클라우드 환경에서의 무중단 수직 확장에 관한 연구

[1*]박준석, [2] 고대식

A study on live vertical scale-up in a cloud environment

[1*]Jun-Seok Park, [2]Dae-Sik Ko

요 약

본 논문에서는 클라우드 자원의 무중단 수직 확장 서비스 제공을 위한 VMP(Virtual Machine Placement) 방안을 제시하였다. 수직확장을 위해서는 물리 서버의 여유 공간을 사전에 확보해야 하기 때문에, 이를 위한 FirstFit 배치 전략 기반의 가상 서버 할당율을 가변적으로 조정하는 "일반-혼합-수직의 모드 전환" 알고리즘을 제시하였으며, 수직 확장 비율, 가상화율, 여유자원율 등을 파라미터로 하여 시뮬레이션을 수행하였다. 시뮬레이션 결과, 수직 확장 비율이 50% 일 경우에 여유 공간을 고려하면 전체적으로 150%의 자원의 필요하나, 제안한 알고리즘의 시뮬레이션 결과로는 최대 125%의 여유 공간만을 필요로 하는 것으로 나타났다.

Abstract

In this paper, we proposed a Virtual Machine Placement (VMP) method to provide live vertical scaling services for cloud resources. Since free space on the physical server must be secured in advance for vertical scaling, a "general-mixed-vertical" mode conversion algorithm based on the FirstFit placement strategy that variably adjusts the allocation ratio of virtual servers to physical servers for this purpose is presented. Simulations were performed using parameters such as vertical scaling ratio, virtualization ratio, and free resource ratio. When the vertical scaling ratio is 50%, considering free space, 150% of resources are required as a whole, but simulation results of the proposed algorithm show that only up to 125% of free space is required.

Keywords: *Vertical scaling, VMP, FirstFit, Hot-Plug, Cloud*

[1*] *Corresponding Author* *세림티에스지㈜ 클라우드 연구소 소장 (junseok.park@selim.kr)*

[2] *목원대학교 전자공학과 교수 (kds@mokwon.ac.kr)*

Received: Nov. 30, 2022, Revised: Dec. 16, 2022, Accepted: Dec. 19, 2022

I. 서론

클라우드에서 제공하는 서비스는 크게 IaaS, PaaS, SaaS 등으로 분류된다, 이중 IaaS 는 모든 클라우드 서비스 제공자(Cloud Service Provider: 이하 CSP)가 제공하는 기본적인 서비스로써, CSP 가 보유하고 있는 물리서버(Physical Machine: 이하 PM)를 비롯한 클라우드 인프라를 통해 생성된 가상서버(Virtual Machine: 이하 VM)를 고객에게 제공하는 것을 일컫는다. IaaS 서비스는 컴퓨팅(Compute), 스토리지(Storage), 네트워크(Network)의 3 가지를 중심으로 제공되며, 이 중 컴퓨팅 서비스는 정해진 사양의 CPU, Memory, Disk 를 운영체제(Operating System)와 함께 제공한다.

클라우드 서비스를 제공하는 CSP 나 기관의 사설 클라우드 입장에서는 최소의 비용으로 최대의 서비스 제공 또는 최대의 자원 활용을 목표로 한다. 특히, IaaS 의 경우는 준비된 물리 자원하에서 최대한 많은 가상자원을 생성하여 제공하여야 하나, 일정 서비스 수준을 유지하기 위해 적정량의 가상화율을 지키게 된다. 이 때, 정해진 가상화율을 기준으로 최대한의 가상자원을 최소한의 물리 자원에 배치함으로써, 허용 가능한 최대한의 효율을 올리고자하는 실질적 방법이 VMP 이다. VMP 는 사용자가 요구하는 사양의 VM 을 어떠한 물리 자원에 배치하는 것이 효과적인가 하는 문제를 다루는 연구 분야이며, 이는 서비스의 수준 및 서비스 제공을 위한 비용과 직결되는 문제로 최근까지 다양한 연구가 진행되고 있다[1][2][3][4].

클라우드를 사용하는 대표적인 이유는 신속성과 탄력성에 있다. 신속성은 사용자의 요구에 즉시 서비스를 제공할 수 있음을 의미하며, 탄력성은 사용자 또는 환경의 변화에 따라 제공하는 서비스를 확장, 축소, 재배치하는 등의 자유로운 운영 환경을 제공하는 것이다. 특히, 클라우드의 확장성은 낮은 비용으로부터 사용자 서비스를 시작하고, 활용이 많아질수록 자원을 비례적으로 증가시킬 수 있으므로, 효율적인 비용 투자가 가능하다. 한편으로, 서비스의 고가용성은 서비스의 중단을 최소화하여 사용자 폭주를 비롯한 다양한 상황이 발생할 경우에도 이를 유지시킬 수 있는 특징을 일컷는다. 서비스의 중단은 서비스 품질의 저하, 고객의 이탈, SLA 미충족에 의한 계약의 위반 등 치명적인 상황을 야기하므로 가능하면 높은 가용성을 유지시키는 것이 필요하다. 하지만, 고가용성을 유지하기 위해서는 그에 따른 비용이 추가적으로 발생하기 때문에 적절한 수준에서 결정하게 된다.

본 연구는 IaaS 로 제공되는 VM 의 서비스 중단 없는 수직 확장 방안에 대한 것으로, 클라우드의 수직 확장 방안에 대해서는 그동안 미미하게 진행된 연구를 보완하고[5][6], 실제 적용 가능성을 제시 및 분석하였다.

II. 클라우드 자원 확장

컴퓨팅 자원의 확장은 수직 확장(Vertical Scaling)과 수평 확장(Horizontal Scaling)으로 구분할 수 있다. 수직 확장은 VM 의 사양을 상향하는 것을 의미하며, 수평 확장은 동일한 VM 을 여러 개 제공하는 것으로 요약할 수 있다. 이러한 차이로 인해서, 내부적으로 확장을 처리하는 방법이 상이하다.

먼저, 수평 확장은 서로 다른 다수의 VM 을 동시에 사용하는 것이기 때문에 VM 의 앞 단에 로드밸런서(Loadbalancer: 이하 LB)를 위치시킨 후, 사용자의 요청을 적절히 분배한다. 사용자의 요청이 많아지면, 동일한 사양의 VM 을 새롭게 생성시킨 후, LB 의 멤버로서 포함시키는 방법으로 확장을 수행한다. 동일한 방법으로 사용자의 요청이 적어지면, VM 을 제거하고, LB 의 멤버에서도 삭제하는 방법으로 축소를 수행한다.

한편, 수직 확장은 VM 을 정지시킨 후, 동일 또는 새로운 PM(VM 의 가상화를 수행하는 서버)에 새로운 사양으로 VM 을 생성한다. 이러한 방법은 사양 변경(Resizing)을 의미하며, 따라서 확장 또는 축소 모두 가능하다. 그러나 운영되고 있는 VM 의 서비스를 일시적으로나마 정지시킨 후, 서비스가 중단되는 단점이 있다.

수직 확장의 또다른 방법은 Hot-Plug 방식이다[7][8]. Hot-Plug 는 운영중인 서버의 중지 없이, Guest OS(VM 의 운영체제)의 CPU 나 Memory 를 추가시키는 방법으로, 현재 대부분의 가상화 엔진(KVM, vSphere, Xen 등)에서 모두 제공하며, 리눅스 및 윈도우즈 등의 Guest OS 에서도 대부분 지원한다. 단, 자원의 축소는 현재로서는 가능하지 않다(일부 유닉스 가상화 솔루션에서는 축소도 가능하나, CSP 에서 주로 사용되는 x86 기반에서는 제공되고 있지 않다). 표 1 에서 수평 확장과 수직 확장, 그리고 무중단 수직 확장에 대한 비교 내용을 나타내었다.

Table 1. Comparison of resource scaling methods
표 1. 자원 확장 방법 비교

	Horizonal Scaling	Vertical Scaling	Hot-plug Scaling
LB	O	X	X
Same spec.	O	X	X
Service stop	X	O	X
Scale limits	X	X	O
Scaling time	~ 5 min	~ 5 min	ms
Scale-in(down)	O	O	X

표 1 에서 보이는 것과 같이, 무중단 수직 확장 방법은 확장이 가능한 용량에 한계가 있고, 자원의 축소가 불가능하다는 단점이 있다. 반면, 1) 사용자 접속 트래픽을 분배하는 LB 가 필요 없고, 2) 자원의 세밀한 용량으로 확장이 가능하므로 동일 사양일 필요가 없으며, 3) 운영중인 서비스의 중단이 발생하지 않고, 4) 거의 실시간으로 필요한 자원을 추가함으로써 자원을 확장하는 것이 가능하다는 장점이 있다.

III. 무중단 수직 확장 서비스를 위한 VMP 알고리즘

3.1 여유 공간과 가상화율

일반적인 정보시스템이나 클라우드 자원 모두 일정 부분의 여유 용량을 비워두고 있다. 2018 정보시스템 하드웨어 규모 산정 지침에 따르면 1.3 배의 용량으로 규모를 산정하고(30% 여유 공간)[9], 또한 MS Azure 의 경우, PM 내의 20~30%의 여유 공간을 기준으로 리밸런싱을 수행한다[10]. 특히 클라우드 서버의 경우, VM 이 배치되는 PM 의 패치나 업그레이드 등을 수행하기 위해서는 서비스의 연속성 제공을 위해 PM 의 VM 을 모두 다른 PM 으로 옮기는 라이브 마이그레이션이 수행해야 하는데, 이때, 옮겨질 PM 에 여유 공간이 필요하다.

여유 공간 확보를 위해 첫째, PM 내에서의 VM 할당 공간 확보, 둘째, 랙 기준의 여유 PM 공간 확보, 셋째, 서버의 집합인 랙 단위 공간 확보 방법이다. 제각기 장단점이 존재하지만, 본 연구는 첫째 PM 내에서의 여유 공간 확보 방법을 기준으로 한다.

하이퍼바이저는 가상화 기능을 통하여 VM 을 물리 자원(PM)에 생성한다. 이 때, 물리 자원의 전체 용량보다 가상으로 생성된 VM 의 총 용량 비율을 가상화율이라고 한다. 일반적으로 CPU 는 물리 자원 용량 이상으로 VM CPU 를 생성할 수 있으나, 메모리의 경우는 물리 자원 용량을 초과하도록 구성하지 않는다. 이는 CPU 의 경우 시분할을 통하여 전체 성능을 낮추는 방식으로 물리 자원 사양 이상의 많은 VM 에 CPU 를 할당할 수 있으나, 메모리의 경우는 물리 자원 이상의 메모리를 제공할 수 없기 때문이다. 오픈스택의 경우 매뉴얼에 따르면 CPU 는 16 배, 메모리는 1.5 배로 권장하고 있으나, 서비스 제공자 입장에서는 서비스의 수준을 유지하기 위해 가상화율을 매우 낮은 수준에서 적용하고 있는 것으로 알려지고 있다. 일례로, 국가정보자원관리원의 경우는 CPU 만 300% 수준이며, AWS 는 HyperThread 만 적용하여 200% 수준이다[11][12].

결과적으로, CSP 가 내부적인 여유 공간 확보 비율은 공개적으로 알려져 있지 않지만, 위와 같은 상황을 가정하면, 20~30% 수준으로 추정해 볼 수 있다. 따라서, 이 여유 공간을 가용성 유지를 위한 임시적인 수직 확장 공간으로 사용할 수 있다.

3.2 무중단 수직 확장 서비스를 위한 VMP 알고리즘

무중단 수직 확장을 위한 PM 배치를 일반 VM 및 수직 확장용 VM 의 요청에 따라, 1) 수직 확장 VM 이 없는 일반 PM, 2) 수직 확장과 일반 VM 이 혼합되어 있는 혼합 PM, 3) 수직 확장 VM 만을 배치시키는 PM 을 지정하고, 일반 PM 으로부터 혼합 PM 또는 수직 확장 PM 으로 전환될 수 있는 방법을 제안한다.

그림 1 은 이러한 3 가지 형태의 PM 유형을 나타낸 것으로, 여유 공간(A1, A2), 일반 공간(B1, B2), 수직 확장 공간(C1, C2)을 일정한 비율로 구성하여 나타낸 것이다.

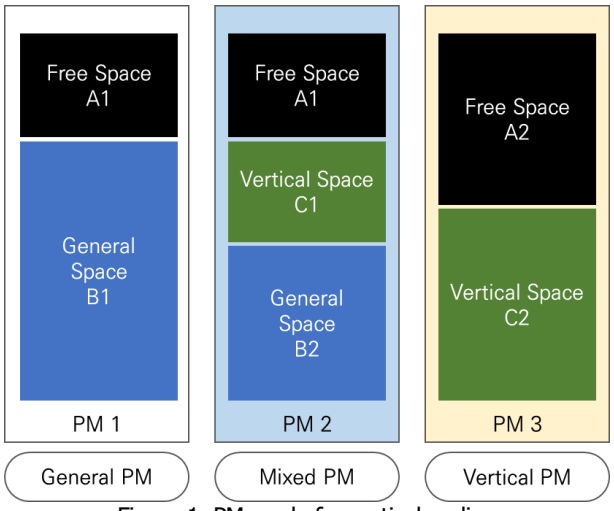

Figure 1. PM mode for vertical scaling
그림 1. 수직 확장 서비스를 위한 PM 지정 및 구성

클라우드 운영 정책에 따라 여유공간 A1 의 크기를 정하면, 나머지 공간을 일반 할당 공간(B1, B2)으로 사용하며, 혼합 PM 의 경우는 여유 공간 크기만큼을 수직 확장 할당 공간으로 지정한다. 수직 확장 PM 은 전체 공간을 여유 공간과 수직 확장 할당 공간으로 50%씩 나누어 할당한다.

이러한 구성 및 배치 전략을 제시하는 이유는 다음과 같다.

첫째, 최초 VM 생성시, 수직 확장용과 일반용으로 분리하여 요청하는 것으로 가정하였다. 이는 수직 확장형은 미려한 성능 저하가 발생하기도 하고, 모든 VM 을 수직 확장이 가능하도록 한다면, 그만큼의 여유 공간 확보해 놓아야 하기 때문이다.

둘째, 수직 확장 할당 공간을 별도로 확보하도록 함으로써, 그에 따른 여유 공간을 동일하게 유지하도록 하였는데, 이는 동시에 다수의 사용자가 수직 확장을 동시에 요청할 경우, 서비스를 수용할 수 없는 상황이 발생할 수 있기 때문이다.

셋째, 혼합 PM 과 수직 확장 PM 의 경우 수직 확장 할당 공간을 여유 공간과 동일한 비율로 설정함으로써, 수직 확장형 VM 을 동시에 최대 2 배에 한해서 수직 확장 서비스를 보장할 수 있도록 하였다.

넷째, 일반 PM 으로 시작하여 수직 확장 VM 의 요청에 따라, 현재의 PM 들에 대한 VM 할당 상태의 비율을 검사하여, 일반 PM 에서 혼합 PM 으로, 혼합 PM 에서 수직 확장 전용 PM 으로 전환할 수 있도록 하여, 가변적인 수직 확장 VM 의 요청에 적응적으로 대응할 수 있도록 하였다.

다섯째, 수직 확장 PM 을 별도로 지정하여 50:50 의 비율로 할당함으로써, 높은 비율의 수직 확장 VM 의 요청에 따른 자원 낭비를 최소화한다.

이를 구현할 수 있는 일반 VM 과 수직 확장 VM 의 요청에 따른 PM 배치 전략을 위한 의사코드를 표 2 에 보인다

Table 2. VMP algorithm pseudocode for live scale-up
표 2. 무중단 수직 확장을 VMP 알고리즘 의사코드

```
main() {
for(전체 VM) {
   if (일반 VM) {
     일반 VM 할당()
   }
   else if(수직 VM) {
     수직 VM 할당()
   }
 }
}

일반 VM 할당() {

 for(전체 PM) {
   if (혼합 PM 이 있을 경우) {
     if (할당 공간이 있을 경우)
       혼합 PM 에 일반 VM 할당
   }
 }
 for(전체 PM) {
   if (일반 PM 이 있을 경우) {
     if (할당 공간이 있을 경우)
       일반 PM 에 일반 VM 할당
   }
 }
}



수직 VM 할당() {

for(전체 PM) {
  if (혼합 PM 이 있을 경우)
     if (할당 공간이 있을 경우)
        혼합 PM 에 수직 VM 할당()
}
for(전체 PM) {
  if (수직 PM 이 있을 경우)
     if (할당 공간이 있을 경우)
        수직 PM 에 수직 VM 할당()
}
for(전체 PM) {
  if (혼합 PM)
     if (혼합 PM 에 일반 VM 이 없을 경우)
        현재 PM 을 수직 PM 으로 변경 후, 수직 VM 할당()
}
for(전체 PM) {
  if (일반 PM)
     if ( 혼합 PM 으로 전환이 가능할 경우 )
        현재 PM 을 혼합 PM 으로 변경 후, 수직 VM 할당()
}
for(전체 PM) {
  if (일반 PM)
     if ( 수직 PM 으로 전환이 가능할 경우 )
        현재 PM 을 수직 PM 으로 변경 후, 수직 VM 할당()
}
}
```

IV. VMP 시뮬레이션

4.1 실험 데이터

실험 데이터는 행정안전부의 “2022 년도 행정공공 클라우드 전환 상세 설계 사업” 의 결과 데이터를 활용하였다. 해당 데이터는 80 개 공공기관의 633 개 정보시스템을 클라우드로 전환하기 위한 상세 설계 데이터로써, 다양한 사양으로 총 1,920 여개의 VM 으로 구성되어 있다. 이러한 데이터는 가정에 의해 시뮬레이션하는 타 연구에서 와는 달리 실제 클라우드로 전환되는 정보시스템의 사양으로써 실효성이 있다고 판단한다.

해당 실험 데이터의 분포는 그림 2 와 같다. 전체 VM 은 1,920 개 이며, 전체 Core 와 메모리의 합은 각각 9,308 core, 33,398 GB 이다. 4core-8GB 와 2core-4GB 의 VM 이 970 개로, 전체 VM 의 50%를 차지하며, 또한 4GB 이하 메모리 사양의 VM 은 536 개로, 전체 VM 의 30%를 차지한다.

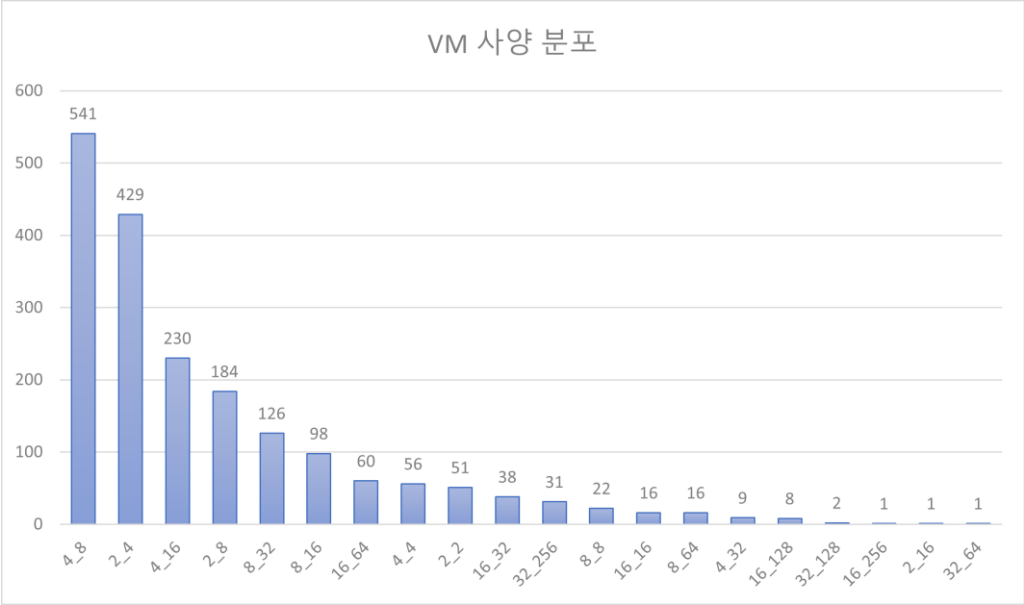

Figure 2. Distribution of VM specifications
그림 2. 실험에 사용된 VM 사양 분포도

4.2 물리 서버 사양

PM 의 사양은 2021 년도 국가정보자원관리원에서 제공하는 정보자원 기술기준에 따른 클라우드 서버의 표준 사양 중 가급에 해당하는 사양을 선정하였다[13]. 해당 서버는 36 코어(2 소켓 합)의 CPU 와 코어당 16GB 의 메모리를 요구하기 때문에, 필요한 메모리는 총 576 GB 이다. 한편, 대부분의 클라우드로 활용되는 x86 의 경우 HyperThread 설정 시, 2 배의 코어를 제공하고(가상화율 200%), 또한 가상화 솔루션에 따라 Oversubscription 이 가능하기 때문에 최종적으로 하나의 물리서버는 적용하는 가상화율에 따라 활용 가능한 용량이 차이가 있다.

앞선, 전체 요청 VM 1,920 개는 총 9,308 core 와 33,398 GB 메모리를 요구하므로, 단순 계산으로 물리 서버의 용량(36 와 576)으로 나누면 Core 로는 259 개, 메모리로는 58 개 이므로, 큰 수인 259 개의 물리 서버가 필요함을 알 수 있다.

가상화율과 여유 공간율에 따른 전체 9,308 코어를 할당하기 위해 필요한 PM 의 수는 수식 1 과 같이 계산할 수 있으며, 계산 결과는 표 3 과 같고, PM 의 개수는 산술적인 최소 수치로 볼 수 있다.

Equation 1. Equation for total number of PMs according to virtualization rate and free space rate
수식 1. 가상화율과 여유 공간율에 따른 전체 PM 수 계산식

$$PMs = \frac{total\ cores}{(1 - free\_rate) * \left(\frac{cores}{PM}\right) * virt\_rate}$$

Table 3. Minimum number of PMs according to virtualization rate and free rate
표 3. 가상화율과 여유공간률에 따른 일반 VM 의 산술적 최소 필요 PM 수

	Cores per PM	20% free_rate	25% free_rate	30% free_rate
200%	72	161	173	195
300%	108	107	115	123
400%	144	81	87	93
500%	180	65	69	74

4.3 시뮬레이션 환경

VMP 시뮬레이션을 위해 의사코드를 토대로 하여 Javascript 를 활용한 코드를 작성하였다. 해당 코드는, 일반 브라우저에서 개발자 모드로 진입하여 바로 실행할 수 있으며, 별도의 컴파일 없이 결과를 바로 확인할 수 있다. 본 시뮬레이션은 크롬 브라우저에서 수행하였다.

PM 의 여유공간률을 20%, 30% 로 지정했을 경우에 대해 각각 수행하였으며, 수직 확장 VM 의 발생 확률은 0%부터, 60%까지 10% 증가시키고, 각각 3 회식 실행하였다. 또한 가상화율을 200% 로 지정했을 경우와 300%로 지정했을 경우에 대해 동일한 방법으로 각각 진행하였다.

PM 은 의사코드에서와 같이, 최초 일반 PM 에서 시작하며, 수직 확장 VM 이 발생시, 현재 PM 의 VM 분포에 따라 가능할 경우 혼합 PM 과 수직 PM 순으로 전환되도록 하였다.

그림 3, 4 는 가상화율을 200%로 하고, 수직확장용 VM 을 요청하는 비율이 0%부터 60%까지 10% 씩 증가한 상황일 경우의 시뮬레이션 결과로써, 여유 공간을 각각 20%, 30%로 지정한 환경에서의 VM 할당에 위한 필요한 PM 의 수를 나타낸다.

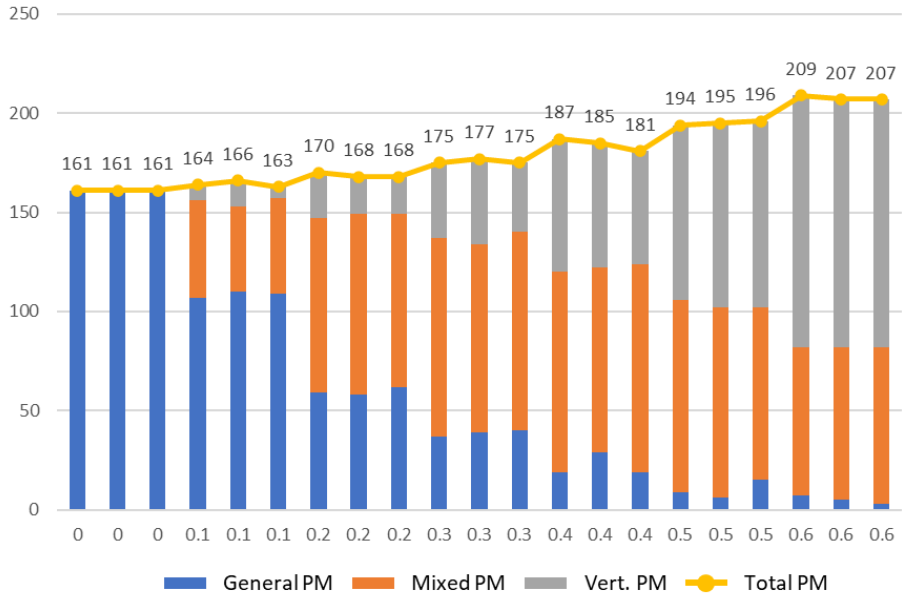

Figure 3. Number of PMs according to vertical VM rate(VR=200%, FR=20%)
그림 3. 수직확장 비율에 따른 PM 수(가상화율 200%, 여유 공간 20%)

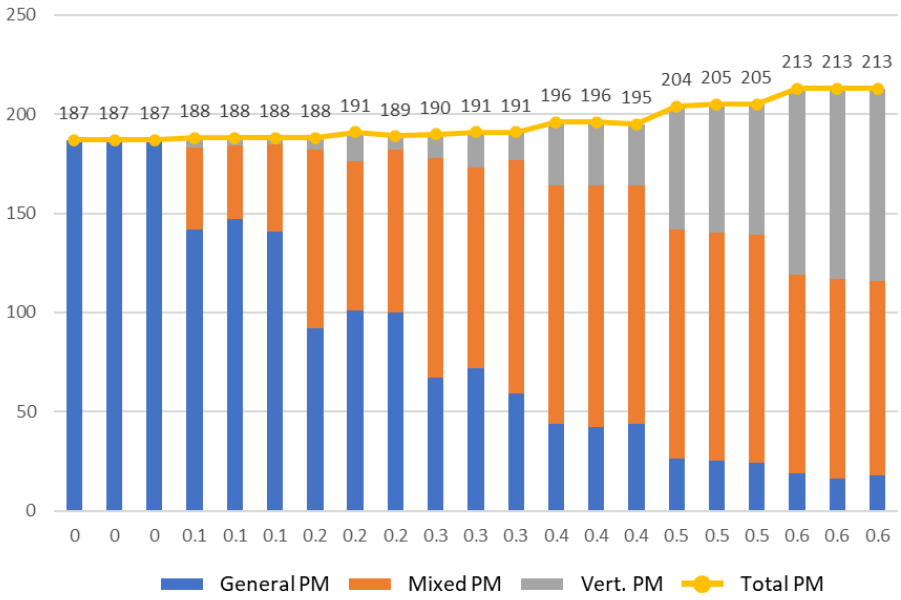

Figure 4. Number of PMs according to vertical VM rateVR=200%, FR=30%)
그림 4. 수직확장 비율에 따른 PM 수(가상화율 200%, 여유 공간 30%)

그림 5, 6 은 가상화율을 300%로 하고, 수직확장용 VM 을 요청하는 비율이 0%부터 60%까지 10% 씩 증가한 상황일 경우의 시뮬레이션 결과로써, 여유 공간을 각각 20%, 30%로 지정한 환경에서의 VM 할당에 위한 필요한 PM 의 수를 나타낸다.

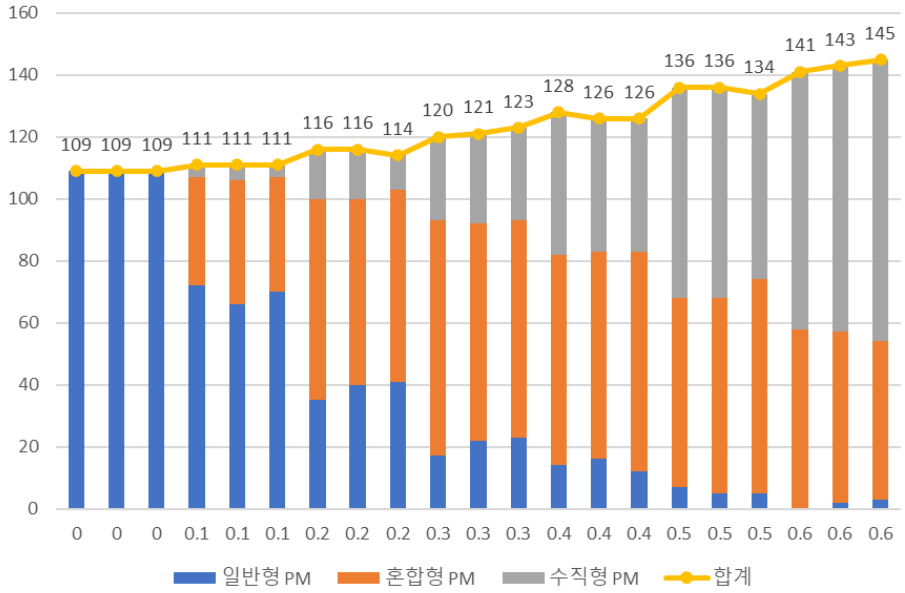

Figure 5. Number of PMs according to vertical VM rate(VR=300%, FR=20%)
그림 5. 수직확장 비율에 따른 PM 수(가상화율 300%, 여유 공간 20%)

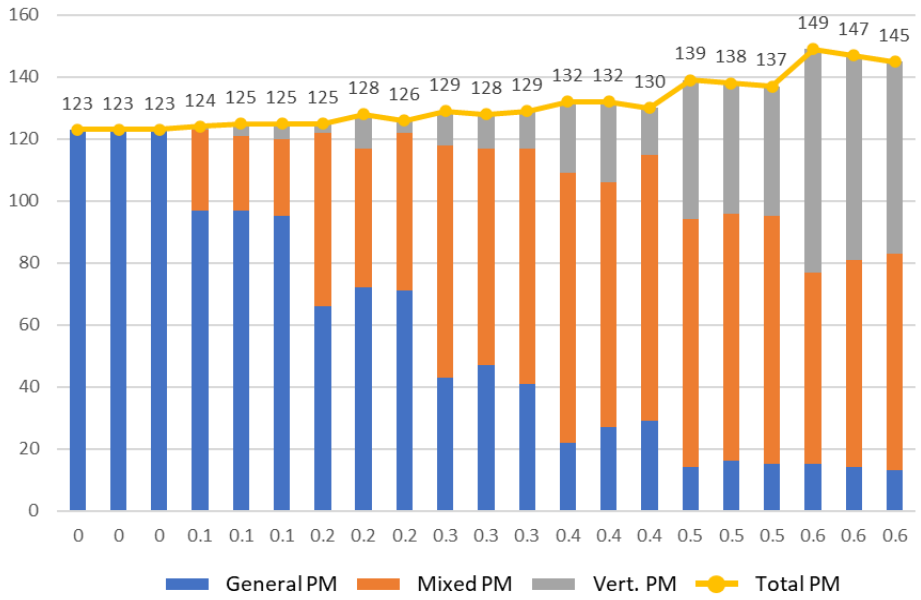

Figure 6. Number of PMs according to vertical VM rate(VR=300%, FR=30%)
그림 6. 수직확장 비율에 따른 PM 수(가상화율 300%, 여유 공간 30%)

4.4 결과 분석

4.4.1 수직 확장 VM 이 없을 경우,

수직 확장을 하지 않는 VM 만을 배치하였을 경우(수직 확장 비율 0%), 그림 3 부터 그림 6 까지의 첫 번째 3 개의 값 모두, 표 3 에서 보이는 최소 PM 수와 거의 동일한 수치를 보이는 것을 알 수 있다. 이는 기존의 VMP 연구에 대비하여 First-Fit 알고리즘만을 적용하여도 PM 활용률이 매우 높게 나타난 것으로, 기존 연구들은 PM 대비하여 VM 의 크기를 상당히 크게 잡아 실제 알고리즘의 효과를 극대화한 것으로 판단된다. 본 시뮬레이션은 실제 데이터를 근간으로 하였으며, 크기가 작은 VM 이 상당수 존재하고, 또한 현재 많이 사용되는 클라우드 서버의 사양이 상대적으로 매우 높아 First-Fit 알고리즘만으로도 VMP 가 상당히 잘되는 것을 알 수 있다.

Table 4. Number and percentage of PMs required based on percentage of scale-up VMs
표 4. 수직 확장 VM 의 비율에 따라 필요한 PM 의 개수 및 비율

ScaleUp VM rate	200-20		200-25		200-30		300-20		300-25		300-30	
	PMs	PM rate	PMs	PM rate	PM	PM rate	PM	PM rate	PM	PM rate	PM	PM rate
0%	161	100%	173	100%	187	100%	109	100%	117	100%	123	100%
10%	164	102%	175	101%	188	101%	111	102%	119	102%	125	102%
20%	168	104%	175	101%	189	101%	116	106%	122	104%	126	102%
30%	175	109%	179	103%	191	102%	121	111%	123	105%	129	105%
40%	185	115%	187	108%	196	105%	126	116%	130	111%	132	107%
50%	195	121%	198	114%	205	110%	136	125%	134	115%	138	112%
60%	207	129%	208	120%	213	114%	143	131%	144	123%	147	120%

4.4.2 수직 확장 비율에 대한 분석

표 4 는 수직 확장 VM 의 비율에 따른 필요한 PM 의 개수와 수직 확장 VM 이 없는 경우를 100%로 할 경우, 상대적 증가되는 PM 의 비율을 나타낸 것이다. 수직 확장 VM 의 비율이 높아질수록 필요한 여유 공간의 크기가 비례하여 증가하므로, 전체 요구되는 PM 의 개수도 거의 비례적으로 증가한다.

수직 확장 비율이 50% 일 경우는, 여유 공간을 고려하면 전체적으로 150%의 자원의 필요하나, 제안한 알고리즘의 시뮬레이션 결과로는 최대 125%(300-20)의 여유 공간만을 필요로 하였다. 이러한 결과는 앞 선, 수직 확장 VM 이 없을 경우와 마찬가지로 저사양의 VM 이 다수 분포됨에 따라 전체 VM 이 효과적으로 배치되는 것으로 판단된다.

4.4.3 여유 공간 비율에 대한 분석

여유 공간의 비율을 크게 가져갈수록, 수직 확장 비율이 0%일 경우에는 많은 PM 이 요구되지만, 수직 확장 비율이 높아질수록 점차 격차가 줄어들고, 50% 이상일 경우, 필요로 하는 PM 수가 거의 비슷해지는 것을 알 수 있다.

4.4.4 가상화율에 따른 분석

가상화율에 따른 요구되는 PM 의 수는 200%와 300%의 가상화율의 차이 만큼의 PM 수가 1.5 배 더 필요한 것을 확인하였으며, 가상화율이 높아도 수직 확장 비율에 대한 분석이나 여유 공간 비율에 대한 분석 결과는 유사하다.

4.4.5 수익성 분석

PM 수에 따른 비용의 증가는 가변적이고 또한 잘 알려져 있지 않기 때문에, 하나의 PM 으로 얻을 수 있는 수익에 대해 분석하였다. 하이퍼쓰레딩만을 적용하여 가상화율 200%인 72-576 용량을 갖는 하나의 서버는, NCP(Naver Cloud Platform)의 2022 년도 월단위 금액 기준으로 계산하면, 코어당 28,000 원이며 메모리는 GB 당 4,000 원 이므로[14], 가상화율 200%, 여유 공간율을 25%로 할 경우 수직 확장 서비스를 위해 추가되는 PM 수와 수직 확장 전의 PM 당 수익 및 2 배의 수직 확장 후의 수익을 계산하면, 그림 7 과 같다.

수직 확장 이전에는 여유 공간의 확보를 위해 추가적인 PM 이 필요하므로, PM 당 수익이 수직 확장 비율에 따라 점차 줄어들지만, 수직 확장이 발생하고 그 크기가 2 배로 수행된다면, 수직 확장 비율이 높아질수록 수익이 늘어남을 알 수 있다. 또한, 수직 확장이 발생한다는 것은 서비스를 위한 요구되는 자원의 용량이 초기 설정보다 크다는 것을 의미하므로, 다시 축소시키는 경우는 적을 것으로 판단하며, 이는 곧 지속적인 수익 증대를 어느 정도 보장할 것으로 기대할 수 있다.

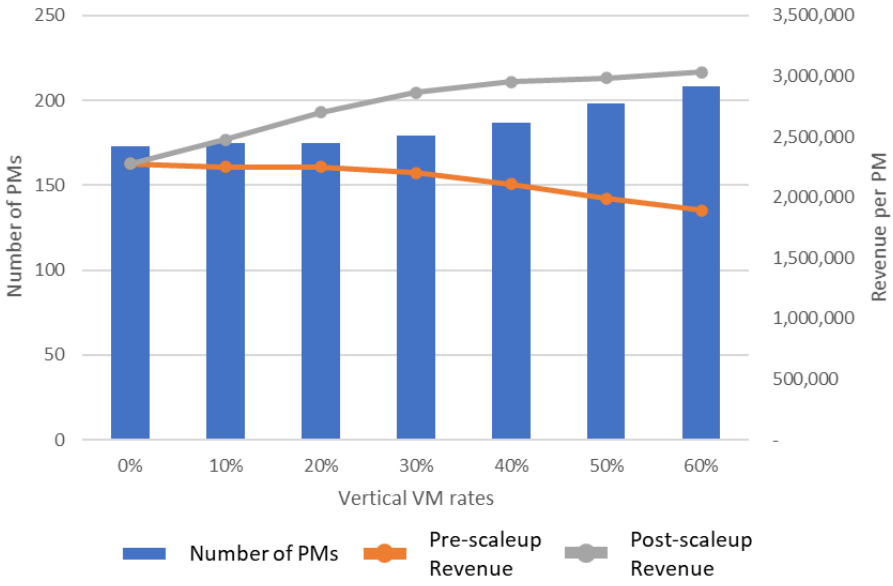

Figure 7. Revenue per PM pre and post vertical scaling
그림 7. 수직 확장 전과 후의 PM 당 수익

V. 결론

본 논문에서는 Hot-Plug 방식의 무중단 수직 확장 서비스를 위한 가상자원 배치 전략을 제시하고, 실 데이터를 기반으로 FirstFit 알고리즘 구현하여 VMP 시뮬레이션을 수행하였다. 시뮬레이션에 사용된 데이터는 633 개의 정보 시스템에 대한 총 1,920 여개의 VM 사양을 사용하였으며, 초기 용량 대비 최대 2 배까지 확장 가능한 용량으로 산정하였다.

PM 에서 설정한 가상화율과 여유 자원율에 따라 할당할 수 있는 VM 의 수가 달라지므로 200%와 300%의 가상화율에 대해 20%, 30%의 여유 자원율을 두었을 경우, 0~60%의 수직확장 VM 의 요청에 따른 요구되는 PM 의 개수를 시뮬레이션을 통하여 구하였다.

제시한 알고리즘에 의한 시뮬레이션 결과, 전체 요청 VM 중에서 수직 확장 비율이 20% 미만일 경우 추가로 필요한 PM 의 수가 미미함을 알 수 있었으며, 수직 확장 비율이 50% 일 경우 전체적으로 150%의 자원의 필요하나, 제안한 알고리즘의 시뮬레이션 결과로는 최대 125%의 여유 공간만을 필요로 하는 것으로 나타났다.

결론적으로, 수직 확장 서비스를 제공하기 위해, 제시한 바와 같은 방법으로 여유 공간을 효율적으로 사용한다면, 수직 확장이 발생하지 않을 경우에는 낮은 수준의 추가 비용이 요구되지만, 수직 확장이 발생할 경우에는 추가적인 부가 수익 창출을 기대할 수 있다.

VI. 참고문헌

[1] Abdulaziz Alashaikh, Eisa Alanazi, and Ala Al-Fuqaha. “2021. A Survey on the Use of Preferences for Virtual Machine Placement in Cloud Data Centers”, ACM Comput. Surv. 54, 5, Article 96 (May 2021), 39 pages.
[2] Nawaf Alharbe, Abeer Aljohani, Mohamed Ali Rakrouki: “A Fuzzy Grouping Genetic Algorithm for Solving a Real-World Virtual Machine Placement Problem in a Healthcare-Cloud”, Algorithms 15(4): 128 (2022)
[3] Feng Shi1 and Jingna Lin, “Virtual Machine Resource Allocation Optimization in Cloud Computing Based on Multiobjective Genetic Algorithm”, Hindawi Computational Intelligence and Neuroscience, Volume 2022
[4] Deafallah Alsadie, “Virtual Machine Placement Methods using Metaheuristic Algorithms in a Cloud Environment – A Comprehensive Review“, International Journal of Computer Science and Network Security, VOL.22 No.4, April 2022
[5] Turowski, M., Lenk, A. “Vertical Scaling Capability of OpenStack.” Service-Oriented Computing - ICSOC 2014 Workshops Lecture Notes in Computer Science, 2015, p. 351-362
[6] L. Lu et al., “Application-driven dynamic vertical scaling of virtual machines in resource pools,” 2014 IEEE Network Operations and Management Symposium (NOMS), 2014, pp. 1-9
[7] VMware, “https://docs.vmware.com/en/VMware-vSphere/7.0/vsphere-esxi-vcenter-server-703-virtual-machine-admin-guide.pdf”
[8] Linux Foundation, “https://www.linux-kvm.org/page/CPUHotPlug”
[9] TTA, “A Guideline for Hardware Sizing of Information Systems”, TTAK.KO-10.0292/R2, 2018-12-19. revision.
[10] Microsoft, “https://docs.microsoft.com/en-us/azure-stack/hci/manage/vm-load-balancing”
[11] Amazon, “https://docs.aws.amazon.com/ko_kr/AWSEC2/latest/UserGuide/cpu-options-supported-instances-values.html”
[12] TechTarget, “https://www.techtarget.com/searchcloudcomputing/blog/The-Troposphere/Amazon-does-not-oversubscribe”
[13] NIRS, “Information resource technology standards”, 2021, http://iot.nirs.go.kr
[14] https://www.ncloud.com/charge/region/ko

저자소개

박준석 (*Jun-Seok Park*)

1998 년 2 월 목원대학교 대학원 전자및컴퓨터공학 석사
2013 년 12 월 세림티에스지㈜ 클라우드 연구소 소장
2022 년 12 월 목원대학교 대학원 지능정보융합학과 박사과정

관심분야 : 클라우드, IT 융합

고대식 (*Dae-Sik Ko*)

1987 년 2 월 경희대학교 전자공학과 석사
1987 년 2 월 경희대학교 전자공학과 박사
1989 년 ~ 현재 목원대학교 전자공학과 교수

관심분야 : 디지털 트윈, 멀티미디어통신, 클라우드컴퓨팅