(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0063988
(43) 공개일자 2021년06월02일

(51) 국제특허분류(Int. Cl.)
G06F 9/50 (2018.01) *G06N 5/04* (2006.01)
(52) CPC특허분류
G06F 9/5077 (2013.01)
G06F 9/5072 (2013.01)
(21) 출원번호 10-2019-0152768
(22) 출원일자 2019년11월25일
심사청구일자 2021년02월01일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
허의남
경기도 용인시 기흥구 이현로29번길 86-21 e편한세상대림아파트 201동 401호
손아영
경상북도 경주시 안강읍 화전길 43-6 전원빌라 303호
(74) 대리인
특허법인지원

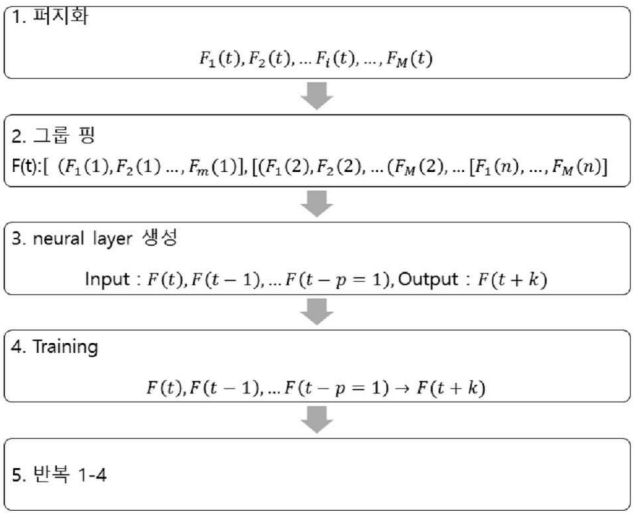
전체 청구항 수 : 총 3 항

(54) 발명의 명칭 **분산 클라우드 환경에서의 지능형 서비스 배치 방법**

(57) 요 약

본 발명은 분산 클라우드 환경에서의 지능형 서비스 배치 방법에 관한 것이다. 본 방법은 서비스 배치시 서비스 배치결정->타겟머신 선정 과정을 실시한다. 배치 결정에 앞서 가상머신(서비스)자원 상태에 따라 결정하게 되는데 서비스를 상태를 예측하여 배치를 결정하고자 한다. 자원 상태를 예측하기 위해 퍼지 시스템에서 실시되며, 이때 러닝 기반 퍼지시스템에서 자원을 예측한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 5/048 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711065146
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기술진흥센터
연구사업명 SW컴퓨팅산업원천기술개발사업
연구과제명 (대학ICT기초연구실) 서비스 이동 지원을 위한 분산형 클라우드 핵심원천기술 연구
기 여 율 1/1
과제수행기관명 경희대학교 산학협력단
연구기간 2019.01.01 ~ 2019.12.31

청구범위유예 : 있음

명 세 서

청구범위

청구항 1

클라우드 환경에서의 서비스 배치를 수행하는 단계; 및

상기 서비스 배치에 따른 자원 관리를 수행하는 단계;

를 포함하는,

분산 클라우드 환경에서의 지능형 서비스 배치 방법.

청구항 2

제1항에 있어서,

상기 서비스 배치에 대한 초기 규칙(rule)을 설정하는 단계; 및

상기 설정된 초기 규칙에 따라 규칙을 조정하는 단계;

를 더 포함하는,

분산 클라우드 환경에서의 지능형 서비스 배치 방법.

청구항 3

제1항에 있어서,

상기 자원 관리에 따라 클라우드 서비스를 제공하는 단계;

를 더 포함하는,

분산 클라우드 환경에서의 지능형 서비스 배치 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 분산 클라우드 환경에서의 지능형 서비스 배치 방법에 관한 것으로, 더욱 상세하게는 러닝 기반 퍼지 시스템에서 자원을 예측에 관한 분산 클라우드 환경에서의 지능형 서비스 배치 방법에 관한 것이다.

배 경 기 술

[0003] 클라우드 컴퓨팅(Cloud Computing)이란 정보처리를 자신의 컴퓨터가 아닌 인터넷으로 연결된 다른 컴퓨터로 처리하는 기술을 말한다. 우리가 사용하고 있는 개인용 컴퓨터(PC)에는 필요에 따라 구매한 소프트웨어가 설치되어 있고 동영상과 문서와 같은 데이터도 저장되어 있다. 문서를 작성하려면 자신의 컴퓨터에 저장되어 있는 문서 작성 프로그램을 구동시켜야 한다. 그러나 클라우드 컴퓨팅은 프로그램과 문서를 다른 곳에 저장해 놓고 내 컴퓨터로 인터넷을 통해 프로그램과 문서가 저장된 곳에 접속해서 이용하는 방식이다.

[0004] 따라서, 필요한 소프트웨어를 내 컴퓨터에 설치할 필요도 없고, 또 주기적으로 업데이트 하지 않아도 되며, 회사 컴퓨터에서 작업을 하던 문서를 따로 저장해서 집으로 가져갈 필요도 없다. 또한 자신의 컴퓨터가 고장을 일으켜도 데이터가 손상될 염려도 없고, 필요한 만큼 쓰고 비용을 지불하면 되므로 사용 빈도가 낮은 소프트웨어

를 비싸게 구입할 필요도 없고, 터무니없이 큰 저장장치를 갖추지 않아도 된다.

[0005] 이러한 클라우드 컴퓨팅 기술에 기반한 응용이 점차 대중화되면서 안정적인 성능 보장이 필요해졌다. 특히, 클라우드의 자원 공유 특성에도 불구하고 CPU, 메모리, 스토리지와 같은 대부분의 클라우드 자원들은 다른 사용자들과 비교적 잘 구분되어 사용되고 있지만, 네트워크 성능에서의 변동으로 인하여 클라우드 기반 응용들의 성능에 대한 변동폭 또한 크다는 문제점이 발생했다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) [특허문헌 1] 한국등록특허 제10-1613513호. 2016.04.12. 등록.

발명의 내용

해결하려는 과제

[0008] 본 발명은 전술한 문제점을 해결하기 위하여 창출된 것으로, 분산 클라우드 환경에서의 지능형 서비스 배치 방법을 제공하는 것을 그 목적으로 한다.

[0009] 본 발명의 목적들은 이상에서 언급한 목적들로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0011] 상기한 목적들을 달성하기 위하여, 본 발명의 일 실시예에 따른 분산 클라우드 환경에서의 지능형 서비스 배치 방법이 개시된다. 상기 방법은 서비스 배치시 서비스 배치결정->타겟머신 선정 과정을 실시한다. 배치 결정에 앞서 가상머신(서비스)자원 상태에 따라 결정하게 되는데 서비스를 상태를 예측하여 배치를 결정하고자 한다. 자원 상태를 예측하기 위해 퍼지 시스템에서 실시되며, 이때 러닝 기반 퍼지시스템에서 자원을 예측할 수 있다.

[0012] 상기한 목적들을 달성하기 위한 구체적인 사항들은 첨부된 도면과 함께 상세하게 후술될 실시예들을 참조하면 명확해질 것이다.

[0013] 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라, 서로 다른 다양한 형태로 구성될 수 있으며, 본 발명의 개시가 완전하도록 하고 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자(이하, "통상의 기술자")에게 발명의 범주를 완전하게 알려주기 위해서 제공되는 것이다.

발명의 효과

[0015] 본 발명의 일 실시예에 의하면, 데이터 센터는 더 작은 자원을 사용하면서 QoS를 보장할 수있다.

[0016] 또한, 본 발명의 일 실시예에 의하면, 분산 클라우드 환경에서 데이터 센터내에서 성능을 향상시키고 효율을 높일 수 있다.

[0017] 본 발명의 효과들은 상술된 효과들로 제한되지 않으며, 본 발명의 기술적 특징들에 의하여 기대되는 잠정적인 효과들은 아래의 기재로부터 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0019] 상기 언급된 본 발명 내용의 특징들이 상세하게, 보다 구체화된 설명으로, 이하의 실시예들을 참조하여 이해될 수 있도록, 실시예들 중 일부는 첨부되는 도면에서 도시된다. 또한, 도면과의 유사한 참조번호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭하는 것으로 의도된다. 그러나 첨부된 도면들은 단지 본 발명 내용의 특

정한 전형적인 실시예들만을 도시하는 것일 뿐, 본 발명의 범위를 한정하는 것으로 고려되지는 않으며, 동일한 효과를 갖는 다른 실시예들이 충분히 인식될 수 있다는 점을 유의하도록 한다.

도 1은 본 발명의 일 실시예에 따른 러닝 기반 퍼지시스템에서 자원을 예측의 실시도를 도시한 도면이다.

도 2는 본 발명의 일 실시예에 따른 러닝기반의 서비스 배치 실시 과정을 도시한 도면이다 .

도 3은 본 발명의 일 실시예에 따른 러닝을 통해 자원을 예측하는 룰 설정을 통해 계산 배치가 실시되는 과정을 도시한 도면이다.

도 4는 본 발명의 일 실시예에 따른 정확도를 도시한 도면이다.

도 5는 본 발명의 일 실시예에 따른 에너지소비량의 결과를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 본 발명은 다양한 변경을 가할 수 있고, 여러 가지 실시예들을 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 이를 상세히 설명하고자 한다.

[0021] 청구범위에 개시된 발명의 다양한 특징들은 도면 및 상세한 설명을 고려하여 더 잘 이해될 수 있을 것이다. 명세서에 개시된 장치, 방법, 제법 및 다양한 실시예들은 예시를 위해서 제공되는 것이다. 개시된 구조 및 기능상의 특징들은 통상의 기술자로 하여금 다양한 실시예들을 구체적으로 실시할 수 있도록 하기 위한 것이고, 발명의 범위를 제한하기 위한 것이 아니다. 개시된 용어 및 문장들은 개시된 발명의 다양한 특징들을 이해하기 쉽게 설명하기 위한 것이고, 발명의 범위를 제한하기 위한 것이 아니다.

[0022] 본 발명을 설명함에 있어서, 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우, 그 상세한 설명을 생략한다.

[0023] 이하, 본 발명의 일 실시예에 따른 분산 클라우드 환경에서의 지능형 서비스 배치 방법을 설명한다.

[0024] 도 1은 본 발명의 일 실시예에 따른 러닝 기반 퍼지시스템에서 자원을 예측의 실시도를 도시한 도면이고, 도 2는 본 발명의 일 실시예에 따른 러닝기반의 서비스 배치 실시 과정을 도시한 도면이며, 도 3은 본 발명의 일 실시예에 따른 러닝을 통해 자원을 예측하는 룰 설정을 통해 계산 배치가 실시되는 과정을 도시한 도면이다.

[0025] 또한, 도 4는 본 발명의 일 실시예에 따른 정확도를 도시한 도면이고, 도 5는 본 발명의 일 실시예에 따른 에너지소비량의 결과를 도시한 도면이다.

[0026] 도 1 내지 도 5를 참조하면, 본 발명은 분산되어 있는 클라우드 환경에서 특 서비스를 받던 사용자가 다른 지역으로 이동하더라도 지리적으로 가까운 새로운 마이크로 데이터센터로부터 기존의 서비스를 연속적으로 받을 수 있는 서비스 배치 방법을 도출하고 자원을 효율적으로 쓰는 것이 목적을 한다. 또한 서비스 배치 시 배치 상황을 결정하고 서비스에 따라 다양한 요소들을 고려하여 서비스의 목적에 따라 성능을 보장하고자 한다.

[0027] 또한, 기존의 연구들은 전체 서비스 배치 방법들은 시간을 줄이고 있었으나 여전히 오버헤드가 발생하고 전체 배치시간은 시간은 여전히 느리다. 네트워크 지연시간이나 CPU 작업을 동시에 하면서 네트워크 입출력을 할 경우, CPU 사용량이 많아질수록 오버헤드에 따른 지연 시간이 있었다. 성능을 예측하고 서비스 배치 효율을 향상시켜 오버헤드를 개선하고자한다.

[0028] 또한, 한 두 가지의 요소만 고려되고 있어 전체적으로 시스템의 부하가 큰 상황에서는 응용 프로그램의 성능을 정확하게 예측하지 못하고 있으며, 동적으로 자원을 할당하지 못한다는 문제점도 가지고 있다. 가상머신의 CPU, 메모리, 네트워크 자원요구를 서비스에 적합하게 동적으로 할당하는 연구가 필요하다. 물리적 머신은 할당된 가상머신 들의 시간에 따라 요구되는 자원을 충분히 보유하지 못하면 성능이 저하되는 Hotspot 문제가 발생하기 때문이다. 서비스 배치시점을 결정하고 서비스 배치 대상을 결정하여 자원을 효율적으로 사용하고자 한다.

[0029] 또한, 분산된 클라우드 환경에서 서비스 배치시 기존 기술들은 대기시간만 고려하거나 실행시간에 기초하여 메모리 사용량만 고려 할뿐 함께 서비스의 목적에 따라 고려되고 있지 않다.

[0030] 서비스 배치 방식을 활용하여 끊김 없는 서비스 제공을 받을 수 있도록 가상머신 할당이 필요한 시점에 수행하려는 워크로드의 특성을 파악하여 가장 효율적으로 수행 시키는 것이 목적이다. 또한 서비스에 따라 가상머신을 성능을 보장하고자 한다.

[0031] 또한 성능을 예측하고 가상머신 상태를 예측하여 서비스 배치 상황 및 배치 머신을 결정 하고자 한다. 지능형

알고리즘을 통해 성능을 예측하고 자원을 충분히 보유하여 QoS를 보장하고자 한다.

[0032] 따라서, 서비스 배치시 서비스 배치결정->타겟머신 선정 과정을 실시한다. 배치 결정에 앞서 가상머신(서비스) 자원 상태에 따라 결정하게 되는데 서비스를 상태를 예측하여 배치를 결정하고자 한다. 자원 상태를 예측하기 위해 퍼지 시스템에서 실시되며, 이때 러닝 기반 퍼지시스템에서 자원을 예측한다.

[0033] 또한, 클라우드 서비스 제공자는 서비스에 따라 자원을 관리할 수 있으며 서비스에 따라 적은 자원으로 성능을 보장 가능 하다. 따라서 클라우드 서비스를 제공하거나 사용하는 기업에서 기반 기술로 활용 가능하다.

[0034] 또한, 클라우드 환경에서의 서비스 배치를 활용하여 자원을 효율적으로 활용하고 성능을 향상시킬 수 있으며 분산 클라우드 환경에서의 서비스를 보장하여 프로바이더의 이익까지 높여줄 수 있다.

[0035]
$$\text{Initial}: F(t) = [F_1(t), F_2(t), \ldots . F_M(t)], t = 1,2 \ldots n$$
$$If\ (F_1(t+k), F_2(t+k), \ldots, F_M(t+k))$$
$$[\ (F_1(t), F_2(t) \ldots, F_m(t)), \ldots, (F_1(t-k), F_2(t-k), \ldots (F_M(t-k)] \qquad (p > k)$$

[0036]
$$\text{After}: (F_1(t), F_2(t) \ldots, F_m(t)), \ldots, (F_1(t-p-1), F_2(t-p-1)) \rightarrow (F_1(t+k), F_2(t+k) \ldots, F_m(t+k))$$

[0037] 상기의 수식은 초기 룰을 설정하고 예측에 따라 룰을 조정하는 과정을 수식으로 표현한 것이다.

표 1

[0038]

변수	설명
r	resource
r violation	서비스 배치에 따른 sla violation
r decision	서비스 배치 결정
t	time
s	서비스 배치 coefficeint
l	time length
r actual	실제 리소스량
r allloc	할당된 리소스 량

[0039] 표 1은 상기 수식에 대한 변수에 대한 설명이다.

[0040] 또한, 결정시 정확도는 높이고, SLA Violation을 낮추기 위해 아래와 같이 수식으로 표현할 수 있다.

[0041]
$$r_{violation}\ (t) = \max\{0, r_i^{actual}(t) - r_i^{alloc}(t)\}$$

[0042]
$$r_{decision}(t+1) = s * r_i^{pred}(t+1) + \frac{1}{l} \sum_{z=t-l+1}^{t} r_{violation}$$

[0044] 또한, S를 통해 예측 error를 감소시키고, $\frac{1}{l} \sum_{z=t-l+1}^{t} r_{violation}$ 를 통해 성능 저하를 감소시키는 역할을 한다.

[0045] 실험은 아래의 표와 같은 VM type 별로 성능 테스트를 하였으며, 실험결과 다른 기법대비 결과를 도 4 내지 도 5에서 확인할 수 있다.

표 2

[0046]

TYPE	CPU spec
High	2500 MIPS, 0.87 GB
Medium	2000 MIPS, 1.7 GB
Small	1250 MIPS, 1.7 GB

Micro	500 MIPS, 613 MB

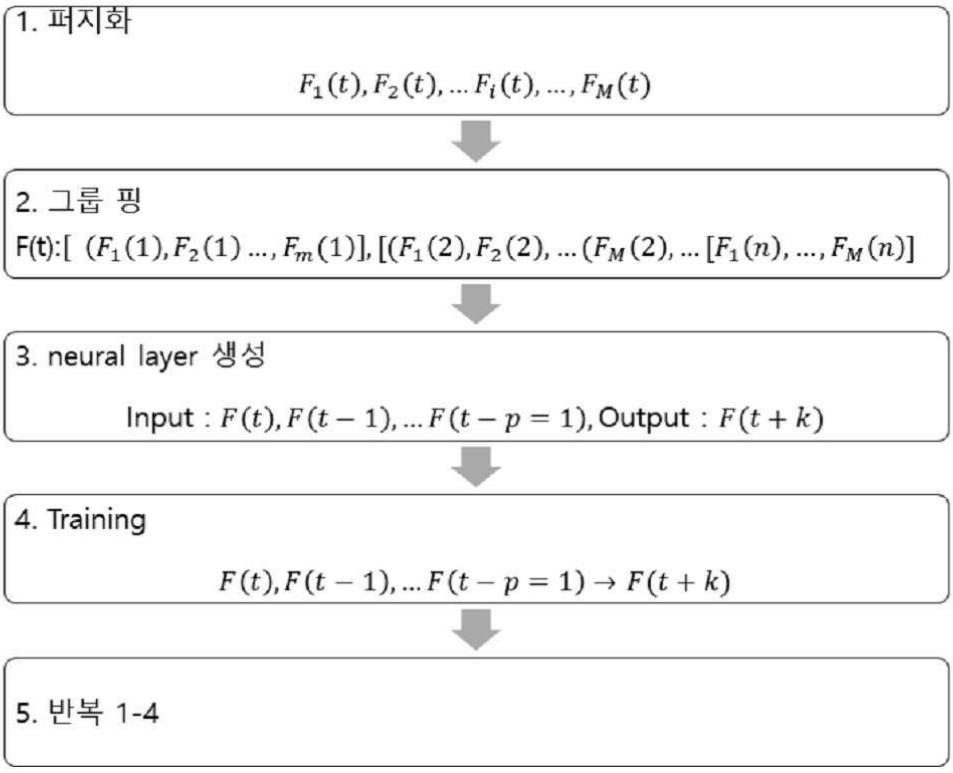
[0047] 이상의 설명은 본 발명의 기술적 사상을 예시적으로 설명한 것에 불과한 것으로, 통상의 기술자라면 본 발명의 본질적인 특성이 벗어나지 않는 범위에서 다양한 변경 및 수정이 가능할 것이다.

[0048] 따라서, 본 명세서에 개시된 실시예들은 본 발명의 기술적 사상을 한정하기 위한 것이 아니라, 설명하기 위한 것이고, 이러한 실시예들에 의하여 본 발명의 범위가 한정되는 것은 아니다.

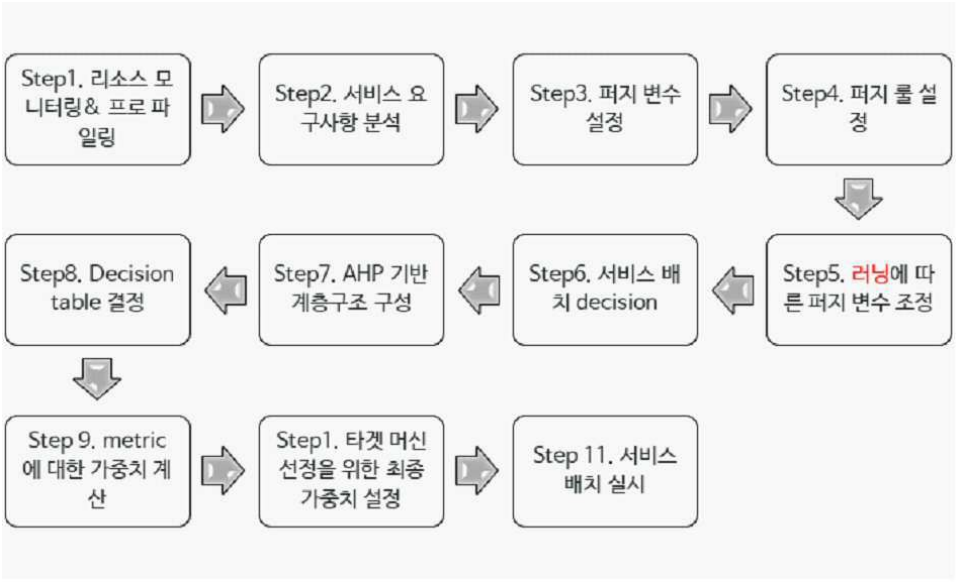
[0049] 본 발명의 보호범위는 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 이해되어야 한다.

도면

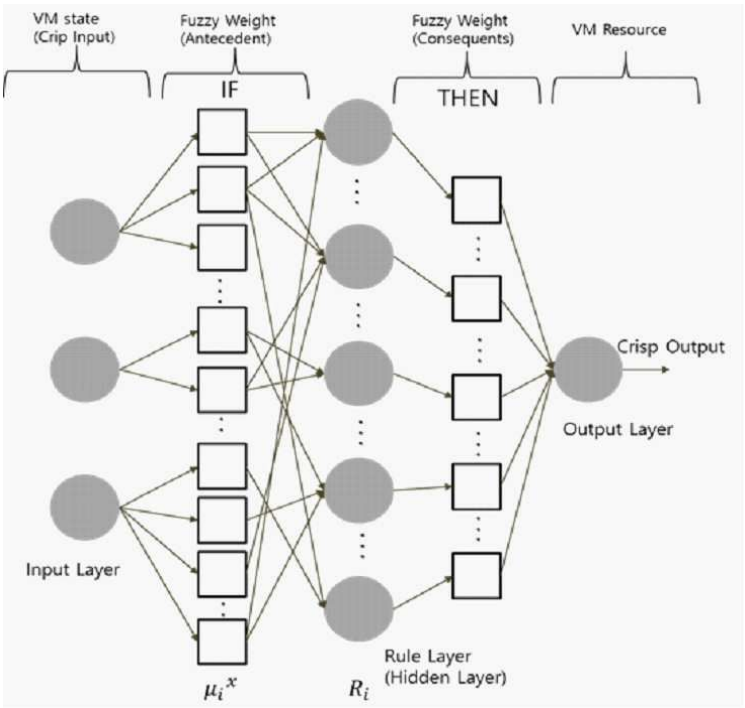
도면1



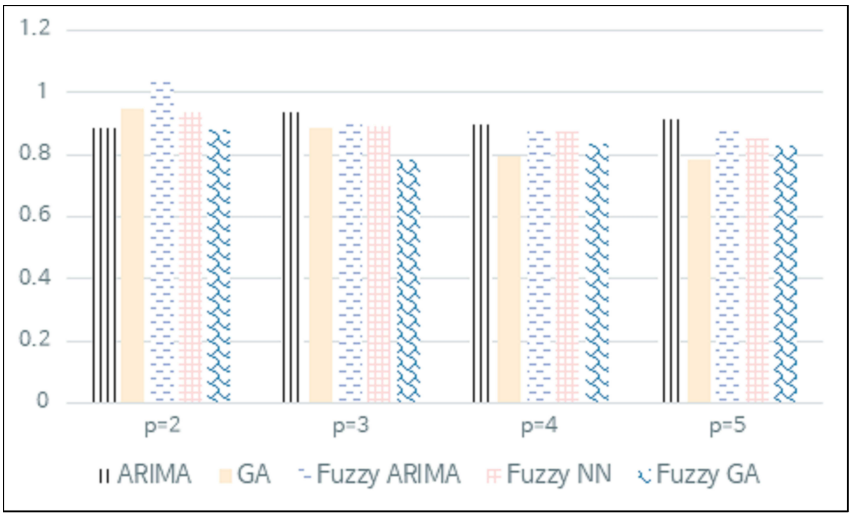
도면2


Step1. 리소스 모니터링& 프로파일링
Step2. 서비스 요구사항 분석
Step3. 퍼지 변수 설정
Step4. 퍼지 룰 설정
Step5. 러닝에 따른 퍼지 변수 조정
Step6. 서비스 배치 decision
Step7. AHP 기반 계층구조 구성
Step8. Decision table 결정
Step 9. metric에 대한 가중치 계산
Step1. 타겟 머신 선정을 위한 최종 가중치 설정
Step 11. 서비스 배치 실시

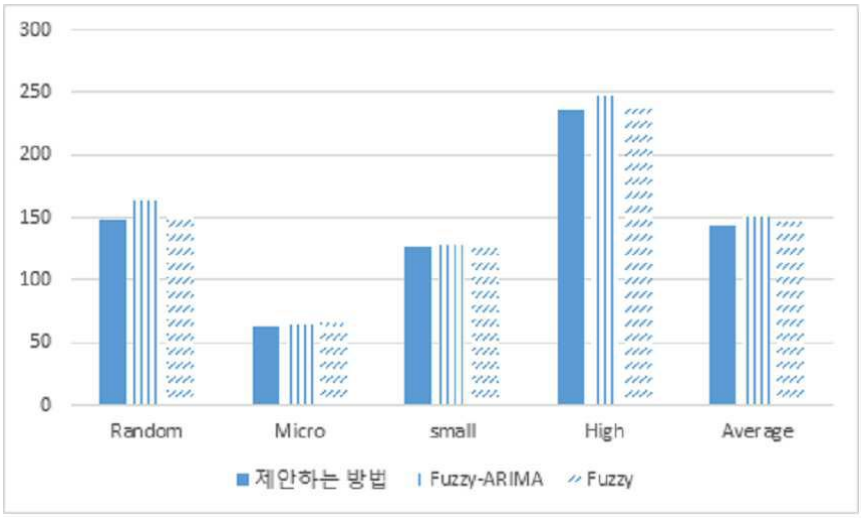
도면3


VM state
(Crip Input)
Fuzzy Weight
(Antecedent)
IF
Fuzzy Weight
(Consequents)
THEN
VM Resource
Crisp Output
Output Layer
Input Layer
$\mu_i{}^x$
R_i
Rule Layer
(Hidden Layer)

도면4


1.2
1
0.8
0.6
0.4
0.2
0
p=2
p=3
p=4
p=5
ARIMA
GA
Fuzzy ARIMA
Fuzzy NN
Fuzzy GA

도면5


300
250
200
150
100
50
0
Random
Micro
small
High
Average
제안하는 방법
Fuzzy-ARIMA
Fuzzy