بررسی الگوریتم‌های توازن بار در محیط رایانش ابری

حسین خیرآبادی، دکتر کیارش میزانیان باغ گلستان

دانشجو، دانشگاه یزد، دانشکده کامپیوتر و برق

عضو هیئت علمی دانشگاه یزد

Hossein.khirabadi@gmail.com

k.mizanian@yazd.ac.ir

خلاصه

امروزه فناوری رایانش ابری به سرعت در حال فراگیر شدن در صنعت و بخش‌های مختلف آموزشی است، فناوری رایانش ابری یک تعریف جدید از مخزنی از منابع مجازی‌سازی شده است. یکی از نگرانی‌های اصلی که رایانش ابری تعادل بار است. با استفاده از الگوریتم‌های توازن‌بار می‌توان تعادل بار را بین سیستم‌های رایانش ابری را بدست آورد که باعث افزایش بهره‌وری از منابع، افزایش رضایت مشتری، کاهش زمان پاسخ، کاهش هزینه و ... می‌شود.

کلمات کلیدی: رایانش ابری، ماشین مجازی ، توازن‌بار، مهاجرت

۱. مقدمه

رایانش ابری یک روش جدید برای ارائه منابع محاسباتی از طریق اینترنت است. موسسه مالی و فناوری و استانداردها[۱]رایانش ابری را این گونه تعریف می‌کند: رایانش ابری مدلی است برای فراهم کردن دسترسی آسان براساس تقاضای کاربر از طریق شبکه به مجموعه‌ی از منابع رایانشی قابل تغییر و پیکربندی (مثل: شبکه‌ها، سرورها، فضای ذخیره‌سازی، برنامه‌های کاربردی و سرویس‌ها) که این دسترسی بتواند با کمترین نیاز به مدیریت منابع و یا نیاز به دخالت مستقیم فراهم‌کننده سرویس به سرعت فراهم شده یا آزاد (رها) گردد. توازن بار در رایانش ابری را می‌توان یکی از بخش‌های مهم این فناوری دانست که منجربه بهربرداری موثر از منابع ، افزایش رضایت مشتری، کاهش زمان پاسخ و ... می‌شود. الگوریتم‌های متعددی برای حل مسئله توازن بار پیشنهاد شده‌اند، که هرکدام دارای مزایا و معایبی هستند و بسته به وضعیت سیستم می‌توانند بار را به صورت متوازن توزیع کنند.

در بخش دوم چالش‌های موود در توازن بار محیط رایانش ابری، در بخش سوم مروری بر انوا ع الگوریتم‌های توازن بار رایانش ابری و سپس در بخش چهارم به نتیجه گیری می‌پردازیم.

۲. چالش‌های توازن‌بار در رایانش ابری

قبل از این که مروری بر انواع الگوریتم‌های توازن بار در محیط رایانش ابری داشته باشیم، نیاز داریم انواع چالش‌های موجود توازن بار در محیط رایانش ابری را شناسایی کنیم. در این بخش چالش‌های که ممکن است در هنگام پیشنهاد یک الگوریتم وجود داشته باشد به طور خلاصه بیان می کنیم.

۱.۲ توزیع فضایی گره‌های ابر

برخی از الگوریتم‌های توازن بار فقط زمانی مؤثر عمل می کنند که گره‌ها نزدیک هم باشند و تأخیر بین آن‌ها ناچیز باشد. زمانی که این الگوریتم‌ها در محیطی که گره‌های آن به‌صورت فضایی توزیع شدن اجرا می‌شوند با چالش‌های روبه‌رو می‌شود که دلیل به وجود آمدن این چالش این است در این الگوریتم‌ها برگ خریدهای نظیر سرعت خطوط شبکه، فاصله بین کلاینت‌ها و گره‌های پردازش کننده وظایف در نظر گرفته نشده است. برای حل این چالش باید الگوریتم به گونه‌ی طراحی شود که تعادل با رابین تمام گره‌های توزیع‌شده به دست آورد و قادر به تحمل تأخیر بالا باشد[۲] .

۲.۲ تکرار

یک الگوریتم ذخیره‌سازی تکرار کامل یک ذخیره‌سازی کارا به‌حساب نمی‌آید، زیرا داده‌ها در تمام گره‌ها تکرار می‌شود. الگوریتم‌های تکرار کامل هزینه‌ی زیادی را تحمیل می کند، به این دلیل که فضای زیادی برای ذخیره‌سازی نیازمند است. الگوریتم ذخیره‌سازی بخشی که نوع دیگر از الگوریتم ذخیره‌سازی است که داده‌ها بر اساس قدرت و ظرفیت در گره‌های مختلف ذخیره می کند[۳]؛ که باعث کارایی بهتر این الگوریتم نسبت به الگوریتم قبلی می‌شود؛ اما الگوریتم بخشی باعث افزایش پیچیدگی در الگوریتم‌های توازن بار می‌شود به این دلیل که آن مجموعه قطعات داده‌ها در گره‌های مختلف در دسترس است.

۳.۲ پیچیدگی الگوریتم

هرچه الگوریتم‌های توازن بار ازنظر پیاده‌سازی و عملیات دارای پیچیدگی کمتر باشند، بیشتر مورد استقبال قرار می گیرند. افزایش پیچیدگی منجر به افزایش پیچیدگی پردازش می‌شود که می‌تواند باعث کاهش کارایی شود. علاوه بر این‌زمانی که الگوریتم‌ها نیاز به اطلاعات و ارتباطات بیشتر برای نظارت و کنترل داشته باشند، تأخیر مشکلات بیشتری ایجاد می کند و بهره‌وری کاهش می‌یابد. ازاین‌رو الگوریتم‌های توازن بار باید به ساده‌ترین شکل ممکن طراحی شوند.

۴.۲ نقطه شکست

کنترل توازن بار و جمع‌آوری داده در مورد نودهای مختلف باید به‌گونه‌ای طراحی شود از یک نقطه شکست در الگوریتم جلوگیری کند. برخی از الگوریتم‌های (الگوریتم‌های متمرکز) می‌توانند مکانیزگی مؤثر و کارآمد برای حل تعادل بار دریک الگوی خاص فراهم کنند. این نوع الگوریتم‌ها یک کنترل‌کننده برای کل سیستم دارند که اگر کنترل‌کننده دچار مشکل شود، کل سیستم از کار می‌افتد.

هر الگوریتم توازن باری که طراحی می‌شود باید بر چالش‌های که گفته شد فائق آید[۴].

۳. انواع الگوریتم‌های توازن بار

الگوریتم‌های توازن بار به دودسته کلی ایستا و پویا تقسیم می‌شوند. الگوریتم‌ها ایستا وظایف را به گره‌ها قابل‌دسترس تخصیص می‌دهند. فرایند انتخاب گره برای ارسال وظیفه بر اساس اطلاعات از قبل در مورد ویژگی و قابلیت گره دارد انجام می‌شود.الگوریتم‌های ایستا تغییراتی پویایی که در زمان اجرا در ویژگی‌های گره به وجود می‌آید را در نظر نمی‌گیرد، بعلاوه نمی‌تواند خود را با تغییرات محیط تطابق دهد[۵].هدف اصلی این الگوریتم‌ها کاهش زمان اجرای کلی است.

۱.۳ الگوریتم زمان‌بندی حداقل حداقل

این الگوریتم با مجموعه‌ای از وظایف شروع می‌شود. این الگوریتم منبعی که کمترین زمان تکمیل را برای تمام وظایف را دارد پیدا می کند، سپس وظیفه با کمترین اندازه به آن منبع تخصیص می‌دهد، (ازاین‌رو به این الگوریتم حداقل حداقل گفته

می‌شود.) و این وظیفه تخصیص داده‌شده را از مجموعه وظایف حذف کند. این روند تا هنگامی‌که تمام وظایف توسط الگوریتم تخصیص داده شود ادامه می‌یابد. این الگوریتم ساده است و بار موجود در منابع را قبل از تخصیص وظیفه در نظر نمی‌گیرد، بنابراین توازن بار مناسبی به دست نمی‌آورد[۶].

۲.۳ الگوریتم توازن بار توزیع یکسان اجرای کنونی

در این روش الگوریتم تلاش می‌کند تا بار تمام ماشین‌های مجازی که در یک مرکز داده به یکدیگر متصل هستند را به‌صورت یکسان حفظ کند. الگوریتم از یک جدول ایندکس از ماشین‌های مجازی و تعداد درخواست‌های که به هریک از آن‌ها تخصیص داده‌شده نگهداری می‌کند[۷].اگر یک درخواست جدید برای تخصیص به یک ماشین مجازی به مرکز داده ارسال شود، الگوریتم در جدول ایندکس را برای پیدا کردن کم‌بارترین ماشین مجازی جستجو می‌کند، اگر بیش از یک ماشین مجازی پیدا شود اولین را انتخاب می‌کند و سپس شناسه آن را به کنترل‌کننده مرکز داده ارسال می‌کند. مرکز داده درخواست را به ماشین مجازی انتخاب‌شده تخصیص می‌دهد و تعداد درخواست‌های تخصیص داده‌شده به ماشین مجازی در جدول ایندکس را یکی افزایش می‌دهد. زمانی که ماشین‌مجازی وظیفه را به پایان می‌رساند، توازن بار آن را با یک درخواست به مرکز داده برای اطلاع‌رسانی ارسال می‌کند و سپس در جدول ایندکس تعداد درخواست‌های تخصیص داده‌شده به ماشین مجازی را یکی کاهش می‌دهد. این محاسبات اضافی یک سربار است که جدول را دوباره و دوباره جستجو می‌شود. این الگوریتم سعی دارد زمان پاسخ، زمان پردازش را بهبود دهد و تحمل خطا ندارد و مشکل یک نقطه شکست را دارد[۸].

۳.۳ الگوریتم توازن بار نظارت فعال

این الگوریتم اطلاعاتی در مورد هر ماشین مجازی و تعداد درخواست‌های فعلی که به هر ماشین‌مجازی تخصیص داده‌شده را نگهداری می‌کند. زمانی که یک درخواست جدید وارد می‌شود ماشین‌مجازی که کمترین بار را دارد شناسایی‌شده و انتخاب می‌شود. اگر بیش از یک ماشین‌مجازی شناسایی شود اولین ماشین انتخاب می‌شود. شماره شناسایی ماشین مجازی انتخاب‌شده به کنترل‌کننده داده مرکزی برگردانده می‌شود و سپس کنترل‌کننده داده مرکزی درخواست به ماشین مجازی که توسط شناسه ارسال‌شده شناسایی کرده ارسال می‌کند؛ و سپس آن را به الگوریتم توازن بار نظارت فعال اطلاع‌رسانی می‌کند[۹]. در این الگوریتم فقط بار ماشین مجازی را در نظر می‌گیرد و به قابلیت‌های ماشین مجازی مانند قدرت پردازش و ظرفیت حافظه اهمیت نمی‌دهد، که ممکن است منجر به افزایش زمان انتظار برای برخی از درخواست‌ها شود و کیفیت سرویس کاهش نیز می‌یابد

۴.۳ الگوریتم توازن بار فرصت‌طلب

این الگوریتم توازن باریک الگوریتم ایستا است و هدفش مشغول نگه‌داشتن هر گره در ابر بدون در نظر گرفتن بار هریک از گره‌هاست. الگوریتم تلاش می‌کند تا کار انتخاب‌شده را به یک ماشین مجازی که به‌صورت تصادفی انتخاب‌شده اعزام کند[۱۰].الگوریتم توازن بار فرصت‌طلب زمان اجرای وظیفه در ماشین مجازی را در نظر نمی‌گیرد که ممکن است موجب شود وظیفه به‌کندی پردازش شود؛ و زمان اتمام کل افزایش یابد.

۴. نتیجه‌گیری

توازن بار یکی از مهمترین مسائل در محیط رایانش ابری محسوب می‌شود. وظیفه اصلی توازن بار استفاده از تمام منابع به گونه‌ی است هیچ یک از منابع پربار در حالیکه برخی دیگر از منابع کم‌بار نباشد. هدف اصلی توازن بار کاهش زمان پاسخ، افزایش رضایت مشتری، بهره‌برداری حداکثری از منابع و ... است. الگوریتم‌های متعددی جهت توازن بار پیشنهاد شده است که هرکدام بسته به وضعیت محیط رایانش ابری مزایا و معایبی دارند و الگوریتم‌های که به ظرفیت و بار منابع توجه می‌کنند عملکرد بهتری دارند.

۵. مراجع

1. Mell P., Grance T.(2011), " The NIST Definition of Cloud Computing (Draft). NIST

2. Shaw S.B, Singh, A.K.(2014),"A Survey on Scheduling and Load Balancing Techniques in Cloud Computing Environment",Computer and Communication Technology (ICCCT),pp.87-95.

3. Buyya R. , Ranjan R. , Calheiros R.(2010), "InterCloud: Utility-oriented federation of cloud computing environments for scaling of application services," Proceeding ICA3PP'10 Proceedings of the 10th international conference on Algorithms and Architectures for Parallel Processing pp.13-31.

4. Foster I., Zhao Y., Raicu I., Lu S.(2008), "Cloud Computing and Grid Computing 360-degree compared," Grid Computing Environments Workshop, pp.99-106.

5. Grosu D., Chronopoulos A.T., Leung M.(2008), "Cooperative load balancing in distributed systems," Parallel and Distributed Processing Symposium., Proceedings International, IPDPS

6. Chen h., Wang F., Helian N., Akanmu G.(2013), "User-Priority Guided Min-Min Scheduling Algorithm For Load Balancing in Cloud Computing," Parallel Computing Technologies (PARCOMPTECH), pp.1-8.
7. Azawi Mohialdeen I.(2013), "Comparative Study of Scheduling Algorithms in Cloud Computing Environment, " Journal of Computer Science , pp. 252-263.
8. Zhang Q., Cheng L., Boutaba R.(2010), "Cloud computing: sate-of-art and research challenges," online: The Brazillian Computer Society.
9. Soni, G, Kalra, M.(2014), "A Novel Approach for Load Balancing in Cloud Data Center," Advance Computing Conference (IACC), pp.807-812.
10. Adhikari j., Patil s(2013) "Double Threshold Energy Aware Load Balancing in Cloud Computing," Computing, Communications and Networking Technologies (ICCCNT),pp.1-6.