریت SLA، شامل مدیریت خودکار SLA و سازش بین چندین پارامتر QoS است. دستیابی به QoS برای رعایت SLA بسیار مهم است، چرا که میزان رضایت‌مندی کاربر به میزان رعایت SLA بستگی دارد.

فراهم‌کنندگان سرویس‌های ابری باید به گونه‌ای قادر به زمانبندی منابع و بکارگیری ابزارهای کاربردی باشند، که موارد ذکر شده در SLA را برطرف و کارایی ابزارهای کاربردی را بهینه کنند و هزینه‌ی تهیه منبع را کاهش دهند. الگوریتم‌های بسیاری در این زمینه وجود دارند که برخی از آنها تنها یک پارامتر در SLA را در نظر گرفته‌اند، اما روش‌هایی که چندین پارامتر را مورد توجه قرار داده‌اند بالاتر است. در بررسی نیازهای SLA، ابزارهای متفاوتی دخیل هستند: از جمله: میزان CPU، RAM، فضای ذخیره‌سازی، پهنای باند و زمان پاسخ. از طرفی در زمان اجرا، ممکن است شرایط غیر قابل پیش‌بینی از قبیل خطا در تخمین و بار کاری پویا رخ دهند، از اینرو روش‌های متفاوتی برای تخصیص منبع مبتنی بر SLA ارائه شده است. برای مثال در روش تخصیص ایستا برای تامین نیازهای SLA، فراهم‌کنندگان برای هر درخواستی یک VM[۹] مجزا راه‌اندازی می‌کنند، در این صورت اگرچه ممکن است نیازهای سطح سرویس مثل زمان پاسخ و ظرفیت برطرف شوند، اما منابع سخت‌افزاری هدر می‌روند که در نهایت منجر به افزایش هزینه زیرساخت می‌شود، چرا که ممکن است درخواست‌ها، تمام منابع را استفاده نکنند. برای برطرف کردن این مشکل روش‌های چند اجاره‌ای[۱۰] مطرح شدند که در آنها یک VM می‌تواند بر اساس میزان منابعش چندین درخواست را سرویس‌دهی کند، اما باید دقت شود که SLA درخواست‌های مختلف رعایت شود و چنانچه بتوان روش‌هایی ارائه داد که این مهم را رعایت کنند به کارایی بالاتری می‌توان دست یافت (تخصیص‌های پویا)[4] .

در روش پیشنهادی یک روش زمانبندی با در نظر گرفتن اولویت درخواست‌های ورودی پیشنهاد شده است که برای استقرار درخواست‌های کاربر روی VMها می‌باشد. به طور کلی زمانبندی یک فرآیند نگاشت وظایف به منابع در دسترس بر پایه نیازمندی‌ها است. مسئله زمانبندی در رایانش ابری،یک مسئله بسیار مهم و از رده مسائل NPمحسوب می‌شود که سعی دارد یک زمانبندی بهینه برای اجرای وظایف و تخصیص بهینه منابع مشخص نماید به شکلی که در زمان کمتر وظایف بیشتری را بتوان پردازش کرد. درخواست‌هایی که توسط کاربر ارائه می‌شود از نظر اولویت، بارکاری و پارامترهای SLA در سطح یکسانی قرار ندارند. به این صورت که اگر کاربر نیاز به پردازش داده داشته باشد آن درخواست به عنوان درخواست الویت بالا و اگر درخواست، اجرای وظایف غیربحرانی باشد به عنوان درخواست الویت پایین کلاس‌بندی می‌شود و به تمام درخواست‌ها بترتیب ورودشان منابع تخصیص داده می‌شوند. از سوی دیگر کاربران علاقه‌مند هستند که کارهایشان در کمترین زمان ممکن و کمترین هزینه به اتمام برسد؛ بعبارتی اولویت مطابق با SLA انجام بپذیرد. از طرف دیگر، سرویس‌دهنده‌ی ابر نیز تمایل دارد میزان استفاده از منابع خود را به حداکثر برساند و همچنین میزان سود خود را افزایش دهد که این دو در تضاد با یکدیگر هستند و معمولاً با روش‌های سنتی اختصاص منابع و مکانیزم‌های زمانبندی موجود همخوانی ندارد. از این رو، روش پیشنهادی با در نظر گرفتن الزاما تذکر شده در SLA، با ارائه‌ی مدل ریاضی قطعی یکپارچه، وظایف را با بیشترین میزان

بهره‌برداری از منابع و با کمترین هزینه و در کمترین زمان، تخصیص می‌دهد. نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی می‌تواند هزینه و زمان اتمام را کاهش دهد.

روند ادامه‌ی مقاله به این شرح می‌باشد که در بخش دوم کارهای مرتبط ارائه می‌شود. بخش سوم به رویکرد پیشنهادی و بخش چهارم به نتایج شبیه‌سازی بر روی روش پیشنهادی اختصاص دارد و در بخش پایانی یعنی بخش پنجم نتیجه گیری از این مقاله ارائه خواهد شد.

۲- کارهای مرتبط

باتوجه به اینکه،زمانبندی و تخصیص منبع یکی از مهمترین نیازهای سیستم‌های ابری است. درسال ۲۰۱۱،وینسنت سی و همکارانش یک زمانبند اکتشافی جدید با در نظر گرفتن اهداف پارامتر SLA متعدد مانند مقدار CPU مورد نیاز، پهنای باند شبکه و ذخیره‌سازی برای استقرار برنامه‌های کاربردی در ابرها ارائه می‌کنند. این مکانیسم شامل یک متعادل کننده بار برای توزیع کارآمد اجرای برنامه‌ها بر روی منابع Cloud است. همچنین لیست منابع در دسترس و لیست ماشین‌های مجازی موجود قبل از تخصیص منبع را بررسی می‌کند. طراحی این مدل شامل پیاده سازی و ارزیابی آن با استفاده از ابزار شبیه‌سازی کلودسیم است[5]. در سال۲۰۱۲ یک استراتژی درخواست جدید براساس نیازهای QoS توسط لیلین یو و همکارانش ارائه شده است. در این استراتژی بررسی می‌شود که آیا درخواست جدید می‌تواند بر اساس نیازهای QoS و قابلیت‌های منابع پذیرفته شود یا نه. در این مدل چهار استراتژی: شروع زمان کار جدید Vm، انتظار، درج و تاخیر جریمه مورد توجه است. نتایج شبیه‌سازی از این الگوریتم پیشنهادی بهبود ٪۴۰ درهزینه را ارائه می‌دهد [6]. الگوریتم مورد مقایسه بعدی توسط لیلین یو و همکارانش در سال ۲۰۱۱ ارائه شده است، در اینجا استراتژی درخواست مشتری برای سرویس‌های نرم‌افزار ERP با توافق بر SLA از پیش تعریف شده است و ثبت پارامترهای QoS مشتری مد نظر قرار گرفته است. هدف آن، افزایش سود ارائه‌کننده SaaS با کاهش هزینه VMها با استفاده از استراتژی‌های تخصیص منبع موثر، از لایه بستر است [7]. درمدل پیش‌بینی نیازهای منبع برای ماکزیمم سازی سودمندی، با توجه به متفاوت بودن SLA مورد بررسی قرار گرفته است(۲۰۱۳) که تخصیص منبع بر اساس بار کاری و معیارهای سطح سرویس (مثل زمان پاسخ و تعداد درخواست‌ها در هر ثانیه) می باشد [8].

در مقاله سوراب کومار گارج، از مشکل تخصیص منابع در یک مرکز داده که انواع مختلفی از بارهای کاری برنامه را اجرا می‌کند، به ویژه برنامه‌های کاربردی غیر تعاملی و تراکنشی، بحث شده است. ما مکانیسم کنترل پذیرش و زمان‌بندی را پیشنهاد می‌کنند که نه تنها استفاده از منابع و سود را به حداکثر می‌رساند، بلکه الزامات SLA کاربران را نیز تضمین می‌کند.

در مدل پیش‌بینی مبتنی بر شبکه عصبی مصنوعی[11] و الگوریتم انتشار به عقب[12] ، سودمندی یک روز هر VM از داده‌های یک هفته از یک مجموعه داده را با شبکه حداقل خطای مربع میانه[13] ریشه نشان می‌دهد. ایده اصلی این است که تقاضای منبع در پنجره زمانی فعلی مانیتور شود تا بتوان درباره تخصیص سرور و نیز ورود کار در پنجره زمانی بعدی تصمیم‌گیری کرد. در هر چرخه زمانبندی، کنترل ورودی و زمانبندی می‌تواند سه عملکرد را انجام دهد که شامل کنترل ورودی، اجرای SLA و مقیاس‌پذیری خودکار اس[9] .همچنین باتوجه به اینکه حفظ توازن بار در محیط رایانش ابری یک مسئله مهم است. در این مقاله ندا نیل ساز دزفولی و مریم رستگارپور یک روش زمانبندی و اختصاص وظایف به منابع با ترکیب الگوریتم کرم شب‌تاب چند هدفه و منطق فازی ارایه کرده‌اند. هدف این روش پیشنهادی، بهبود زمان گردش کار و هزینه ارتباطی در محیط رایانش ابر است[10]. در سال ۱۴۰۰ یک روش پیش‌دستانه با هدف تشخیص زودهنگام وضعیت میزبان‌ها توسط صدیقه باقری و همکارانش ارائه شده است که مقدار مصرف پردازنده هر میزبان در آینده، توسط روش ماشین یادگیری افراطی(ELM) پیش‌بینی می‌شود و با استفاده از سه آستانه تطبیقی وضعیت آتی میزبان‌ها مشخص می‌شود، سپس در صورت نیاز انتقال بار در بین ماشین‌های مجازی پربار و کم بار انجام می‌شودالبته احتمال پربار شدن آنها بعد از تخصیص کمینه در نظر گرفته می‌شود[11]. لایانیا و همکارانش در سال ۲۰۲۰ دو الگوریتم زمانبندی کار، یعنی الگوریتم آفلاین TBTS و الگوریتم زمانبندی آنلاین-SLA LB برای سیستم‌های ناهمگن پیشنهادکرده‌اند که در آن معیارهای عملکرد مانند طول زمانبندی، جریمه، هزینه افزایش و همچنین ضریب استفاده از VM الگوریتم پیشنهادی در مقایسه با الگوریتم‌های موجود حتی در وضعیت مقیاس‌پذیری مجموعه داده و ماشین‌های مجازی، بهترعمل می‌کنند [12]. تیان ژنگ و همکارانش در سال ۲۰۲۱ درمقاله‌ای الگوریتم SLAMIG را پیشنهاد کرده‌اند که شامل گروه‌بندی مهاجرت همزمان و زمان‌بندی مهاجرت آنلاین است. که مشخصه، پهنای باند موجود را برای بهبود مهاجرتهای سریال و موازی به حداکثر می‌رساند.

نتایج تجربی آنها نشان داد که SLAMIG می‌تواند به طور مؤثری تعداد مهلت‌های مهاجرت ازدست رفته را کاهش دهد و در عین حال به عملکرد مهاجرت خوب در زمان کل مهاجرت، میانگین زمان اجرا، زمان خرابی، داده‌های انتقال یافته با زمان اجرا الگوریتم قابل قبول دست یابد[13] . براساس مطالعه ساختار سیستم رایانش ابری و نحوه عملکرد آن، نمونه‌هایی ازالگوریتم‌های زمانبندی و تخصیص منبع در قالب یک جدول نمایش داده می‌شود (جدول ۱). در جدول ۱ یک مقایسه کلی از پژوهش‌های انجام شده در زمینه SLA بهمراه ویژگی‌های الگوریتم‌ها، پارامترهای ارزیابی و بسترهای شبیه‌سازی ارائه می‌شود.

جدول۱: مقایسه الگوریتم‌های مورد مطالعه مبتنی بر SLA

مرجع	ویژگی‌ها	پارامترهای ارزیابی	بسترشبیه‌سازی
وینسنت سی[5]	- تأمینمنابعقابلاعتمادبرایسرویسابری توسط متوازن کننده - بررسی لیست منابع در دسترس و لیست ماشین‌های مجازی موجود قبل از تخصیص	بهره‌وری منابع، زمان پایان	کلودسیم
یو و همکارانش [6]	در این استراتژی بررسی شده است که آیا درخواست جدید می‌تواند بر اساس نیازهای QoSو قابلیت‌های منابع پذیرفته شود یا نه	هزینه، زمان پاسخ، نرخ دسترسی	کلودسیم
یو و همکارانش[7]	- افزایش سود فراهم‌کننده SaaSبا کاهش هزینه VMها با استفاده از استراتژی‌های تخصیص منبع موثر، از لایه بستر است - کاهشمصرفانرژیمرکز داده،قابلیتاطمینانسرویس باتوازنتوان	هزینه، زمان شروع، نرخ جریمه	کلودسیم
جین سونگ تسای[8]	- تخصیصVM بر اساس محاسبه یک تابع سودمندی و بار کاری - تعیین میکند کدام VMها باید ایجاد، تخریب و یا تغییر سایز دهد.	زمان پردازش، هزینه	استفاده از داده های واقعی
سوراب کومار گارج[9]	- سودمندی یک روز هر VM از یک مجموعه داده را با حداقل خطای مربع میانه ریشه نشان می‌دهد.	بهره وری مرکزداده	شبیه‌سازی یک مرکزداده واقعی
بویا کارج کومار[10]	کاهش هزینه ارتباطی و کاهش زمان اتمام کار، برای نگاشت کارها به ماشین مجازی در محیط رایانش ابری میباشد.	هزینه ارتباطی، زمانگردش کار، زمان پاسخ ، انرژی مصرفی و قابلیت اطمینان	متلب
ندا نیل ساز دزفولی و مریم رستگارپور[10]	هدف روش پیشنهادی، بهبود زمان گردش کار و هزینه ارتباطی در محیط رایانش ابر است.	تعدادمتوسطSLA، زمان مهاجرت، تعدادزمانبهینهسازی، زمان عملیاتی، زمان سربار	کلودسیم
صدیقه باقری و همکارانش[11]	هدف تشخیص زودهنگام وضعیت میزبان‌ها است که مقدار مصرف پردازنده هر میزبان در آینده	بهره‌وریCPU ,تعداد مهاجرتVMها، انرژی مصرفی و زمان پاسخ	کلودسیم
لایانیا و همکارانش [12]	طراحی و تست یک سیستم زمانبند در چهار مرحله	طول زمانبندی، بهره وری ابرو هزینه	کامپایلرGCC
تیان ژنگ و همکارانش [13]	کاهش سربار مهاجرت	زمان مهاجرت، زمان پردازش، انتقال داده و انرژی مصرفی	CloudSimSDN-NFV

۳- رویکرد پیشنهادی

در این بخش ابتدا به ارائه چارچوبی از روش پیشنهادی مبتنی بر زمانبندی و تخصیص منابع با در نظر گرفتن اولویت SLA در رایانش ابری اشاره خواهد شد، سپس به تشریح رویکرد مورد نظر برای تخصیص منابع با شناخت درخواست‌ها و منابع در حال پردازش و جاری پرداخته می‌شود؛ به این صورت که با ایجاد یک مدل پیشنهادی در چند سطح مختلف و با تکنیک اولویت‌گذاری و وزن‌دهی و همچنین مطابق با توافق‌نامه سطح خدمات سرویس به ارزیابی و فرموله‌سازی هزینه و انرژی مصرفی در سرور پرداخته خواهد شد.

۳-۱- چارچوب و مدل روش پیشنهادی

مسئله تخصیص منابع و زمانبندی کارها در محیط رایانش ابری، چالش مهمی است که تأثیرمستقیمی بر کلیات سرویس ارائه شده دارد. تخصیص منابع عبارت است از تصمیم‌گیری درباره اینکه چگونه، چه مقدار، کجا و چه زمان، منابع در دسترس برای کاربر ایجاد می‌شود. به طور نمونه، کاربران درباره نوع و میزان منابع درخواستی تصمیم می‌گیرند، سپس فراهم‌کنندگان، منابع درخواستی را بر روی گره‌هایی در مراکز داده قرار می‌دهند [14]. برای اجرای برنامه‌های کاربردی به طور کارا، نوع منابع باید با ویژگی‌های حجم بار مطاقبت داشته باشد و مقدار آن نیز باید محدودیت‌ها را (مانند مهلت زمانی تکمیل کار) به طور شایسته ارضا نماید. در یک محیط الاستیک مانند ابر، که کاربران می‌توانند منابع را به طور پویا درخواست نمایند یا بازگردانند، در نظر گرفتن اینگونه تنظیمات دارای اهمیت زیادی است. پس از تخصیص منابع محاسباتی نیاز به اتخاذ تصمیماتی درباره زمانبندی کارها می‌باشد [15]. زمانبندی وظایف یعنی فرآیند نگاشت کارها به منابع در دسترس بر پایه نیازمندی‌ها و ویژگی کارها است. مسئله زمانبندی کار در رایانش ابری، یک مسئله بسیار مهم و از رده مسائل NP محسوب می‌شود که سعی دارد یک زمانبندی بهینه برای اجرای وظایف و تخصیص بهینه منابع مشخص نماید، به شکلی که در زمان کم، حجم بیشتری از کارها را بتوان پردازش کرد. زمانی‌که منابع به کاربر داده می‌شود، برنامه کاربردی یک تصمیم زمانبندی اتخاذ می‌نماید. در بسیاری موارد، برنامه کاربردی شامل کارهای چندگانه است که منابع تخصیص داده شده به آن‌ها داده می‌شود. برنامه کاربردی همچنین باید مطمئن شود که به هر کار مقدار منابع کافی و مناسب داده می‌شود، یا اشتراک آن‌ها منصفانه خواهد بود. در مسئله پیشرو، به طور کلی، m ماشین مجازی و n وظیفه موجود است. باید توجه داشت که این m ماشین در c سلول (سرور) مستقر شده است و از هر کدام، یک یا بیشتر موجود است. هر کار، دارای یک فرآیند خاص است و توالی انجام عملیات مختلف به منظور تکمیل آن کار مشخص است. همچنین هر یک از کارها دارای تعدادی کار فرعی یا وظیفه می‌باشد. انجام هر وظیفه از هر کار بر روی هر ماشین، دارای زمان پردازش خاص خود است. ظرفیت سخت‌افزاری و نرم‌افزاری سرورها و ماشین‌های مجازی مشخص شده است و هرماشین تنها یک عملیات یا وظیفه را در هر لحظه می‌تواند انجام دهد. فرض بر این است که در صورت شروع یاعملیات، قطع آن مجاز نیست. در این مسئله، عملیات مربوط به پردازش هر کار بر روی ماشین‌های مجازی مورد نیاز برای پردازش در سلولی که ماشین مستقر است، زمانبندی می‌گردد. قبل از تشریح رویکرد و چارچوب روش پیشنهادی ابتدا به صورت گرافیکی یک دید کلی از روش پیشنهادی مطابق با شکل (۱) ارائه شده است سپس به تشریح روش و گام‌های مختلف آن پرداخته شده است.

مطابق با معماری ارائه شده ۵ لایه مختلف برای پیشبرد هدف تعریف می‌شود، که هر لایه به صورت مجزا تشریح شده است.

- لایه اول درخواست سرویس کاربران

بر اساس جریان بسته شبکه، ماشین‌های مجازی از بزرگ به کوچک مرتب می‌شوند.

- لایه دوم: پلت فرم نظارت بر تخصیص منابع

۱) شناسایی تمام درخواست‌های سرویس و تعیین اینکه آیا سرویس آنلاین است.

۲) مرتب کردن بار سرور توسط وزن.

۳) مقدار بار مرتب شده را به یک پایگاه داده ذخیره می‌کند.

- لایه سوم : پلت فرم تخصیص منابع

۱)دریافت درخواست کاربر

۲) ابتداماشین X را به دست می‌آورد.

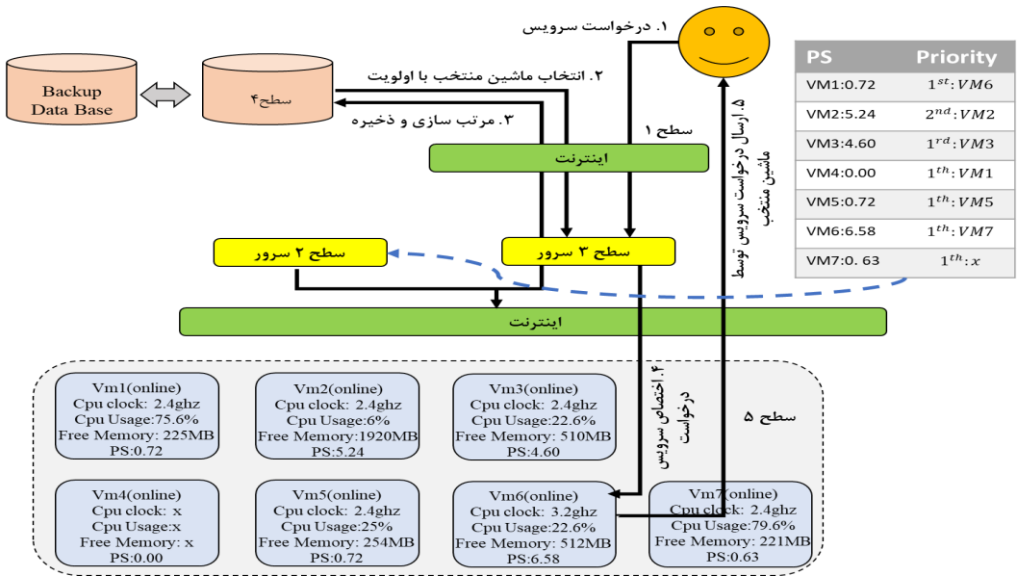
۳)درخواست کاربر را به ماشین X براساس روش تعادل بار پرس و جو تعیین می‌کند.

- لایه چهارم

بارگذاری اطلاعات اولویت‌بندی شده سرورها در هر۱دهم ثانیه که در پایگاه داده ذخیره شده است.

- لایه پنجم

تهیه پشتیبان برای ذخیره‌سازی سرویس‌های ابر در گروهی از سرویس‌هابه درخواست‌های کاربر پاسخ می‌دهد.



شکل ۱: چارچوب اولیه از روش پیشنهادی.

مطابق با معماری ارائه شده، درخواست سرویس مطابق با توافق‌نامه سطح سرویس در لایه سوم درخواست کاربر، دریافت می‌شود; سپس با توجه به درخواست انجام شده ماشین هدف با توجه بر اولویت و محاسبه وزن تخصیص انتخاب می‌شود و داده‌ای مبتنی بر تخصیص به لایه دوم ارسال می‌شود. در این لایه با آنالیز ماشین‌های آنلاین که بر اساس الگوریتم نوبت گردشی، وزن‌دهی شده‌اند مرتب‌سازی ماشین‌ها بر اساس اولویت انجام و در یک پایگاه داده ذخیره می‌شود. سپس در لایه پنجم ماشین منتخب به درخواست سرویس ارائه شده تخصیص داده می‌شود و پاسخ سرویس درخواست شده داده می‌شود. همچنین در لایه چهارم ذخیره‌سازی اطلاعات و تعاملات در زمان مشخص شده بروزرسانی و ذخیره می‌شود. اما نکته‌ای که باید به آن پرداخت نحوه محاسبه نرخ خدمات و وزن‌دهی به ماشین‌ها بر اساس فعالیت‌هایشان است. در حقیقت هر ماشینی در یک بازه زمانی مشخص شده یک وزن بر اساس میزان فعالیت خود دارد که این امر به صورت روابط (۱) و (۲) محاسبه می‌شود. μ_i و λ به ترتیب نشان‌دهنده نرخ خدمات و نرخ ورود است که برای محاسبه وزن زمانبندی (β) استفاده می‌شود. همچنین $\emptyset_i$ درصد وزن بار مشترک را محاسبه می‌کند. این مرحله به فاز تحلیل کمک می‌کند تا به طور موثر میزان وزن هر ماشین را ارزیابی کند. متغییرهای مورد استفاده به صورت جدول (۲) نشان داده شده است.

در شکل (۲) روند الگوریتم تخصیص منابع و زمانبندی بر اساس الگوریتم نوبت گردشی و اولویت‌داری با توجه به خدمات سطح سرویس در بستر مورد مطالعه نشان داده شده است.

جدول۲: متغییرهای مورد استفاده در فرمول (۱) و (۲).

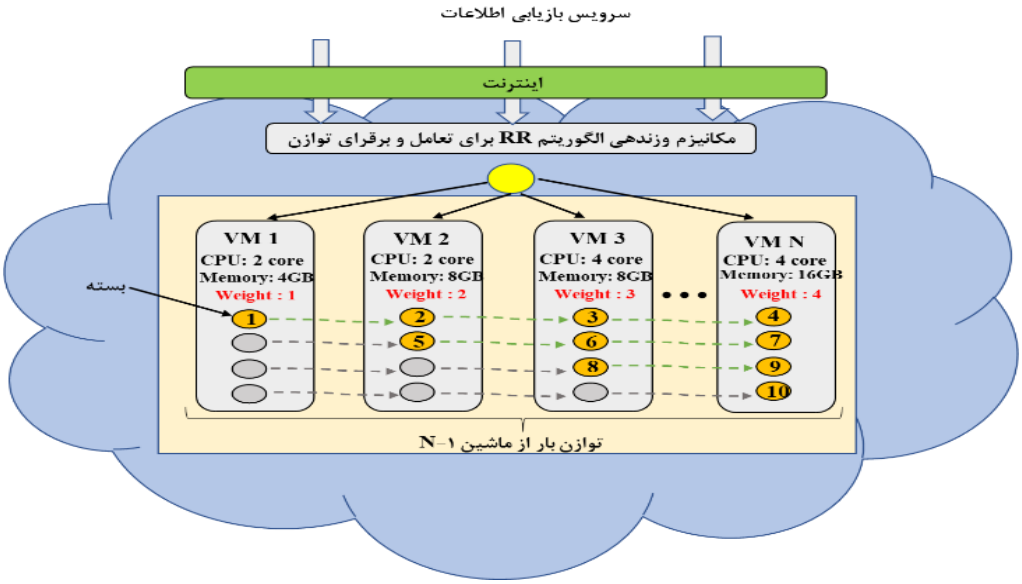
متغییر	توضیحات
$\emptyset_i$	درصد وزن نشان‌دهنده بار مشترک
λ	نرخ ورود
μ_i	نرخ خدمات
β	وزن زمانبندی

$$\Phi_i = \frac{\mu_i}{\lambda}\left(1-\sqrt{\frac{\lambda}{\mu_i \beta}}\right) \qquad (۱)$$

$$\beta = \frac{\frac{1}{\lambda}\left[\sum_{i=1}^{N}\sqrt{\mu_i}\right]^2}{\frac{1}{\lambda^2}\left[\sum_{i=1}^{N}\mu_i\right]^2 - \frac{2}{\lambda}\left[\sum_{i=1}^{N}\mu_i\right]+1} \qquad (۲)$$

۳-۲- مدلسازی ریاضی مسئله

در این بخش بر مدلسازی ریاضی مسئله تخصیص منابع و زمانبندی کارها در محیط رایانش ابری تمرکز شده است. اهداف این مسئله، کمینه‌سازی هزینه انجام کارها و همچنین بهینه‌سازی مصرف انرژی می‌باشد.



شکل۲: روند الگوریتم تخصیص منابع و زمانبندی بر اساس الگوریتم نوبت گردشی.

نمادهای بکارگرفته شده و متغیرهای تصمیم جهت مدلسازی مسئله به شرح زیر می‌باشند.

- اندیس وظیفه: $i = 1,2, \ldots, s$
- اندیس کار: $j = 1,2, \ldots, t$
- اندیس نوع ماشین: $p = 1,2, \ldots, m$
- اندیس شماره ماشین: $r = 1,2, \ldots, N_{pk}$
- اندیس سرور: $k = 1,2, \ldots, c$

۳-۲-۱- پارامترهای مدل

- t_{ijp} : مدت زمان انجام وظیفه i از کار j بر روی ماشین نوع p
- DTI_k: هزینه انتقال ورود داده به سرورk (هزینه در مگابایت$/MBs)
- DTI_O :هزینه انتقال خروج داده از سرورk (هزینه درمگابایت$/MBs)
- Com_k :هزینه محاسبات در سرورk (هزینه در میلیون دستور اجرا شده$/S/Mis)
- Bg_i:بودجه برای انجام کارj
- cpu_{ij} :میزان سی پی یو مورد نیاز برای انجام وظیفه i از کار j
- mem_{ij} :میزان حافظه مورد نیاز برای انجام وظیفه iاز کار j
- $E_{K,max}$: بیشینه مصرف انرژی سرورk
- $E_{K,\ idle}$: مصرف انرژی ثابت در سرور k
- E_K :مصرف انرژی پویا در سرور k
- ES_K :انرژی شروع/اخیر سرور k
- gs_{ij} : زمان تکمیل وظیفه i از کار j
- gt_j: زمان تکمیل کار j
- Cst_{ijk} : هزینه انجام وظیفه i از کار j در سرور k

۳-۲-۲- متغییرهای تصمیم

$$Y_{ijk} = \begin{cases} 1 & \text{اگر وظیفه } i \text{ از ، کار } j \text{ ، در سرور } k \text{ انجام شود} \\ 0 & otherwise \end{cases} \quad (۳)$$

$$M_{rpk} = \begin{cases} 1 & \text{ماشین } r \text{ ام نوع } p \text{ در سرور } k \text{ تخصیص داده شده} \\ 0 & otherwise \end{cases} \quad (۴)$$

$$Z_{ijrpk} = \begin{cases} 1 & \text{وظیفه } i \text{ کار از } j \text{ به ماشین شماره } r \text{، نوع p در سرور } k \\ 0 & otherwise \end{cases} \quad (۵)$$

$$Act_k = \begin{cases} 1 & \text{اگر سرور } k \text{ استفاده شود} \\ 0 & otherwise \end{cases} \quad (۶)$$

۳-۲-۳- مدل ریاضی

در اینجا تابع هدف اول (رابطه‌ی (۷)) مربوط به محاسبه هزینه انجام تمام وظایف از تمام کارها در تمام سرورها است. باید توجه داشته باشیم که هزینه محاسبات و ذخیره‌سازی با توجه به ظرفیت هر سرور و همچنین ترافیک شبکه و غیره در سرورهای مختلف، متفاوت می‌باشد. تابع هدف دوم (رابطه‌ی (۸))، انرژی مصرفی را درسرورهایی که فعال می‌باشند محاسبه می‌کند. در نظر گرفتن

این نکته ضروری است که هزینه قدرت مصرفی منابع ارتباطی، خنک‌کننده‌ها و تهویه مطبوع ماژول‌ها (واحدها) در سراسر سرورها و تجهیزات ارتباطی/شبکه در مرکز داده مستهلک هستند و در نتیجه نسبتا مستقل از حجم کار کاربر فرض می‌شوند. به طور دقیق‌تر، این هزینه‌ها درمعادله هزینه و قدرت مرکز داده گنجانده نمی‌شوند.

$$F_1 = \sum_{r=1}^{N_{pk}} \sum_{p=1}^{m} \sum_{k=1}^{c} M_{rpk} \qquad (۷)$$

$$F_2 = \sum_{i=1}^{s} \sum_{p=1}^{t} \sum_{k=1}^{c} Cst_{ijk} \cdot Y_{ijk} \qquad (۸)$$

با توجه به توابع هدف تعریف شده چند محدودیت نیز تعریف کرده‌ایم که به صورت زیر می‌باشد.

$$\sum_{r=1}^{N_{pk}} \sum_{p=1}^{m} \sum_{k=1}^{c} Z_{ijrpk} = 1 \; \forall \, i,j \qquad (۹)$$

محدودیت (۹) تضمین می‌کند که هر وظیفه از هر کار تنها به یک ماشین از یک نوع و در یک سرور تخصیص داده شود.

$$gs_{ij} \geq t_{ijp} \; \forall \, i,j,p \qquad (۱۰)$$

$$gt_j \geq gs_{ij} \; \forall \, i,j \qquad (۱۱)$$

محدودیت (۱۰) تضمین می‌کند که زمان تکمیل هر وظیفه حداقل برابر با زمان انجام آن وظیفه می‌باشد. همچنین محدودیت (۱۱)زمان تکمیل هر وظیفه را نشان می‌دهد. یعنی زمان تکمیل کار j باید حداقل برابر بازمان تکمیل وظیفه آن باشد.

$$Cst_{ijk} = cpu_{ij} \cdot Com_k + \quad \forall \, ij,k \qquad (۱۲)$$
$$+ mem_{ij} \cdot (DTI_k + DTO_k)$$

محدودیت(۱۲) یک فراهم کننده سرویس ابر عمومی، هزینه سه مورد را دریافت می‌کند: محاسبات و انتقال داده. انتقال داده از/به اینترنت، آدرس‌های IP خصوصی یا عمومی و غیره صورت می‌پذیرد و با توجه به ناچیز بودن هزینه انتقال داده‌ها بین سرورهای ابر، در مدلسازی، معادله هزینه می‌تواند نادیده گرفته شود و هزینه آن را صفر تنظیم نمود.

$$\sum_{k=1}^{c} \sum_{i=1}^{s} Cst_{ijk} \cdot Y_{ijk} \leq Bg_j \; \forall \, j \qquad (۱۳)$$

محدودیت (۱۳) برای کنترل بودجه تعریف می‌شود و تضمین می‌کند که هزینه انجام تمام وظایف از تمام کارها ازبودجه تعیین شده تجاوز نمی‌کند.

$$\sum_{i=1}^{s} \sum_{j=1}^{t} Y_{ijk} \cdot E_K \leq E_{k,\max} \cdot Act_k - E_{k,idle} - ES_k \quad \forall \, k \qquad (۱۴)$$

در محدودیت (۱۴) هر سرور یک حد انرژی $E_{k,max}$ دارد که نمی‌تواند از آن تجاوز کند و خدمت یا میزبانی ماشین‌های مجازی باتوجه به ظرفیت باقی مانده صورت می‌پذیرد. یک انرژی ثابت یا idle به علاوه انرژی شروع/اخیر سرور k، مصرف انرژی اخیر آن سرور را مشخص می‌نماید. در شکل ۳ یک بلوک دیاگرام کلی از مدل روش پیشنهادی را ارائه می‌کند که در آن به ترتیب مراحل اجرا ارائه می‌شود.

۴- نتایج شبیه‌سازی

در این بخش برای نشان دادن اجرایی بودن روش پیشنهادی یک سیستم کوچک مورد بررسی قرار داده شده است. برای شبیه‌سازی از کلودسیم استفاده شده است، در حقیقت مدل ارائه شده به صورت پویا است و منعطف با نیازمندی‌های مختلف می‌باشد. این امر باعث می‌شود روش ارائه شده از نقطه نظرهای مختلف مورد بررسی قرار بگیرد. در اینجا ناهمگنی سخت‌افزاری و نمونه‌های VM مورد توجه قرار می‌گیرد. ماشین‌های مجازی مورد استفاده برای اجرای درخواست سرویس براساس انواع نمونه‌های M3 آمازون مدلسازی شده‌اند. خصوصیات ماشین مجازی به شرح جدول (3) می‌باشد:

جدول۳:مشخصات نمونه‌های ماشین مجازی

	پردازنده	حافظه اصلی (GB)	فضای ذخیره‌سازی(GB)
m3.medium	۱	۳,۷۵	۱×۴
m3.large	۲	۷,۵	۱×۳۲
m3.xlarge	٤	۱۵	۲×۴۰

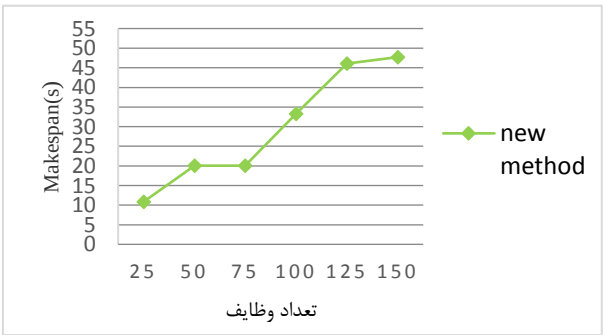
مقادیر جدول (۴)برای پارامترهای انتخاب شده و تنظیمات اولیه برای شبیه‌سازی است. آزمایشات انجام شده بر اساس تعداد درخواست‌های مختلف در یک بازه زمانی مشخص و یکسان انجام شده است.

شکل (۳):بستر مدل پیشنهادی

جدول۴ پارامترها و تنظیمات اولیه شبیه‌سازی روش پیشنهادی.

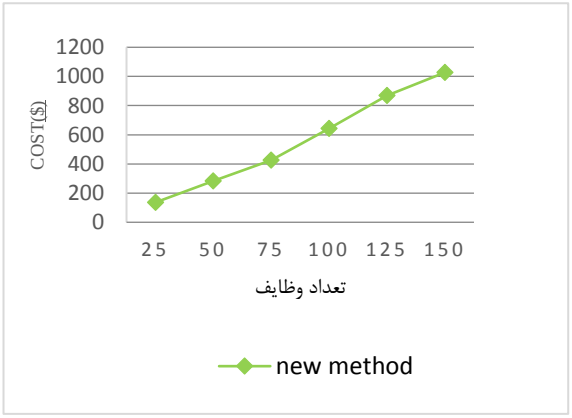
پارمتر/مقدار	توضیحات
LCG-2005-1	اندازه وظایف
150_۲۵	تعداد کل درخواست‌ها
مکان اشتراکی و زمان اشتراکی	سیاست زمانبندی تخصیص
۱	تعداد مرکزداده
۱۵	تعداد هر ماشین مجازی(Vm)

شکل (۴) یک مقایسه کلی از نتایج حاصل از طول زمانبندی نسبت به تعداد وظایف مختلف را ارائه می‌دهد.در این نمودار محور x، y به ترتیب بیانگر تعداد وظایف و طول زمانبندی می‌باشد. با توجه به شکل(۴) می‌توان به این نتیجه رسید که تعداد وظایف بر طول زمانبندی تاثیر گذار می‌باشد. برای اجرای وظایف با سایز بزرگتر، زمان اجرای بیشتری نیاز دارند. در این نمودار، مینیمم و ماکزیمم طول زمانبندی بترتیب به میزان 10.85 و 47.72 ثانیه می‌باشد.



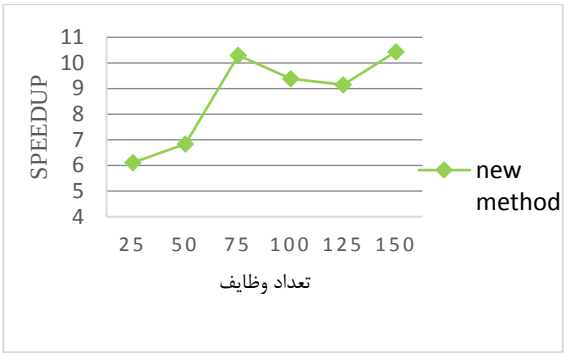
شکل۴:نمودار مقایسه‌ای طول زمانبندی در تعداد وظایف مختلف.

هزینه پردازش، ازمسائل ارزیابی الگوریتم‌ها می‌باشد که در شکل (۵) عملکرد الگوریتم ارائه شده نسبت به تعداد وظایف متغیر ارائه شده است. در این نمودار محور x، y به ترتیب بیانگر تعداد وظایف و هزینه پردازش می‌باشد. برای اجرای وظایف با سایز بزرگتر، زمان اجرای بیشتری نیاز دارند. در این نمودار، مینیمم و ماکزیمم هزینه پردازش بترتیب به میزان ۱۳۶.۱۲ و ۱۰۲۷.۷۸ دلار می‌باشد.



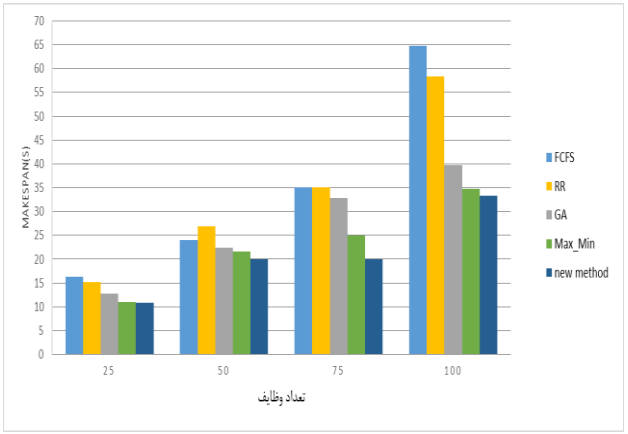
شکل۵:نمودار مقایسه‌ای روش ارائه شده از دید هزینه.

شکل(۶) یک مقایسه کلی از نتایج حاصل از سرعت پردازش در وظایف مختلف را ارائه می‌دهد. سرعت پردازش،عملکرد الگوریتم ارائه شده در شکل(۶) ارائه شده است. شکل(۶) بیانگر این مورد است که با افزایش شمار وظایف ما شاهد افزایش در سرعت پردازش هستیم، اما به طور کلی میتوان گفت نرخ تغییرات سرعت نسبت به شمار بزرگ وظایف نسبتا ثابت خواهد بود. در این نمودار، مینیمم و ماکزیمم سرعت پردازش بترتیب به میزان ۱۰.۴۴و ۶.۱۱ می‌باشد.



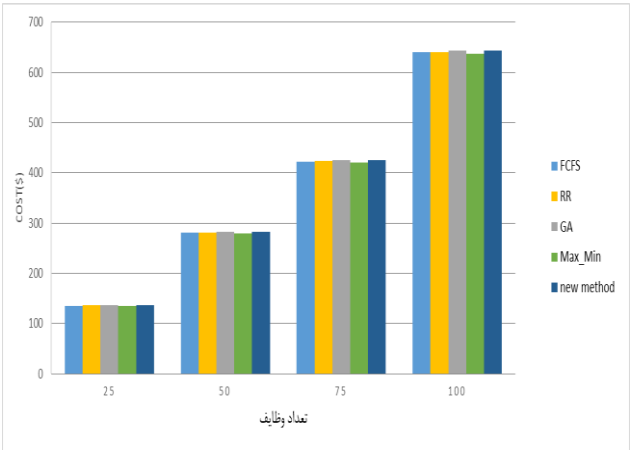
شکل۶:نمودار مقایسه‌ای روش ارائه شده از دید سرعت.

در ادامه این بخش نتایج بدست آمده از رویکرد روش پیشنهادی (PM[14]) را با چهار الگوریتم (GA[15]،Max_min[16]،RR[17] و FCFS) [16, 17, 18] ارائه شده است.



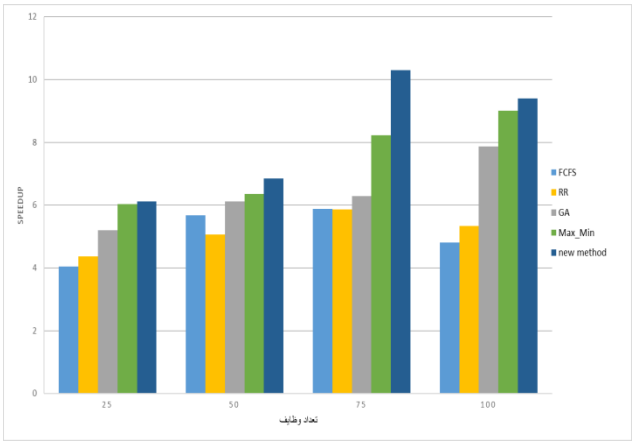
شکل ۷: نتایج حاصل از مقایسه روش پیشنهادی با دیگر روش‌های مشابه از دید طول زمانبندی.

در شکل(۷) محور x و y به ترتیب بیانگر تعداد وظایف و طول زمانبندی وظایف می‌باشد. با توجه به شکل میتوان به این نتیجه رسید که تعداد وظایف بر طول زمانبندی تاثیر گذار می‌باشد. همچنین الگوریتم PM عملکرد قابل قبولی نسبت به الگوریتم‌های RR, FCFS GA, و Max_min ارائه دارد.



شکل۸:نتایج حاصل از مقایسه روش پیشنهادی با دیگر روش‌های مشابه از دید سرعت هزینه.

در شکل (۸) محور x و y به ترتیب بیانگر تعداد وظایف و طول زمانبندی وظایف می‌باشد. با توجه به شکل میتوانبه این نتیجه رسید که تعداد وظایف بر هزینه پردازش تاثیر گذار می‌باشد.
همچنین الگوریتم PMعملکرد قابل قبولی نسبت به الگوریتم‌های RR, FCFS GA, و Max_min ارائه دارد.



شکل۹:نتایج حاصل از مقایسه روش پیشنهادی با دیگر روش‌های مشابه از دید سرعت پردازش.

در شکل (۹)میزان سرعت پردازش الگوریتم PMنسبت به لگوریتم‌های RR, FCFS, GA, و Max_min مورد بررسی قرار گرفته شده است. شکل (۹)نشان می‌دهد که روش پیشنهادی ارائه شده عمل توزیع وظایف و اجرا را سریع‌تر از الگوریتم‌های FCFS RR, GA, و Max_min انجام می‌دهد. دلیل این امر کاهش مؤثر میزان طول زمانبندی در روش پیشنهادی می‌باشد; بنابراین بالا بودن سرعت در روش پیشنهادی دور از ذهن نخواهد بود. همان طور که شکل (۹) نشان داده شده است؛ روش پیشنهادی بیشترین سرعت (۱۰.۳) و روش FCFSکمترین سرعت (۴) را دارا می باشد.

۵- نتیجه‌گیری

مسئله زمانبندی و تخصیص منابع یک چالش بزرگ در محیط‌های ابری می‌باشد و رابطه مستقیمی با زمان پردازش و هزینه کاربران دارد؛ از این رو سوی سیستم تخصیص منبع باید با درخواست‌های پیش‌بینی نشده سروکار داشته باشد. همچنین پس از تخصیص منابع محاسباتی نیاز به اتخاذ تصمیماتی درباره زمانبندی وظایف دارد. مسئله زمانبندی وظایف در رایانش ابری، مسئله‌ای بسیار مهم محسوب می‌شود که سعی دارد یک زمانبندی بهینه برای اجرای وظایف مشخص نماید. در واقع هدف این سیستم‌ها، مشخص کردن یک منبع پردازشی از مجموعه منابعی که یک کار برای پردازش به آن نیاز دارد می‌باشد، به شکلی که در زمان کمتر و هزینه مناسب وظایف بیشتری را بتوان پردازش کرد. سیستم زمانبندی، وظایف مختلفی را در سیستم ابر جهت افزایش نرخ

تکمیل کارو افزایش سرعت از منابع و در نتیجه افزایش توان محاسباتی، کنترل می‌کند. در این مقاله با توجه به چالش‌های ذکر شده یک مدل چند سطحی از تخصیص و زمانبندی وظایف در رایانش ابری ارائه شده است، که به کمک آن علاوه بر در نظر گرفتن اولویت کارها و SLA، هزینه، زمان اتمام وظایف و سرعت پردازش بهینه‌سازی می‌کند. در حقیقت به کمک رویکرد ارائه شده وضعیت‌های جاری ماشین‌های مجازی بررسی شده و بهترین تخصیص ممکن در شرایط مختلف صورت خواهد گرفت. برای ارزیابی الگوریتم پیشنهاد شده از نمونه‌های مختلف ماشین مجازی و مراکز داده در شبیه‌ساز کلودسیم استفاده شده است که ابزار مناسبی برای شبیه‌سازی محیط ابر است. اگرچه ارزیابی طول زمانبندی نسبت به چهارالگوریتم FCFS، RR،Max_min و GA کاهش قابل توجهی را نشان می‌دهد، یک بهبود نسبی در هزینه پردازش اجرای وظایف ارائه شده است. همچنین سرعت پردازش بهبود قابل قبولی نسبت به الگوریتم‌های FCFS، RR، Max_min و GA ارائه شده است. در کارهای آینده، ما قصد داریم استحکام الگوریتم را افزایش دهیم و پیچیدگی زمانی الگوریتم ارائه شده رو مورد بررسی قرار گیرد. همچنین با در نظر گرفتن اهداف بهره‌وری انرژی در تخصیص و استفاده از منابع بررسی خواهیم شد.

مراجع

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, "Above the Clouds: A BerkeleyView of Cloud Computing February,"*Electrical Engineering and Computer Sciences University of California at Berkeley*,2009 .

[2] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions,"*Future Generation Computer Systems*,vol. 29 ,no. 7 ,pp. 1645-1660, 2013.

[3] D. Liu, X. Sui, L. Li, Z. Jiang, H. Wang, Z. Zhang, Y. Zeng "A cloud service adaptive framework based on reliable resource allocation,"*Future Generation Computer Systems*,vol. 89 ,pp. 455-463, 2018.

[4] A. Hadjali, H. Mezni, S. Aridhi, A. Tchernykh, "Uncertainty in Cloud Computing: Concepts, Challenges and Current Solutions,"*International Journal of Approximate Reasoning*,vol. 111 ,pp. 53-55, 2019.

[5] V. C. Emeakaroha, I. Brandic, M. Maurer, I. Breskovic, "SLA-aware application deployment and resource allocation in clouds,"*IEEE 35th Annual Computer Software and Applications Conference Workshops*,pp. 298-303, 2011 .

[6] L. Wu, S. K. Garg, R. Buyya, "SLA-based admission control for a Software-as-a-Service provider in Cloud computing environments,"*Journal of Computer and System Sciences*,vol. 78 , no. 5 ,pp. 1280-1299, 2012.

[7] L. Wu, S. K. Garg, R. Buyya, "SLA-Based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments," in *IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID)* ,Newport Beach, CA, USA, 2011.

[8] J. T.Tsai, J. C. Fang, J. H. Chou, "Optimized task scheduling and resource allocation on cloud computing environment using improved differential evolution,"*Computers&Operations Research*,pp. 3045-3055, 2013.

[9] B. Rajkumar, S. K. Garg, S. K. Gopalaiyengar, "SLA-Based Resource Provisioning for Heterogeneous Workloads in a Virtualized Cloud Datacenter,"*International Conference on Algorithms and Architectures for Parallel Processing(ICA3PP)*,pp. 371-384, 2011.

[10] Nilsaz, N., & Rastgarpour, M. (2021). A New Load-Balancing Algorithm Using Fuzzy Logic and Multi-Objective Firefly Algorithm in Cloud Computing Environment. Journal of Soft Computing and Information Technology, 10(2), 14-26.

[11] Bagheri, S., Mostafavi, S., & Adibnia, F. (2021). An ELM-based Load Balancing Algorithm for Cloud Computing Platforms. Journal of Soft Computing and Information Technology, 10(2), 39-52.

[12] M. Lavanya, B. Shanthi, S. Saravanan, "Multi objective task scheduling algorithm based on SLA and processing time suitable for cloud environment,"*Computer Communications*,vol. 151 ,pp. 183-195, 2020.

[13] T .Z .He, A .N. Toosi,B. Rajkumar, "SLA-aware multiple migration planning and scheduling in SDN-NFV-enabled clouds,"*The Journal of Systems&Software*,vol. 176 ,2021 .

[14] C. Morariu, O. Morariu, S. Raileanu, T. Borangiu, "Machine learning for predictive scheduling and resource allocation in large scale manufacturing systems,"*Computers in Industry*,vol. 120 ,2020 . .

[15] M. H. Malekloo, N. Kara, M. El, "An energy efficient and SLA compliant approach for resource allocation and consolidation in cloud computing environments," no. 17 ,pp. 9-24, 2018.

[16] J. Gu, J. Hu, T. Zhao, G. Sun, "A New Resource Scheduling Strategy Based on Genetic Algorithm in Cloud Computing Environment,"*JOURNAL OF COMPUTERS*,vol. 7 ,pp. 42-52, 2012.

[17] N. Sharma, Dr. S. Tyagi, "A Comparative Analysis of Min-Min and Max-Min Algorithms based on the Makespan Parameter,"*International Journal of Advanced Research in Computer Science*,vol. 8 ,2017 .

[18] K. Mahajan, A. Makroo and D. Dahiya, "Round Robin with Server Affinity: A VM Load Balancing Algorithm for Cloud Based Infrastructure,"*Journal of Information Processing Systems*,vol. 3 , 2013 ..

پاورقی‌ها:

[1] Cloud computing

[2] Infrastructure as a Service (IaaS)

[3] Platform as a Service (PaaS)

[4] Software as a Service (SaaS)

[5] Virtualize

[6] Quality of Service (Qos)

[7] Service Level Agreement (SLA)

[8] Market oriented

[9] Virtual machine

[10] Multi Tenancy

[11] Artificial Neural Network (ANN)

[12] Back Propagation (BP)

[13] Root Mean Square Error (RMSE)

[14] proposed method

[15] Genetic algorithm

[16] Max_min

[17] Round Robin