2nd. International Conference on
Information Technology, Communications and Telecommunications (irICT2016)
1-2 March 2016 - Iran, Tehran

Archive of SID

Optimal Allocation of VM Resources According to Quality of Services Metrics for Reducing Cost in the Cloud Computing

Yaghoob Siahmargooei, M. Kazem Akbari, S. Alireza Hashemi Golpayegani, Saeed Sharifian

Department of Computer Engineering and Information Technology, Amirkabir University of Technology
E-mail: {y.s, akbarif, sa.hashemi}@aut.ac.ir
The Electrical and Electronics Engineering Department, Amirkabir University of Technology
E-mail: sharifian_s@aut.ac.ir

Abstract. Resource management and reduce the cost are the most important issues in cloud computing. In this article we try to optimize the amount of required processing resources of cloud infrastructure customers, in addition reduce the cost of rental resources. Hence, we want to determine the optimal number of virtual machines and cloud services on their proper alignment with compliance standards such as response time and quality of service violate. For this purpose, the work has been done to help the problem and formulating it with mixed integer programming (MIP) mode. After specifying decision variables, we use TPC-W benchmark tests. we calculate the relationships between attributes. Then, proposed model and implemented similar models were compared. The results showed that better performance in terms of cost reduction and implementation model with service quality standards have been compared to other methods.

Keywords: cloud computing, resource management, quality of service, resource allocation, performance metrics, MIP.

تخصیص بهینه منابع ماشین‌های مجازی براساس معیارهای کیفیت خدمات با هدف کاهش هزینه در رایانش ابری

[۱]یعقوب سیاه‌مرگویی، [۲] محمدکاظم اکبری، [۳] سیدعلیرضا هاشمی گلپایگانی، [۴]سعید شریفیان

[۱] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، y.s@aut.ac.ir
[۲] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، akbarif@aut.ac.ir
[۲] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، sa.hashemi@aut.ac.ir
[۴] دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، sharifian_s@aut.ac.ir

چکیده

مدیریت منابع و کاهش هزینه از مهمترین مسائل در رایانش ابری است. در این مقاله سعی داریم با بهینه سازی مقدار منابع پردازشی مورد نیاز مشتریان زیرساخت ابری، هزینه اجاره منابع را کاهش دهیم. از این رو، هدف ما یافتن تعداد ماشین‌های مجازی بهینه و نحوه چیدمان مناسب سرویس‌های ابری روی آن‌ها با رعایت معیارهای کیفیت خدمات مانند زمان پاسخ و خطای ارائه سرویس است. برای این منظور پس بررسی کارهای انجام شده، اقدام به طرح مسئله و فرموله کردن آن با کمک مدلسازی برنامه‌ریزی عدد صحیح مختلط نمودیم. پس از مشخص کردن متغیرهای تصمیم، با استفاده از آزمایش محک TPC-W روابط بین مشخصه‌ها را محاسبه نمودیم. سپس مدل پیشنهادی را پیاده سازی کرده و با مدل‌های مشابه مقایسه نمودیم. نتایج ارزیابی نشان داد مدل پیشنهادی عملکرد بهتری از نظر کاهش هزینه و تطبیق با معیارهای کیفیت خدمات نسبت به دیگر روش‌ها داشته است.

واژه های کلیدی: رایانش ابری، مدیریت منابع، کیفیت خدمات، تخصیص منابع، معیارهای کارایی، بهینه سازی، برنامه ریزی عدد صحیح مختلط.

۱- مقدمه

پردازش، بزرگ‌ترین چالش پیش روی فناوری اطلاعات در سال‌های اخیر بوده است. با افزایش کاربران و سرویس گیرندگان و نیز رشد روز

www.SID.ir

Archive of SID

افزون اطلاعات، لحظه به لحظه نیاز به پردازنده و حافظه بیشتری برای پردازش این حجم عظیم اطلاعات خواهیم داشت. آنچه مسلم است لزوم توجه سازمان‌ها و شرکت‌های تجاری به مسئله قدرت پردازشی مورد نیاز برای ارائه خدمات به مشتریان است. شرکت‌ها درصدد برون سپاری خدمات خود به فراهم کنندگان خدمات نرم افزاری هستند. چرا که از طرفی دانش کافی برای تمرکز بر ابزارهای نوین نرم افزاری و سخت افزاری را ندارند و از طرف دیگر هزینه‌ی ایجاد و نهادینه کردن زیرساخت‌ها، توجیه اقتصادی خوبی ندارد [1].

یکی از جدیدترین مدل‌های محاسباتی که قادر به پردازش حجم بالایی از اطلاعات است رایانش ابری[1] نام دارد. یک معماری توزیع شده باقابلیت بهره‌گیری از دستاوردهای اینترنت برای ارائه خدمات باکیفیت به مشتریانی که حجم وسیعی از اطلاعات و محاسبات را روزانه بکار می‌گیرند. زیرساخت ابری دیگر نیاز سازمان‌ها و شرکت‌ها به مراکز عظیم داده و صرف هزینه‌های فراوان برای توسعه سخت افزاری و زیرساختی را از بین برده است [2].

شرکت‌ها به همان اندازه‌ای که از منابع محاسباتی نیاز دارند آن‌ها را در اختیار می‌گیرند و به همان اندازه پول پرداخت می‌کنند. شرکت‌ها و سازمان‌ها بعنوان مشتریان زیرساخت ابری باید بتوانند سرویس‌های الکترونیکی مورد نیاز خود را به‌درستی تعریف و پیاده سازی کرده و آن‌را جهت اجرا روی زیرساخت ابری آماده نمایند. همچنین برای کسب رضایت مشتری و ارائه کیفیت خدمات[2] مناسب‌تر، ضروری است بین کاربران و سازمان قرارداد سطح سرویس[3] وجود داشته باشد. با مشخص بودن این قرارداد سازمان موظف است بر اساس فاکتورهای مشخص شده سرویس درخواستی را در زمان مقرر به درخواست‌کننده ارائه دهد و در صورت تخطی از این مقادیر و عدم سرویس‌دهی سازمان متحمل جریمه‌های مالی مشخصی خواهد شد. لذا به‌منظور کاهش چنین هزینه‌هایی، رعایت کیفیت خدمات ضروری است. چه‌بسا تبعیت از اصول مشخص شده به رضایت بیشتر مشتریان منجر شود. فراهم کننده خدمات رایانش ابری نیز موظف است انواع خدمات و منابع قابل ارائه خود را مشخص نماید. انواع ماشین‌های مجازی در دسترس مشتری، خصوصیات و منابع هریک از جمله مواردی هستند که باید برای مشتری خدمات رایانش ابری مشخص باشد.

۲- طرح مسئله

مطالعات حاکی از آن است که اکثر مشتریان خدمات رایانش ابری طبق اصل پارتو[4]، در ۸۰٪ اوقات، تنها از ۲۰٪ منابع موجود خود استفاده می‌کنند که این مسئله نشان دهنده عدم توجه به مقدار منابع پردازشی مورد نیاز حتی در زمان اجاره خدمات رایانش ابری است [3].

حال نیازمند آن هستیم که مشخص نماییم، چگونه می‌توان با هدف کاهش هزینه اجاره منابع، معیارهای کارایی و کیفیت خدمات را ثابت نگه داریم و از تخطی‌شدن قرارداد سطح سرویس جلوگیری نماییم؟ آیا مدل و رابطه‌ای برای یافتن پاسخ بهینه برای مقادیر مختلف که تاثیر زیادی در تصمیم گیری دارند وجود دارد؟ این‌ها سوالات مهمی هستند که سرمنشأ شکل گیری مدل پیشنهادی می‌باشند. آنچه معلوم است، نیاز سازمان‌ها به داشتن یک مرجع روشن برای محاسبه دقیق مقدار منابع مورد نیاز پردازشی برای دوره‌های گوناگون براساس سرویس‌ها و جریان‌های کاری موجود در سازمان است. چرا که هر سازمان یا شرکت فعال در زمینه فناوری اطلاعات درحال استفاده از سرویس‌های مختلف در فرآیندهای مشخص برای ارائه خدمات به مشتریان و کارکنان خود است. از این‌رو نیاز به یک مدل دقیق جهت بهینه کردن تصمیمات و کاهش هزینه‌ها و حرکت در مسیر تبعیت از کیفیت خدمات و هماهنگی و هم‌راستایی با زیرساخت‌های رایانش ابری است.

۳- کارهای مرتبط

تحقیقات فراوانی در حوزه مدیریت و تامین منابع پردازشی در رایانش ابری طی سال‌های اخیر انجام شده است. با نگاهی اجمالی، اصلی‌ترین مباحث مورد علاقه محققان را عنوان می‌کنیم. این موارد عبارتند از، (۱) *مدیریت منابع پردازشی فیزیکی و زیرساختی براساس بازخورد کارایی و بهره وری*، که سعی در تطبیق تدریجی رویکرد تامین منابع پردازشی بر اساس مقدار کارکرد آن‌ها دارد. (۲) *مدل‌های مبتنی بر پیش بینی*، که با کمک الگوریتم‌ها و روش‌های آماری درصدد تعیین وضعیت آینده‌ی نزدیک برای درخواست‌ها و نیاز منابع پردازشی هستند. (۳) *تامین منابع با استفاده از محاسبات برون خط و پیچیده*، که با کمک ابزارها و روش‌های شناخته شده در حوزه هوش مصنوعی بسمت خودکارسازی و خودمختاری تامین منابع در فراهم کنندگان زیرساخت ابری پیش رفته اند. (۴) *مدل‌های بهینه سازی و ریاضی*، که با فرموله کردن مسئله بمنظور یافتن جواب بهینه از میان فضای پاسخ گسترده اقدام به یافتن مقادیر مورد نیاز برای افزایش صحت تخصیص منابع می‌نمایند. در این مقاله تمرکز اصلی روی مبحث آخر خواهد بود. چرا که راه‌حل پیشنهادی با رویکرد فرموله کردن مسئله بصورت مدل بهینه سازی ارائه بدنبال یافتن تصمیمات بهینه و متناسب با شرایط درنظر گرفته شده خواهد بود.

بمنظور بررسی تحقیقات برجسته می‌توان به کارهای [7-4] اشاره نمود که با استفاده از مدلسازی ریاضی مقدار منابع مورد نیاز در هر لحظه را براساس معیارهای کیفیت محاسبه کرده اند. هرچند در مورد تعداد و نوع ماشین‌های مجازی مورد نیاز و نیز چیدمان مناسب سرویس‌ها روی هر یک از ماشین‌های مجازی اشاره‌ای نشده است. رویکرد مورد استفاده در [8] نیز تنها به چیدمان سرویس‌های اینترنتی اشاره کرده است که ماهیت و خصوصیات سرویس‌های مختص رایانش ابری در آن لحاظ نگردیده است. در [9,10] با درنظر گرفتن اهمیت

[1] Cloud Computing
[2] QoS (Quality of Service)
[3] SLA (Service Level Agreement)
[4] Pareto

www.SID.ir

Archive of SID

قرارداد سطح سرویس و معیارهای موجود در آن اقدام به ارائه یک مدل جهت تعیین منابع پردازشی و تأمین آن‌ها در محیط ابری شرکت Amazon کرده است و تمرکزی روی هزینه و کاهش اجاره منابع نشده است. از میان کارهای انجام شده مدل پیشنهاد شده در [11]، با درنظر گرفتن تمام موارد با اهمیت از دیدگاه حاضر، با استفاده از یک مدل ریاضی جهت بهینه سازی و یافتن بهترین تصمیم برای دانستن مقدار منابع مورد نیاز براساس فاکتور هزینه، کارایی و کیفیت خدمات می‌توان هماهنگی کاملی در شناخت مسئله را درک کرد. همچنین با انجام آزمایش‌ها و نتایج قابل توجه در SPECWeb، می‌توان این کار را از پر اهمیت ترین تحقیقات در این حوزه قلمداد کرد. مدل پیشنهادی مقاله حاضر سعی دارد با تکیه بر دستاوردهای کارهای اخیر تلاش کند مدل ملموس تری براساس تطبیق با آنچه در واقعیت اتفاق می‌افتاد پیشنهاد نماید. راه حلی که براحتی بتوان با محیط مورد نظر و ساختار رایانش ابری و معیارهای سازمانی تطبیق داد.

۴- مدل پیشنهادی

پس از شناخت دقیق مسئله، اینک نیاز به فرموله کردن آن با کمک روش‌ها و ابزارهای مورد نیاز است. هدف، کاهش هزینه اجاره زیرساخت رایانش ابری یک سازمان باکمک بهینه سازی است. سازمانی که از دید فراهم کننده خدمات ابری یک مشتری محسوب می‌شود. بهینه سازی با کمک مدل سازی ریاضی به تلاش برای توسعه یک مدل بهینه برای یک سامانه مشخص گفته می‌شود. در اینجا مدل ریاضی جهت توصیف یک سامانه یا مسئله با استفاده از زبان ریاضی و قضیه‌ها و نمادهای موجود در آن بکار گرفته شده است. حال با استفاده از برنامه ریزی عدد صحیح مختلط[5] (MIP) اقدام به مدل‌سازی مسئله به زبان رسمی و ریاضی خواهیم کرد. عناصر اصلی در روش مورد استفاده شامل تابع هدف، اندیس‌ها و مجموعه‌ها، متغیرهای تصمیم، مشخصه‌های مستقل و وابسته و در نهایت محدودیت‌های مسئله خواهند بود.

۴-۱ تابع هدف

به‌طور صریح، هدف نهایی مورد نظر در مدل پیشنهادی کمینه کردن هزینه کل سازمان است. منظور از هزینه کل سازمان، مجموع هزینه‌های اجاره زیرساخت خدمات رایانش ابری و هزینه تخطی از قرارداد سطح سرویس است. تابع هدف به‌صورت رابطه (۱) تعریف می‌شود. مشخصه وابسته OC بیانگر هزینه کل سازمان است که نهایتاً باید کمینه گردد.

$$\text{Minimize} \quad OC \qquad (۱)$$

با توجه به اینکه مشخصه‌های اصلی تأثیر گذار روی هزینه نهایی نشأت گرفته از منابع مصرفی، زمان پاسخ[6] و تعداد خطاهای[7] درخواست سرویس است، لذا برای کمینه کردن هدف نهایی، نیاز به کمینه کردن موارد زیر داریم: (۱) کمینه کردن هزینه‌ی ماشین‌های مجازی اجاره شده از فراهم کننده، (۲) کمینه کردن زمان پاسخ به درخواست سرویس‌ها، (۳) کمینه کردن تعداد خطاهای رخ داده برای پاسخ به درخواست سرویس‌ها بود.

۴-۲ اندیس‌ها و مجموعه‌ها

به‌منظور شمارش و نام‌گذاری موجودیت‌های مورد استفاده در مسئله از اندیس‌ها و مجموعه‌ها استفاده می‌کنیم. با بررسی مفروضات مسئله و بهره گرفتن از معلومات و مجهولات می‌توان موجودیت‌های اصلی تاثیر گذار در مسئله را یافت. سرویس‌ها، جریان‌های کاری، منابع پردازشی مختلف، دوره‌ها و پنجره‌های زمانی و ماشین‌های مجازی از موجودیت‌های تاثیرگذار در مسئله حاضر هستند. پس به‌طور خلاصه اندیس‌ها و مجموعه‌های مدل پیشنهادی در جدول (۱) بیان شده است.

جدول (۱) : اندیس‌ها و مجموعه‌ها

عنوان	محدوده	توضیحات
v	1,…,V	اندیس انواع ماشین‌های مجازی
i	0,…,I	اندیس نمونه برای هر یک از انواع ماشین‌های مجازی
r	1,…,R	اندیس مجموعه منابع پردازشی
s	1,…,S	اندیس سرویس‌ها
w	1,…,W	اندیس جریان‌های کاری
t	1,…,T	اندیس پنجره‌های زمانی برای درخواست‌های سرویس

۴-۳ متغیرهای تصمیم

برای دستیابی به اهداف ذکر شده و درنهایت هدف اصلی مدل، نیاز به تصمیماتی داریم که آن‌ها را در قالب متغیرهای تصمیم بیان خواهیم کرد. از متغیرهای تصمیم در مرحله حل، به‌عنوان ورودی مدل استفاده خواهد شد تا بتوان در هر لحظه با تصمیم گیری مناسب، به بهینه‌ترین حالت ممکن رسید. همان‌طور که در جدول (۲) مشاهده می‌کنید مهم‌ترین تصمیم در این مدل، تعداد ماشین‌های مجازی اجاره شده خواهد بود که هریک دارای میزان منابع پردازشی و ذخیره سازی معینی باشد. بر اساس اهداف مشخص شده مدل به ما می‌گوید که در هرلحظه با توجه به مشخصه‌ها و محدودیت‌ها و شرایط حاکم چه تعداد ماشین مجازی از هر نوع و با چه میزان منابع پردازشی مورد نیاز است. با توجه به مقدار منابع مورد نیاز در پنجره زمانی مورد نظر می‌توان تعیین کرد که چه تعداد از ماشین‌های مجازی باید روشن یا خاموش باشند و چه سرویس‌هایی باید روی کدام ماشین‌های مجازی اجرا شوند.

(۲) : متغیرهای تصمیم

متغیر تصمیم	توضیحات
X_v	تعداد نمونه‌های تخصیص داده شده از ماشین مجازی v

[5] MIP (Mixed Integer Programming)
[6] Response Time
[7] Fault

www.SID.ir

Archive of SID

$A_{v,i,s}$	وضعیت اجرای سرویس s روی ماشین مجازی i از نوع v (متغیر دودویی)
$M_{v,r}$	مقدار ظرفیت منبع r در ماشین مجازی v

۴-۴ مشخصه ها

اکنون نیاز است مشخصه‌های موجود در مسئله مورد بررسی قرار گیرد. مشخصه‌هایی که هریک به‌نوعی در مدل مؤثر بوده و دارای ماهیت و خصوصیات متفاوتی هستند. لیست کامل مشخصه‌های مورد استفاده در مدل در جدول (۳) موجود است.

جدول (۳) : مشخصه‌ها

مشخصه	توضیح
$D_{s,t}$	تعداد درخواست‌های سرویس s در پنجره زمانی t
$N_{r,s}$	مقدار منبع مورد نیاز نوع r برای هر بار اجرای سرویس s
$SN_{s,r}$	مقدار منبع مورد نیاز از نوع r برای راه اندازی سرویس s
$VN_{v,r}$	مقدار منبع مورد نیاز از نوع r برای راه اندازی ماشین مجازی v
$PB_{w,s,s'}$	احتمال اجرای سرویس 's بعد از سرویس s در جریان کاری w
C_r	هزینه اجاره برای هر واحد از منبع r
C_v	هزینه پایه برای اجاره ماشین مجازی اولیه از نوع v
Q	فاکتور قرارداد سطح سرویس (SLA) (حداقل درصد مجاز برای پاسخ به سرویس‌ها در زمان اجرای آن‌ها)
E_s	مدت زمان تخمینی برای اجرا و پاسخ به یک درخواست از سرویس s
DF_r	ضریب بروز تأخیر در اجرای سرویس‌ها در اثر کمبود منبع r در پنجره زمانی t
FF_r	ضریب بروز خطا در لحظه اجرای سرویس‌ها در اثر کمبود منبع r در پنجره زمانی t
DS	هزینه جریمه برای تأخیر در زمان پاسخ به سرویس‌ها در پنجره زمانی t
FS	هزینه جریمه برای خطا در زمان پاسخ به سرویس‌ها در پنجره زمانی t

۴-۵ محاسبه روابط متغیرهای کارایی

حال بمنظور محاسبه روابط بین متغیرها و نیز یافتن تاثیر هریک از مشخصه‌ها روی تابع هدف، نیاز به بررسی شرایط واقعی مسئله را از طریق آزمایش داریم. در اینجا برای محاسبه تاثیرات مشخصه‌های کیفیت خدمات و کارایی با انجام آزمایش روی ماشین‌های مجازی میزان روابط موثر در تاخیر و خطا محاسبه خواهد شد. هدف از این آزمایش یافتن روابط حاکم بر زمان پاسخ و تعداد خطا خواهد بود. اثرات مورد بررسی، زمان پاسخ به درخواست اجرای یک سرویس و همچنین بروز خطا در صورت عدم پاسخ به درخواست مورد نظر خواهد بود. بستر مورد آزمایش محک TPC-W[8] می‌باشد که براساس تطبیق با نیازهای ما یعنی سرویس‌های با زمان پاسخ کوتاه، انتخاب شده است. در اینجا تمرکز بر روی یک سرویس قرار گرفته است تا براساس تابع توزیع پوآسن[9]، در هر ثانیه ۵ کاربر آن‌را درخواست کنند، بدون توجه به اینکه درخواست سرویس این کاربران در ثانیه فعلی پاسخ داده خواهد شد یا خیر، در هر ثانیه ۵ کاربر دیگر سرویس را درخواست می‌کنند. پس به‌طور متوالی در هر ثانیه ۵ کاربر در حال درخواست سرویس هستند. حداکثر کاربران ورودی به سیستم ۱۰۰۰ کاربر خواهد بود، لذا در حالت طبیعی و در زمانی که تمام کاربران در همان ثانیه و بدون هیچ وقفه‌ای پاسخ درخواست سرویس را دریافت نمایند، پس از گذشت مدت زمان ۲۰۰ ثانیه کار سیستم و این آزمایش به پایان خواهد رسید. با استفاده از روش طراحی آزمایش‌ها[10]، ۱۰ سناریو برای منابع موجود در ماشین مجازی سرویس دهنده که محک روی آن قرار گرفته است، در نظر گرفته شده است. همچنین مقدار تحمل کاربران در تأخیر برای اجرای سرویس و پاسخ سرور زمان ۳۰ ثانیه در نظر گرفته شد، لذا درخواست‌هایی که پس از این مدت پاسخ داده شوند شامل موارد عدم پاسخ و تخطی از قرارداد سطح سرویس محسوب خواهند شد. جهت اجرای این آزمایش از ابزار تولید ترافیک LoadUIWeb به‌منظور تولید ترافیک و درخواست‌های شبیه سازی شده به سمت سرور محک TPC-W استفاده شد. خروجی این آزمایش در واقع مدت زمان پاسخ به ۱۰۰۰ درخواست و تعداد خطاهای رخ داده برای هریک از سناریوها می‌باشد. در هر سناریو متوسط زمان پاسخ و حداکثر زمان پاسخ به همراه تعداد درخواست‌های بی‌پاسخ محاسبه شده است. همچنین نتایج حاصل از این آزمایش در جدول (۴) قابل مشاهده است.

جدول (۴) : نتایج حاصل از آزمایش معیارهای کیفیت خدمات

سناریو	منابع ماشین مجازی مورد آزمایش			متوسط زمان پاسخ	حداکثر زمان پاسخ	عدم پاسخ
	Core	CPU (GHz)	RAM (Mb)			
1	2	2.5	2048	0.4	4.84	0
2	2	2.5	1024	2.5	15.03	0
3	2	2.5	512	6.9	30.44	4
4	1	2.5	2048	0.7	6.42	0
5	1	2.5	1024	3.4	13.4	0
6	1	2.5	512	6.4	28.7	0
7	1	1.2	512	4.7	21.3	0
8	1	1.2	2048	2	9.8	0
9	1	0.6	2048	26.3	60	951
10	2	2.5	256	250	250	1000

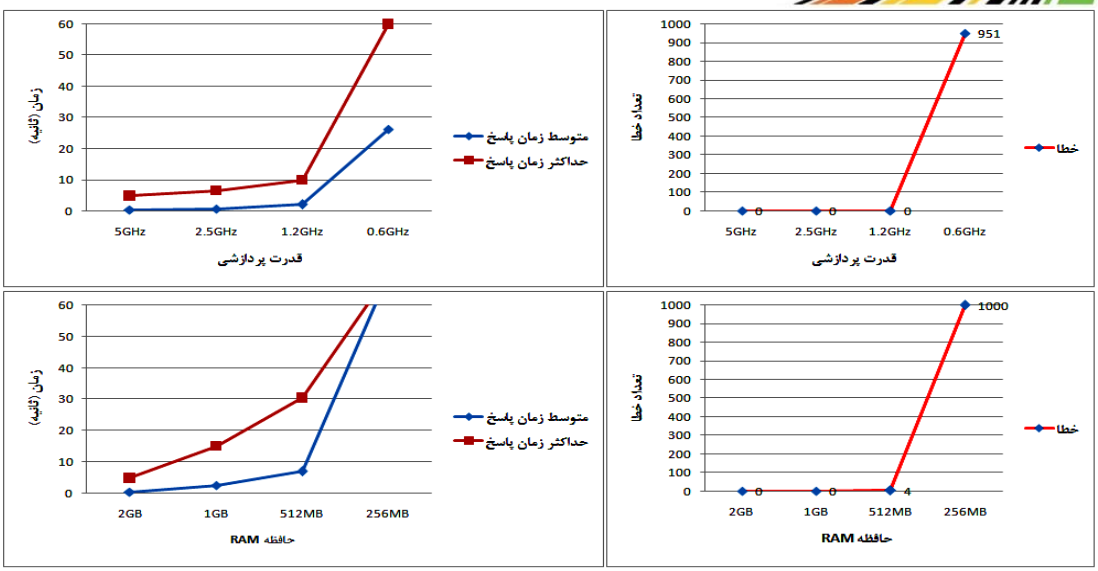
به‌منظور تحلیل این آزمایش می‌توان گفت مقدار منابع پردازشی موجود در ماشین مجازی تأثیر فراوانی بر زمان پاسخ به درخواست‌های کاربران خواهد داشت. لذا در نظر گرفتن منابع پردازشی کافی از مهم‌ترین اولویت‌ها خواهد بود. همان‌طور که در شکل (۱) مشاهده می‌کنید، با کاهش مقدار منابع تخصیص داده شده به ماشین مجازی زمان پاسخ و تعداد خطا به‌طور قابل ملاحظه‌ای افزایش می‌یابد که می‌توان روابط را برهمین اساس بنا کرد. مقادیر مندرج در این نمودار بر اساس منابع انتخاب شده در سناریوها در نظر گرفته شده است، بطوریکه با ثابت فرض کردن یکی از منابع مثل پردازنده، شرایط آزمایش را در صورت کم یا اضافه کردن حافظه بررسی نمودیم.

[8] Benchmark

[9] Poison

[10] DOE (Design of Experiments)

www.SID.ir



شکل (۱) : زمان پاسخ و تعداد خطای پاسخ به درخواست‌ها

۴-۶ متغیرهای وابسته

به‌منظور ایجاد روابط ریاضی و محاسبه ارتباط بین مشخصه‌ها از متغیرهای وابسته استفاده می‌کنیم. با کمک اندیس‌ها و متغیرهای تصمیم و نتایج حاصل از آزمایش قبل، روابط بین متغیرها بمنظور رسیدن به تابع هدف نهایی ساخته خواهد شد. روش تعریف به‌صورت سلسله مراتبی از پایین به بالا خواهد بود، لذا هریک از مشخصه‌های ذکر شده در هر مرحله، بخشی از مشخصه بعدی خواهد بود. در نهایت به مشخصه اصلی مسئله یعنی هزینه کل خواهیم رسید.

رابطه (۲) مقدار منبع مورد نیاز از نوع r برای اجرای تمام درخواست‌های رسیده از سرویس s را در پنجره زمانی t تعریف می‌کند. این مقدار را جهت استفاده‌های بعدی $RN_{s,r,t}$ می‌نامیم.

$$RN_{s,r,t} = (N_{r,s} \times D_{s,t}) + SN_{s,r} \qquad (۲)$$

در رابطه (۳) ماتریس مجاورت جریان کاری w مشخص شده است که احتمال حرکت هر سرویس به دیگر سرویس‌ها در جریان کاری w را نشان می‌دهد. هر درایه از این ماتریس بیانگر احتمال اجرای سرویس 's پس از سرویس s می‌باشد. بر اساس جریان کاری مشخص به‌راحتی می‌توان ماتریس مجاورت متناظر با آن جریان کاری را محاسبه نمود.

$$T_w = \begin{bmatrix} PB_{1,1} & \dots & PB_{1,s'} & \dots & PB_{1,S} \\ \dots & \dots & \dots & \dots & \dots \\ PB_{s,1} & \dots & PB_{s,s'} & \dots & PB_{s,S} \\ \dots & \dots & \dots & \dots & \dots \\ PB_{S,1} & \dots & PB_{S,s'} & \dots & PB_{S,S} \end{bmatrix} \qquad (۳)$$

رابطه (۴) مقدار منبع مورد نیاز از نوع r برای اجرای تمام درخواست‌های سرویس s در پنجره زمانی بعدی (t+1) را نشان می‌دهد. این مقدار که با $RN_{s,r,t+1}$ مشخص می‌شود، از روی احتمال اجرای سرویس‌ها پس از یک دیگر محاسبه می‌شود. در صورت نیاز به پیش‌بینی، این مقدار به $RN_{s,r,t}$ اضافه خواهد شد.

$$RN_{s,r,t+1} = \sum_{w=1}^{W} \sum_{s'=1}^{S} \sum_{s=1}^{S} \sum_{r=1}^{R} \left((N_{r,s'} \times D_{s,t} \times PB_{w,s,s'}) + SN_{s',r} \right) \qquad (۴)$$

در رابطه (۵) هزینه کل منابع مورد نیاز برای سرویس‌های درخواستی در پنجره زمانی t نشان داده شده است. این مقدار که با NC_t نمایش داده می‌شود، به آن هزینه خالص نیز گفته می‌شود.

$$NC_t = \sum_{r=1}^{R} \left(\sum_{w=1}^{W} \sum_{s=1}^{S} N_{r,s} \times D_{s,t} \right) \times C_r \qquad (۵)$$

در رابطه (۶) هزینه کل ماشین‌های مجازی اجاره شده از فراهم کننده خدمات رایانش ابری و منابع آن‌ها در پنجره زمانی t را مشاهده می‌کنید. این مقدار با PC_t نشان داده می‌شود.

$$PC_t = \sum_{v=1}^{V} \sum_{i=0}^{I} (C_v + \sum_{r=1}^{R} M_{v,r} \times C_r) \qquad (۶)$$

در رابطه (۷) ضریب بروز تأخیر در پاسخ به درخواست اجرای سرویس‌ها در پنجره زمانی t آورده شده است. این مقدار که با DP_t نمایش داده می‌شود، که حاصل نتایج بدست آمده از آزمایش بخش قبل و مشخصه‌های کارایی و کیفیت خدمات است. در [12] نیز از روش مشابهی برای تخمین زمان مورد نیز جهت اجرای یک سرویس برمنبای پردازنده و حافظه با استفاده از محک‌های مختلف، پیشنهاد شده است. لذا بر اساس نتایج به دست آمده در آزمایش بخش ۴-۵ و تحقیقات صورت گرفته در منبع ذکر شده، می‌توان رابطه‌های (۷) و (۸) را با کمک رگرسیون محاسبه کرد.

$$DP_t = \sum_{r=1}^{R} \left(\sum_{w=1}^{W} \sum_{s=1}^{S} \frac{E_s}{60} + \sum_{w=1}^{W} \sum_{s=1}^{S} {D_{s,t}}^2 + \sum_{w=1}^{W} \sum_{s=1}^{S} N_{r,s} \times \left(1 - \frac{\left(\sum_{v=1}^{V} \left(\sum_{i=0}^{I} M_{v,r} \times A_{v,i,s} \right) - VN_{v,r} \right)}{\sum_{w=1}^{W} \sum_{s=1}^{S} RN_{s,r,t}} \right) \times Q \times DF_r \right) \qquad (۷)$$

در رابطه (۸) ضریب بروز خطا در پاسخ به درخواست اجرای سرویس‌ها در پنجره زمانی t را مشاهده می‌کنید. در FP_t نیز مانند رابطه قبل، مشخصه‌های ذکر شده به همراه فاکتور تأثیر خطا برای هر منبع مؤثر هستند.

Archive of SID
www.SID.ir

Archive of SID

$$FP_t = \sum_{r=1}^{R}\left(\sum_{w=1}^{W}\sum_{s=1}^{S}\frac{E_s}{60} + \sum_{w=1}^{W}\sum_{s=1}^{S} {D_{s,t}}^2 + \sum_{w=1}^{W}\sum_{s=1}^{S} N_{r,s_r} \times \left(1 - \frac{\left(\sum_{v=1}^{V}\left(\sum_{i=0}^{I} M_{v,r} \times A_{v,i,s}\right) - VN_{v,r}\right)}{\sum_{w=1}^{W}\sum_{s=1}^{S} RN_{s,r,t}}\right) \times Q \times FF_r\right) \quad (۸)$$

در رابطه (۹) هزینه تخمینی کل برای مجموع تأخیرها در لحظه پاسخ به سرویس‌ها در پنجره زمانی t به‌صورت DC_t بیان می‌شود. این مقدار با استفاده از حاصل ضرب فاکتور مقدار تأخیر و جریمه تأخیر محاسبه می‌شود.

$$DC_t = DP_t \times DS \quad (۹)$$

در رابطه (۱۰) هزینه تخمینی کل برای مجموع خطاهای رخ داده برای پاسخ به سرویس‌ها در پنجره زمانی t را مشاهده می‌کنید که با FC_t نشان داده می‌شود.

$$FC_t = FP_t \times FS \quad (۱۰)$$

همانطور که اشاره شد، رابطه (۱۱) هزینه کل سازمان برای ارائه سرویس به درخواست‌ها در پنجره زمانی t را نشان می‌دهد. این مقدار حاصل مجموع هزینه کل ماشین‌های مجازی اجاره شده و هزینه کل تخمینی برای مجموع تأخیرها و خطاها در پنجره زمانی t خواهدبود. در نهایت، محدودیت‌های مسئله در مدل ریاضی به‌صورت روابط (۱۵ تا ۱۲) بیان شده‌است.

$$OC_t = PC_t + DC_t + FC_t \quad (۱۱)$$

$$A_{v,s} \in \{0,1\} \quad (۱۲)$$

$$0 \le PB_{w,s,s'} \le 1 \quad (۱۳)$$

$$SN_{s,r,t} + VN_{v,r,t} \ge 0 \quad (۱۴)$$

$$\sum_{v=1}^{V}\sum_{i=1}^{I}\sum_{s=1}^{S} D_{s,t} \times A_{v,i,s} = \sum_{w=1}^{W}\sum_{s=1}^{S} D_{s,t} \quad (۱۵)$$

۷-۴ پیاده سازی و اجرای مدل

اینک برای اجرای مدل پیشنهادی، نیاز است ابتدا ورودی مدل را مشخص نماییم. بمنظور سهولت در محاسبات از سه نوع ماشین مجازی که در زیرساخت ابر تجارت شرکت Amazon بکار گرفته می‌شود، استفاده نمودیم. مشخصات هریک از ماشین‌های مجازی در جدول (۵) آورده شده است. ورودی دیگر مدل، تاریخچه درخواست‌های سرویس در محک TPC-W است که بعنوان بستر آزمایشی درنظر گرفته شده بود. با درنظر گرفتن تاریخچه ۸ ساعته از درخواست‌ها و انتخاب پنجره زمانی به اندازه ۲ دقیقه، مشخصه‌های مدل با داده‌های ورودی مقداردهی شدند. سپس با کمک ابزارهای بهینه سازی GAMS و LINGO اقدام به بهینه سازی تعداد ماشین‌های مجازی مورد نیاز برای داده‌های ورودی نمودیم. نتیجه این‌کار یافتن بهترین مقادیر برای متغیرهای تصمیم خواهد بود که بصورت ماتریس CN و VN در رابطه‌های (۱۶) و (۱۷) نشان داده شده است. این ماتریس‌ها بترتیب نشان دهنده تعداد بهینه ماشین‌های مجازی مورد نیاز و نیز چیدمان بهینه سرویس‌ها روی هریک از ماشین‌های مجازی برای هر پنجره زمانی هستند. براساس متغیرهای تصمیم، مقدار منابع هریک از ماشین‌های مجازی مشخص شده در ماتریس CN نیز از طریق جدول (۵) قابل محاسبه است.

$$CN = \begin{bmatrix} X_{1,t} & \dots & X_{v,t} & \dots & X_{V,t} \end{bmatrix} \quad (۱۶)$$

$$VN = \begin{bmatrix} A_{1,1,1} & \dots & A_{v,i,1} & \dots & A_{V,I,1} \\ A_{1,1,s} & \dots & A_{v,i,s} & \dots & A_{V,I,s} \\ A_{1,1,S} & \dots & A_{v,i,S} & \dots & A_{V,I,S} \end{bmatrix} \quad (۱۷)$$

در رابطه (۱۷) ماتریس مجموعه سرویس‌های اجرا شونده روی نمونه‌های ماشین‌های مجازی نشان داده شده است. هر یک از درایه‌ها یک عدد دودوبیی است که بیانگر اجرا یا عدم راه اندازی و اجرای سرویس s روی ماشین مجازی i از نوع v خواهد بود.

جدول (۵) : نمونه ماشین‌های مجازی مدل بر اساس Amazon EC2

شناسه	ماشین مجازی	CPU (Mips)	RAM (Mb)	هزینه اجاره بر ساعت
۱	t1.micro	1.7	660	$0.020
۲	c1.medium	7	1825	$0.145
۳	c3.large	9	3700	$0.150

۵- ارزیابی و نتیجه گیری

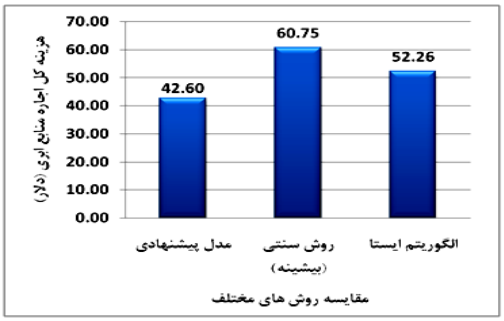
در این بخش، بمنظور ارزیابی نتایج حاصل از اجرای مدل اقدام به مقایسه نتایج با دیگر روش‌های مطرح نمودیم. فراهم کنندگان خدمات رایانش ابری راهکارهای نسبتاً مناسبی را برای کاهش هزینه‌های مشتریان خود ارائه کرده اند. روش‌های مبتنی بر آستانه[11] برای تخصیص لحظه به لحظه منابع پردازشی امروزه در زیرساخت ابری Amazon و Rightscale در حال استفاده است. لذا به جهت بررسی اعتبار مدل پیشنهادی و اطمینان از صحت عملکرد آن، اقدام به پیاده سازی روش تخصیص منابع براساس آستانه نمودیم و با دادن ورودی‌های مدل پیشنهادی به آن نتایج حاصل را جمع آوری کردیم. از آنجا که هدف کاهش هزینه اجاره خدمات رایانش ابری می‌باشد، هزینه اجاره تعداد ماشین‌های مجازی را برای سه حالت استفاده از مدل پیشنهادی، استفاده از روش مبتنی بر آستانه و روش سنتی یعنی تخصیص به میزان بدترین حالت با یکدیگر مقایسه نمودیم. همانطور که در شکل (۲) مشاهده می‌کنید، درصورت استفاده از مدل پیشنهادی هزینه اجاره منابع برای داده‌های ورودی ۴۲,۶ دلار خواهد بود، درصورت استفاده از الگوریتم ایستا و مبتنی بر آستانه این مقدار ۵۲,۲۶ دلار خواهد بود. همچنین درصورت عدم استفاده از روش‌های تخصیص بهینه و درنظر گرفتن بدترین شرایط ممکن برای منابع این مقدار ۶۰,۷۵ دلار خواهد بود. هرچند در روش سنتی به دلیل استفاده از منابع رزرو شده بجای درخواست لحظه‌ای منابع، اجاره با مقدار ۲۵٪ تخفیف همراه است اما بازهم بیشترین هزینه را دربر گرفته است. همانگونه که روشن است، هزینه منابع مورد نیاز درصورت استفاده از مدل پیشنهادی، بطور قابل ملاحظه‌ای کاهش یافته است. همچنین

[11] Threshold-based

www.SID.ir

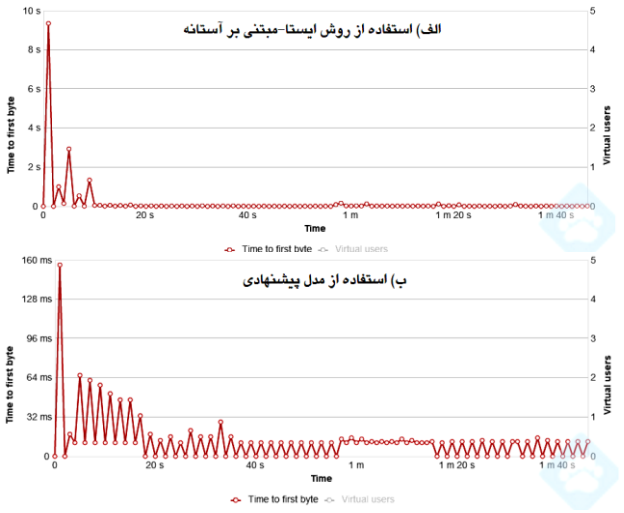
Archive of SID

بدلیل اینکه مدل پیشنهادی چیدمان بهینه ای برای سرویس‌ها روی ماشین‌های مجازی محاسبه کرده است، کارکرد منابع پردازشی هریک از ماشین‌های مجازی نیز در حال متعادل قرار خواهد گرفت.



شکل (۲) : زمان پاسخ و تعداد خطای پاسخ به درخواست‌ها

همچنین معیارهای کارایی و کیفیت سرویس نیز در هر وضعیت مورد بررسی قرار گرفت. شکل (۳)-الف نشان دهنده زمان پاسخ به درخواست‌ها در حین درخواست سرویس‌ها به ماشین‌های مجازی برای روش مبتنی بر آستانه است. همانطور که مشاهده می‌کنید، زمان پاسخ به درخواست‌ها در بدترین حالت ممکن تا ۱۰ ثانیه رسیده است. در شکل (۳)-ب نیز مقادیر زمان پاسخ برای مدل پیشنهادی نشان داده شده است. همانطور که مشخص است، در بدترین حالت ممکن زمان پاسخ به درخواست‌ها و تاخیر در اجرای سرویس‌ها تنها ۱۶۰ میلی ثانیه گزارش شده است. هرچند این مقادیر برای مدل پیشنهادی بسیار مناسب است اما درصورت استفاده از روش سنتی و تخصیص بیشترین مقدار منابع پردازشی مورد نیاز زمان پاسخ می‌تواند به کمتر از این مقدار نیز برسد. اما بدلیل بالا بودن هزینه اجاره منابع به هیچ عنوان معقول بنظر نمی‌رسد.



شکل (۳) : زمان پاسخ و تعداد خطای پاسخ به درخواست‌ها

۶- مراجع

[1] Seth, A., et al. "Integrating SOA and Cloud Computing for SME Business Objective." **WSEAS Transactions on Computers**, *USA* 3 (2012).

[2] Yeo, S., and H-HS L. "Using mathematical modeling in provisioning a heterogeneous cloud computing environment." **Computer** 44.8 (2011): 55-62.

[3] Chen, Y., et al. "An efficient resource management system for on-line virtual cluster provision." **Cloud Computing, 2009. CLOUD'09. IEEE International Conference on**. IEEE, 2009.

[4] Chandra, A., et al. "Dynamic resource allocation for shared data centers using online measurements." **Quality of Service—IWQoS 2003**. Springer Berlin Heidelberg, 2003. 381-398.

[5] Dutreilh, X., et al. "Using reinforcement learning for autonomic resource allocation in clouds: towards a fully automated workflow." *ICAS 2011*, **The Seventh International Conference on Autonomic and Autonomous Systems**. 2011.

[6] Roy, N., et al. "Efficient autoscaling in the cloud using predictive models for workload forecasting." **Cloud Computing (CLOUD), 2011 IEEE International Conference on**. IEEE, 2011.

[7] Villela, D., et al. "Provisioning servers in the application tier for e-commerce systems." **ACM Transactions on Internet Technology (TOIT)** 7.1 (2007): 7.

[8] Chen, G., et al. "Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services." **NSDI**. Vol. 8. 2008.

[9] Maurer, M., et al. "Enacting SLAs in clouds using rules." **Euro-Par 2011 Parallel Processing**. Springer Berlin Heidelberg, 2011. 455-466.

[10] Mi, H., et al. "Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers." **Services Computing (SCC), 2010 IEEE International Conference on**. IEEE, 2010.

[11] Ardagna, D., et al. "Dual time-scale distributed capacity allocation and load redirect algorithms for cloud systems." *Journal of Parallel and Distributed Computing* 72.6 (2012): 796-808.

[12] Stewart, C., and Kai S. "Performance modeling and system management for multi-component online services." **Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2**. USENIX Association, 2005.

www.SID.ir