(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0070929
(43) 공개일자 2015년06월25일

(51) 국제특허분류(Int. Cl.)
G06F 9/50 (2006.01) *G06F 15/16* (2006.01)
(21) 출원번호 10-2014-0072360
(22) 출원일자 2014년06월13일
심사청구일자 2014년06월13일
(30) 우선권주장
1020130157161 2013년12월17일 대한민국(KR)

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
허의남
경기도 용인시 기흥구 사은로126번길 46, 316동 1802호 (보라동, 현대모닝사이드1차아파트)
손아영
서울특별시 동대문구 경희대로 26, 전자정보대학 331호 (회기동)
(74) 대리인
특허법인 신지

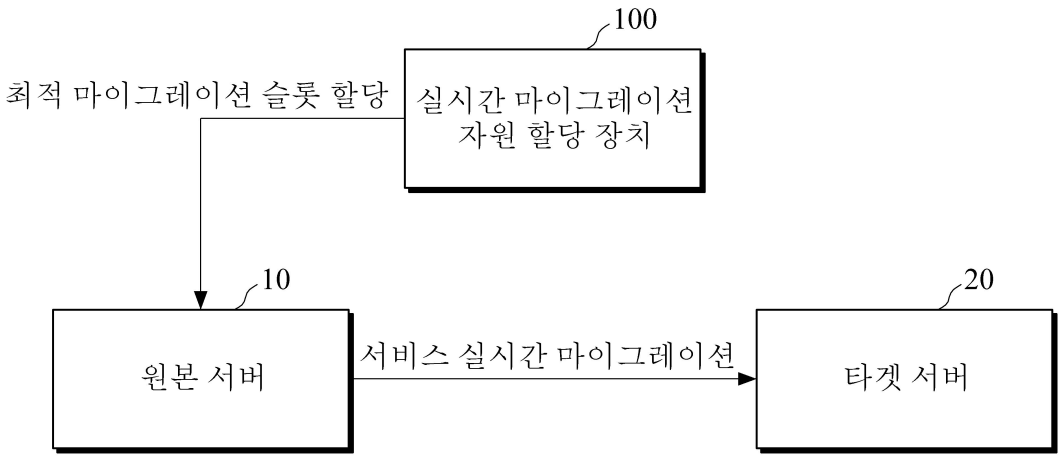
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **실시간 마이그레이션 자원 할당 시스템 및 할당 방법**

(57) 요 약

본 발명에 따른 실시간 마이그레이션 자원 할당 시스템은 하나 이상의 가상 단말을 할당하고, 마이그레이션된 서비스를 가상 단말에서 제공하는 타겟 서버, 하나 이상의 가상 단말을 할당하고, 마이그레이션 슬롯이 할당되며, 할당된 마이그레이션 슬롯을 통해 가상 단말에서 제공하는 상기 서비스를 상기 타겟 서버로 마이그레이션하는 원본 서버 및 거절 가능성 및 대기 시간을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정하고, 결정된 마이그레이션 슬롯 개수에 기초하여 원본 서버에 마이그레이션 슬롯을 할당하는 실시간 마이그레이션 슬롯 할당 장치를 포함한다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업
과제고유번호 17110-01260
부처명 미래창조과학부
연구관리전문기관 한국정보통신산업진흥원
연구사업명 대학IT 연구센터육성 지원사업
연구과제명 실시간 모바일 클라우드 서비스 플랫폼 연구개발
기 여 율 1/1
주관기관 경희대학교 산학협력단
연구기간 2013.06.01 ~ 2013.12.31

명 세 서

청구범위

청구항 1

하나 이상의 가상 단말을 할당하고, 마이그레이션된 서비스를 가상 단말에서 제공하는 타겟 서버;

하나 이상의 가상 단말을 할당하고, 마이그레이션 슬롯이 할당되며, 상기 할당된 마이그레이션 슬롯을 통해 가상 단말에서 제공하는 상기 서비스를 상기 타겟 서버로 마이그레이션하는 원본 서버; 및

거절 가능성 및 대기 시간을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정하고, 상기 결정된 마이그레이션 슬롯 개수에 기초하여 상기 원본 서버에 마이그레이션 슬롯을 할당하는 실시간 마이그레이션 슬롯 할당 장치;

를 포함하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 2

제1항에 있어서,

상기 실시간 마이그레이션 슬롯 할당 장치는,

상기 거절 가능성의 임계값(Threshold)를 설정하고, 상기 설정된 거절 가능성의 임계값 범위 내에서 상기 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 3

제1항에 있어서,

상기 실시간 마이그레이션 슬롯 할당 장치는,

서비스 전이에 따른 각 상태의 거절 가능성의 합산을 통해 거절 가능성을 산출하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 4

제3항에 있어서,

상기 거절 가능성은,

$P_r(R) = \sum_{j=0}^{K} P_{N,j}$ 에 의해 산출되며,

$P_r(R)$ 는 전체 거절 가능성을 나타내고, $P_{N,j}$ 은 가용 전체 슬롯 개수가 N 일 때, 마이그레이션 개수가 K인 경우의 거절 가능성을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 5

제1항에 있어서,

상기 실시간 마이그레이션 슬롯 할당 장치는,

동시에 하나의 마이그레이션 슬롯만 허용된 제1 케이스 및 동시에 둘 이상의 마이그레이션 슬롯이 허용된 제2 케이스 각각에 대해 대기 시간을 산출하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 6

제5항에 있어서,

상기 제1 케이스의 대기 시간은

$$T_{i,K}^{D}=\begin{cases}\dfrac{\lambda_{\mathrm{m}}-K\mu_{\mathrm{m}}}{\lambda_{\mathrm{m}}\mu_{\mathrm{m}}}, & 0\le K\le \rho_{\mathrm{m}},\\ 0, & \rho_{\mathrm{m}}<K\le N\end{cases}$$ 및

$\rho_m=\lambda_m/\mu_m$ 에 의해 산출되며,

상기 λ_m 은 마이그레이션 서비스 발생 비율을 나타내고, 상기 μ_m 은 평균 마이그레이션 시간을 나타내며, 상기 ρ_m 는 서비스 이용률을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 7

제5항에 있어서,

상기 제2 케이스의 대기 시간은

$$T_{N,j}^{D}=e^{\mu_s(\lambda_m-\mu_m)/\mu_m\lambda_m}\left(\frac{\lambda_m-j\mu_m}{\mu_m\lambda_m}\right)+\left(1-e^{-\mu_s(\lambda_m-\mu_m)/\mu_m\lambda_m}\right)\left(\frac{1}{\lambda_s}+\frac{1}{\mu_s}-\frac{j+1}{\lambda_m}\right)$$

에 의해 산출되며,

상기 λ_m 은 마이그레이션 서비스 발생 비율을 나타내고, 상기 μ_m 은 평균 마이그레이션 시간을 나타내며, 상기 μ_s 는 평균 서비스 시간을 나타내고, 상기 λ_s 는 서비스 비율을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 8

제5항에 있어서,

상기 실시간 마이그레이션 슬롯 할당 장치는,

상기 거절 가능성의 임계값 범위 내에서 상기 제1 케이스 및 상기 제2 케이스의 대기 시간 및 거절 가능성을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수(K)를 산출하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 9

제8항에 있어서,

상기 마이그레이션 슬롯 개수(K)는

$$\min_{K}\quad \sum_{i=2K+1}^{N} P_{i,K}T_{i,K}^{D}+\sum_{j=0}^{K-1}P_{N,j}T_{N,j}^{D}$$

$$\text{s.t.}\quad \sum_{i=0}^{K}P_{N,i}\le Th_R.$$

를 통해 산출되며,

상기 $P_{i,K}$ 는 제1 케이스의 각 상태의 거절 가능성을 나타내고, $T_{i,K}^{D}$는 제1 케이스의 대기 시간을 나타내며,

$P_{N,j}$ 는 제2 케이스의 각 상태의 거절 가능성을 나타내고, 상기 $T_{N,j}^{D}$는 제2 케이스의 대기 시간을 나타내며, 상기 $P_{N,i}$ 은 거절 가능성을 나타내고, 상기 Th_R 은 거절 가능성의 임계값을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 시스템.

청구항 10

거절 가능성을 산출하는 단계;

대기 시간을 산출하는 단계;

상기 거절 가능성의 임계값을 설정하는 단계;

상기 대기 시간 및 상기 거절 가능성을 고려하여 예측 대기 시간을 산출하는 단계;

상기 예측 대기 시간을 최소화하는 최적 마이그레이션 슬롯 개수를 설정하는 단계; 및

상기 설정된 최적 마이그레이션 슬롯 개수에 기초하여 마이그레이션 슬롯을 할당하는 단계;

를 포함하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 11

제10항에 있어서,

상기 예측 대기 시간을 최소화하는 최적 마이그레이션 슬롯 개수를 설정하는 단계는,

거절 가능성의 임계값(Threshold)를 설정하고, 상기 설정된 거절 가능성의 임계값 범위 내에서 상기 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 12

제10항에 있어서,

상기 거절 가능성을 산출하는 단계는,,

서비스 전이에 따른 각 상태의 거절 가능성의 합산을 통해 거절 가능성을 산출하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 13

제12항에 있어서,

상기 거절 가능성은,

$$P_r(R) = \sum_{j=0}^{K} P_{N,j}$$ 에 의해 산출되며,

$P_r(R)$ 는 전체 거절 가능성을 나타내고, $P_{N,j}$ 은 가용 전체 슬롯 개수가 N이며 마이그레이션 개수가 j인 경우의 거절 가능성을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 14

제10항에 있어서,

상기 대기 시간을 산출하는 단계는,,

동시에 하나의 마이그레이션 슬롯만 허용된 제1 케이스 및 동시에 둘 이상의 마이그레이션 슬롯이 허용된 제2 케이스 각각에 대해 대기 시간을 산출하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 15

제14항에 있어서,

상기 제1 케이스의 대기 시간은

$$T_{i,K}^{D}=\begin{cases}\dfrac{\lambda_{\mathrm{m}}-K\mu_{\mathrm{m}}}{\lambda_{\mathrm{m}}\mu_{\mathrm{m}}}, & 0\le K\le\rho_{\mathrm{m}},\\ 0, & \rho_{\mathrm{m}}<K\le N\end{cases}$$ 및

$\rho_m=\lambda_m/\mu_m$ 에 의해 산출되며,

λ_m 은 마이그레이션 서비스 발생 비율을 나타내고, μ_m 은 평균 마이그레이션 시간을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 16

제14항에 있어서,

상기 제2 케이스의 대기 시간은

$$T_{N,j}^{D}=e^{\mu_s(\lambda_m-\mu_m)/\mu_m\lambda_m}\left(\frac{\lambda_m-j\mu_m}{\mu_m\lambda_m}\right)+\left(1-e^{-\mu_s(\lambda_m-\mu_m)/\mu_m\lambda_m}\right)\left(\frac{1}{\lambda_s}+\frac{1}{\mu_s}-\frac{j+1}{\lambda_m}\right)$$

에 의해 산출되며,

상기 λ_m 은 마이그레이션 서비스 발생 비율을 나타내고, 상기 μ_m 은 평균 마이그레이션 시간을 나타내며, 상기 μ_s 는 평균 서비스 시간을 나타내고, 상기 λ_s 는 서비스 비율을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 17

제14항에 있어서,

상기 예측 대기 시간을 최소화하는 최적 마이그레이션 슬롯 개수를 설정하는 단계는 상기 거절 가능성의 임계값 범위 내에서 상기 제1 케이스 및 상기 제2 케이스의 대기 시간 및 거절 가능성을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수(K)를 산출하는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

청구항 18

제17항에 있어서,

상기 마이그레이션 슬롯 개수(K)는

$$\min_{K}\quad\sum_{i=2K+1}^{N}P_{i,K}T_{i,K}^{D}+\sum_{j=0}^{K-1}P_{N,j}T_{N,j}^{D}$$

$$\text{s.t.}\quad\sum_{i=0}^{K}P_{N,i}\le Th_R.$$

를 통해 산출되며,

상기 $P_{i,K}$ 는 제1 케이스의 각 상태의 거절 가능성을 나타내고, $T^{D}_{i,K}$는 제1 케이스의 대기 시간을 나타내며, $P_{N,j}$ 는 제2 케이스의 각 상태의 거절 가능성을 나타내고, 상기 $T^{D}_{N,j}$는 제2 케이스의 대기 시간을 나타내며, 상기 $P_{N,i}$ 은 거절 가능성을 나타내고, 상기 Th_R 은 거절 가능성의 임계값을 나타내는 것을 특징으로 하는 실시간 마이그레이션 자원 할당 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 클라우드 컴퓨팅 환경에 관한 기술로, 보다 상세하게는 클라우드 컴퓨팅 환경에서의 자원할당에 관한 기술이다.

배 경 기 술

[0002] 클라우드 환경은 가상 물리서버의 새로운 패러다임이다. 특히, 클라우드 환경과 연계된 가상 머신은 하나의 물리적 장치를 둘 이상의 가상머신으로 분할함으로써, 자원 활용 및 공유의 효율성을 극대화 할 수 있다. 가상 머신(Virtual Machine)을 할당하는 과정은 일반적으로 가상 머신 할당 시 발생하는 작업부하의 특성을 파악하여 가장 효율적인 작업부하를 가질 수 있도록 자원을 할당한다. 그리고, 필요한 경우 성능 최적화 및 저전력 운용을 위해 가상머신 마이그레이션(Migration)을 활용하여 작업 효율을 높이고 있다. 가상 머신에서 실시간 마이그레이션(Live Migration)은 클라우드 컴퓨팅 환경에서 나타나는 중요한 컴포넌트이며, 데이터 센터에서 가상 머신 마이그레이션은 성능 및 비용을 향상시킬 수 있는 방법이 될 수 있다. 가상 머신을 마이그레이션 하는 것은 두 개의 물리 호스트를 필요로 하는데 하나는 가상 머신이 본래 존재하던 호스트이고, 다른 하나는 가상 머신이 마이그레이션할 타겟 호스트이다. 그러나, 가상 머신 마이그레이션은 추가적인 자원(slot)을 필요로 하며, 자원(Slot)을 모두 사용하고 있을 경우 새로운 서비스 요청에 대한 거절 가능성이 높아진다.

[0003] 데이터 센터 내에서의 마이그레이션은 가상 머신의 메모리에 상주하는 콘텐츠가 하나의 물리적 호스트에서 다른 하나의 물리적 타겟 호스트로 이동하는 과정이다. 마이그레이션은 크게 오프라인 마이그레이션(Offline Migration) 및 실시간 마이그레이션(Live Migration)으로 구분할 수 있다. 오프라인 마이그레이션은 원본 호스트에서 실행중인 가상 머신이 중지되고 중지된 가상 머신 메모리의 데이터(콘텐츠)가 타겟 호스트로 이동된 후, 타겟 호스트에서 가상 머신이 다시 시작되는 마이그레이션이다. 오프라인 마이그레이션에서, 가상 머신이 타겟 호스트로 이동하면, 원본 호스트의 메모리는 프리 상태가 된다. 반면에, 실시간 마이그레이션은 원본 호스트의 가상 머신이 동작하는 상태에서 타겟 호스트로 메모리의 데이터가 이동한다. 처음에 모든 페이지들이 원본 호스트에서 타겟 호스트로 움직이며, 동시에 이전에 전송된 단계가 복사되는 동안 CPU 상태와 다른 남아있는 메모리 페이지들이 전송되는데 이 페이지들은 더티 페이지(Dirty Page)가 된다. 이후, 원본 호스트는 동작을 중지하고, 타겟 호스트가 가상 머신이 동작하는 주요 호스트가 된다. 실시간 마이그레이션은 타겟 호스트로 가상 머신이 이동하는 과정에서 실행중인 가상 머신이 종료되지 않고, 끊임없이 변화되는 메모리 페이지가 전송되기 때문에, 더 작은 서비스 다운타임(Downtime)을 가지게 된다.

[0004] 실시간 마이그레이션은 원본 호스트에서 타겟 호스트로 가상 머신이 마이그레이션 하는 과정에서, 원본 가상 머신 메모리의 데이터를 원본 가상 머신 실행 중에도 끊김 없이 이동시키기 위해 이동 슬롯을 필요로 한다. 만약 할당된 이동 슬롯의 최대 개수보다 많은 마이그레이션 요청이 수신되면, 수용 가능한 이동 슬롯의 개수가 부족하게 되어 가상 머신의 마이그레이션 요청이 거절되거나 대기 시간(Waiting)이 길어질 수 있다. 대기 시간은 바로 QoS(Quality of Service)에 영향을 주게 되기 때문에, QoS를 만족시키기 위해서는 이동 슬롯의 개수를 늘리는 것이 바람직하다. 하지만, 이동 슬롯은 물리적 자원이 할당된 가상 머신이기 때문에, 이동 슬롯의 개수를 늘릴 경우, 타겟 호스트의 자원이 낭비되거나 부족해지는 문제가 발생할 수 있다. 따라서, 물리적 자원의 효과적인 사용을 위해서는 한정된 물리적 자원 내에서 최적의 이동 슬롯을 선택하는 것이 필요하다.

[0005] 대한민국 공개특허 제10-2013-0074952호는 클라우드 컴퓨팅 시스템의 자원을 효율적으로 사용하기 위한 가이드를 제공하는 장치 및 방법에 대한 발명으로, 가상 단말 할당 자원의 비용을 고려하여 최적화하는 기술을 포함하고 있다. 하지만, 이는 실행시간에 기초하여 추정된 메모리 사용량만을 고려할 뿐, 대기 시간이나 거절 가능성에 따른 서비스 편의성은 고려되지 않았다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 공개특허 제10-2013-0074952호

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자 하는 과제는 가상 단말의 실시간 마이그레이션 과정에서 한정된 자원 내에서 QoS를 만족시킬 수 있는 최적의 이동 슬롯을 선택하기 위한 자원 할당 방법에 관한 기술을 제공하는 것이다.

과제의 해결 수단

[0008] 본 발명에 따른 실시간 마이그레이션 자원 할당 시스템은 하나 이상의 가상 단말을 할당하고, 마이그레이션된 서비스를 가상 단말에서 제공하는 타겟 서버, 하나 이상의 가상 단말을 할당하고, 마이그레이션 슬롯이 할당되며, 할당된 마이그레이션 슬롯을 통해 가상 단말에서 제공하는 상기 서비스를 상기 타겟 서버로 마이그레이션하는 원본 서버 및 거절 가능성 및 대기 시간을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정하고, 결정된 마이그레이션 슬롯 개수에 기초하여 원본 서버에 마이그레이션 슬롯을 할당하는 실시간 마이그레이션 슬롯 할당 장치를 포함한다.

[0009] 실시간 마이그레이션 슬롯 할당 장치는 거절 가능성의 임계값(Threshold)를 설정하고, 설정된 거절 가능성의 임계값 범위 내에서 상기 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정한다. 그리고, 실시간 마이그레이션 슬롯 할당 장치는 서비스 전이에 따른 각 상태의 거절 가능성의 합산을 통해 거절 가능성을 산출한다. 또한, 실시간 마이그레이션 슬롯 할당 장치는 동시에 하나의 마이그레이션 슬롯만 허용된 제1 케이스 및 동시에 둘 이상의 마이그레이션 슬롯이 허용된 제2 케이스 각각에 대한 대기 시간을 산출할 수 있다. 실시간 마이그레이션 슬롯 할당 장치는 거절 가능성의 임계값 범위 내에서 제1 케이스 및 제2 케이스의 대기 시간 및 거절 가능성을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수(K)를 산출한다.

[0010] 본 발명에 따른 실시간 마이그레이션 자원 할당 방법은 먼저, 거절 가능성을 산출하고, 대기 시간을 산출한다. 그리고, 거절 가능성의 허용 범위인 거절 가능성의 임계값을 설정한다. 다음으로, 대기 시간 및 거절 가능성을 고려하여 예측 대기 시간을 산출한다. 그리고, 산출된 예측 대기 시간을 최소화하는 최적 마이그레이션 슬롯 개수를 결정하고, 결정된 최적 마이그레이션 슬롯 개수에 기초하여 마이그레이션 슬롯을 할당한다.

[0011] 예측 대기 시간을 최소화하는 최적 마이그레이션 슬롯 개수를 설정하는 단계는 거절 가능성의 임계값(Threshold)를 설정하고, 설정된 거절 가능성의 임계값 범위 내에서 상기 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수를 결정한다. 그리고, 거절 가능성을 산출하는 단계는 서비스 전이에 따른 각 상태의 거절 가능성의 합산을 통해 거절 가능성을 산출할 수 있다.

[0012] 대기 시간을 산출하는 단계는, 동시에 하나의 마이그레이션 슬롯만 허용된 제1 케이스 및 동시에 둘 이상의 마이그레이션 슬롯이 허용된 제2 케이스 각각에 대한 대기 시간을 산출한다. 그리고, 예측 대기 시간을 최소화하는 최적 마이그레이션 슬롯 개수를 설정하는 단계는 거절 가능성의 임계값 범위 내에서 제1 케이스 및 상기 제2 케이스의 대기 시간 및 거절 가능성을 고려하여 산출된 예측 대기 시간을 최소화하는 마이그레이션 슬롯 개수(K)를 산출한다.

발명의 효과

[0013] 본 발명에 따른 가상 단말의 실시간 마이그레이션 자원 할당 시스템 및 방법은 거절 가능성 및 대기 시간을 모두 고려하여 마이그레이션 슬롯의 개수를 설정함으로써, 한정된 자원 내에서 서로 상반되는 대기 시간과 거절 가능성의 협상을 통해 서비스 품질을 만족시키는 서비스 마이그레이션을 가능하게 한다.

도면의 간단한 설명

[0014] 도 1은 본 발명에 따른 실시간 마이그레이션 자원 할당 시스템의 일 실시예를 나타내는 도면이다.

도 2는 본 발명의 일 실시예에 따른 실시간 마이그레이션 슬롯 할당 장치의 서비스 전이도를 나타내는 도면이다.

도 3a는 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치의 제1 케이스의 대기 시간을 설명하기 위한 도면이다.

도 3b는 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치의 제2 케이스의 대기 시간을 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 실시간 마이그레이션 자원 할당 방법을 나타내는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0015] 이하, 본 발명의 실시예를 첨부된 도면들을 참조하여 상세하게 설명한다. 본 명세서에서 사용되는 용어 및 단어들은 실시예에서의 기능을 고려하여 선택된 용어들로서, 그 용어의 의미는 발명의 의도 또는 관례 등에 따라 달라질 수 있다. 따라서 후술하는 실시예에서 사용된 용어는, 본 명세서에 구체적으로 정의된 경우에는 그 정의에 따르며, 구체적인 정의가 없는 경우는 당업자들이 일반적으로 인식하는 의미로 해석되어야 할 것이다.

[0016] 도 1은 본 발명에 따른 실시간 마이그레이션 자원 할당 시스템의 일 실시예를 나타내는 도면이다.

[0017] 도 1을 참조하면, 본 발명에 따른 실시간 마이그레이션 자원 할당 시스템은 실시간 마이그레이션 슬롯 할당 장치(100), 원본 서버(10) 및 타겟 서버(20)를 포함할 수 있다. 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 원본 서버(10)에서 서비스를 제공하고 있는 하나 이상의 가상 머신(Virtual Machine)을 타겟 서버(20)로 마이그레이션(Migration)한다. 원본 서버(10) 및 타겟 서버(20)는 특정한 서버로 한정되는 것은 아니며, 서비스를 마이그레이션하기 위한 가상 단말의 슬롯을 보유하고 있는 서버가 원본 서버(10)가 되고, 서비스를 제공하는 가상 단말의 슬롯이 마이그레이션하는 대상 서버가 타겟 서버(20)가 되며, 원본 서버(10) 및 타겟 서버(20)는 서로 교체될 수 있다.

[0018] 원본 서버(10)에서 한정된 연산 자원(Resource) 내에서 다수의 가상 머신의 슬롯(Slot)을 할당하여 사용자에게 서비스를 제공한다. 하지만, 원본 서버(10)의 연산 자원에 여유가 부족하거나, 사용자에게 보다 효과적으로 서비스를 제공할 수 있는 서버가 있는 경우, 해당 서버를 타겟 서버(20)로 하여 서비스를 마이그레이션한다. 실시간 마이그레이션 슬롯 할당 장치(100)가 원본 서버(10)에서 타겟 서버(20)로 서비스를 마이그레이션하는 방법은 오프라인 마이그레이션과 실시간 마이그레이션으로 분류할 수 있다. 오프라인 마이그레이션은 마이그레이션에 소요되는 시간이 짧으나, 서비스를 제공하는 가상 머신이 정지된 상태로 마이그레이션이 수행되기 때문에 서비스가 끊기는 다운타임이 존재하게 된다. 반면에, 실시간 마이그레이션은 가상 머신이 실행중인 상태에서 마이그레이션을 수행하기 때문에 다운타임은 없으나, 소요시간이 증가하게 된다.

[0019] 실시간 마이그레이션 슬롯 할당 장치(100)가 실시간 마이그레이션을 수행할 때, 마이그레이션을 수행할 가상 머신의 슬롯 개수가 늘어나면, 한번에 마이그레이션을 수행할 수 있는 서비스 요청이 늘어날 수 있으나, 타겟 서버(20)로 새로운 서비스 요청이 수신될 시 가용한 가상 머신의 슬롯이 부족하여 거절(Rejection)이 발생할 수 있다. 예를 들어, 원본 서버(10) 및 타겟 서버(20)의 슬롯 개수가 4개 할당되어 있고, 원본 서버(10)에서 총 네 개의 서비스를 제공하고 있을 때, 원본 서버(10)의 슬롯 4개 모두를 동시에 타겟 서버(20)로 마이그레이션 할 경우 타겟 서버(20)의 슬롯 4개 모두가 마이그레이션 상태에 놓이기 때문에, 타겟 서버(20)로 새로운 서비스 요청이 수신되면 거절이 발생할 수 있다.

[0020] 반면에, 원본 서버(10)의 슬롯 중에서 2개씩 마이그레이션을 수행한다면, 마이그레이션 소요 시간 동안 타겟 서버(20)에 두 개의 슬롯에 여우가 있기 때문에, 새로운 서비스 요청이 타겟 서버(20)로 수신되어도 정상적으로 서비스가 가능하다. 하지만, 이 경우, 원본 서버(10)에 남아있는 두 개의 슬롯은 처음 두 개의 마이그레이션이 종료될 때까지 마이그레이션을 수행하지 못하고 대기 시간이 발생하게 된다. 따라서, 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 동시에 마이그레이션을 수행할 수 있는 최적의 마이그레이션 슬롯 개수(K)를 설정한다. 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 거절가능성이 기 설정된 임계값을 만족시키는 범위 내에서 대기 시간을 최소화시키는 마이그레이션 슬롯 개수(K)를 설정한다. 실시간 마이그레이션 슬롯 할당 장치(100)는 거절 가능성 및 대기 시간을 포아송 프로세스(Possion Process) 및 지수분포를 이용해 고려하여 최적의 마이그레이션 슬롯 개수를 설정한다. 실시간 마이그레이션 슬롯 할당 장치(100)가 최적의 마이그

레이션 슬롯 개수 K를 설정하는 과정은 후술하는 도 2 내지 도 3a/도 3b에서 추가적으로 설명하도록 한다.

[0021] 도 2는 본 발명의 일 실시예에 따른 실시간 마이그레이션 슬롯 할당 장치의 서비스 전이도를 나타내는 도면이다.

[0022] 도 2를 참조하면, 본 발명의 일 실시예에 따른 실시간 마이그레이션 슬롯 할당 장치(100)의 서비스 전이도(Transition Diagram, 200)를 통해 서비스 마이그레이션 과정에서 거절 가능성을 산출한다. 서비스 전이도(200)는 데이터 센터 내의 가능한 흐름을 사용한 것이며, 이를 사용하여 거절 가능성과 대기 시간에 각 상태의 가능성을 기반으로 공식화한다. 실시간 마이그레이션 슬롯 할당 장치(100)는 거절 가능성과 대기 시간 사이에서 협상을 통해 최적의 K 값을 산출한다. 서비스 전이도(200)에서, 가상 머신 서비스가 끝나면 전체 가상머신의 수는 1 감소하고, 새로운 요청이 수신되면 전체 가상머신 슬롯의 수는 1증가하게 된다. 마이그레이션이 끝나게 되면, 전체 가상머신과 마이그레이션 슬롯의 수는 1 감소하게 된다. 그리고, 마이그레이션 요청이 수신되면 전체 가상머신 슬롯과 마이그레이션 슬롯의 수는 1 증가하게 된다.

[0023] 서비스 전이도(200)를 구성하는 각각의 구성요소 (I,J)에서 I는 사용하고 있는 전체 슬롯 즉, 서비스를 제공중인 가상 단말 슬롯을 나타내며, J는 마이그레이션 하고 있는 슬롯을 나타낸다.

표 1

서비스 전이도(200)의 균형방정식

State	Rate that the process leaves = rate that it enters	Category
$(0,0)$	$\lambda_s P_{0,0} = \mu_s P_{1,0}$	1
$(N,0)$	$N\mu_s P_{N,0} = \lambda_s P_{N-1,0}$	2
$(i,0);$ $0 < i < N$	$(i\mu_s + \lambda_s + i\lambda_m)P_{i,0} =$ $\lambda_s P_{i-1,0} + (i+1)\mu_s P_{i+1,0} + \mu_m P_{i+1,1}$	3
$(2K,K);$ $2K = N$	$K\mu_m P_{N,K} = \lambda_m P_{N-1,K-1}$	4
$(N,j);$ $0 < j < \frac{N}{2}$	$((N-2j)\mu_s + j\mu_m)\,P_{N,j} = \lambda_s P_{N-1,j} + (N-2j+1)\lambda_m P_{N-1,j-1}$	5
$(i,j);$ $i \neq N, j < K,$ $0 < j < \frac{i}{2}$	$((i-2j)\mu_s + j\mu_m + \lambda_s + (i-2j)\lambda_m)\,P_{i,j} = (i-2j+1)\lambda_m P_{i-1,j-1}$ $+\lambda_s P_{i-1,j} + (i-2j+1)\mu_s P_{i+1,j} + (j+1)\mu_m P_{i+1,j+1}$	6
$(i,K);$ $0 < K < \frac{i}{2}$	$((i-2K)\mu_s + K\mu_m + \lambda_s)\,P_{i,K} =$ $(i-2K+1)\lambda_m P_{i-1,K-1} + \lambda_s P_{i-1,K} + (i-2K+1)\mu_s P_{i+1,K}$	7

[0024]

표 2

[0025] 변수 정의

변수명	설명	
K	마이그레이션 과정동안 제한할 VM 수(마이그레이션 슬롯 개수)	
N	현재 사용되는 전체 자원(가용 전체 슬롯 개수)	
λ_s	서비스 발생 비율(새로운 사용자 도달률)	포아송 프로세스
λ_m	마이그레이션 서비스 발생 비율	
μ_s	평균서비스 시간	지수분포
μ_m	평균 마이그레이션 시간	
Th_R	거절 가능성이 만족해야할 임계값	
$E[T^D]$	예측 대기 시간(지수분포에 근거)	

[0026] 표 1은 서비스 전이도(200)의 균형방정식을 나타낸다. 안정 상태를 가정하여 rate in = rate out(안정 상태: 시스템이 한 상태를 떠나면 언젠가는 돌아온다는 안정 상태를 가정) 균형 방정식을 만들 수 있다. 상태 (0,0)을 나타내는 카테고리(category) 1(201)에서 서비스 발생 비율(λ_s)과 (0,0)에서의 거절가능성($P_{0,0}$)의 곱은 평균 서비스 시간(μ_s)과 (1,0)에서의 거절가능성의 곱과 같다. 카테고리 1(201)이 나타내는 상태 (0,0)은 서비스를 제공중인 가상 단말 슬롯 및 마이그레이션 하고 있는 슬롯 모두 0인 상태, 즉 어떤 서비스 및 마이그레이션도 수행하고 있지 않은 상태를 나타낸다. 그리고, 상태 (1,0)은 서비스를 제공중인 가상 단말 슬롯이 1개이고 마이그레이션 하고 있는 슬롯은 0인 상태로서, 상태 (0,0)에서 (1,0)으로 변화된 상태는 하나의 서비스가 발생한 것을 나타낸다. 그리고, 카테고리 2의 상태(N,0)에서 $N\lambda_s P_{N,0}$은 $\mu_s P_{N-1,0}$과 동일하다.

[0027] 상태 (1,0)에서 상태 (2,1)로 변화되는 것은 하나의 가상 단말 슬롯을 통해 서비스를 제공하는 중에 마이그레이션이 발생한 것으로, 마이그레이션 또한 서비스이므로 서비스가 1 증가한다. 이를 기반으로 상태 (3,1)의 거절가능성을 계산하면, 수학식 1과 같다.

수학식 1

[0028]
$$
\begin{aligned}
&(0,0)\ \lambda_s P_{0,0} = \mu_s P_{1,0},\\
&(1,0)\ (\mu_s + \lambda_s + \lambda_m) P_{1,0} = \lambda_s P_{0,0} + 2\mu_s P_{2,0} + \mu_m P_{2,1},\\
&(2,0)\ (2\mu_s + \lambda_s + 2\lambda_m) P_{2,0} = \lambda_s P_{1,0} + 3\mu_s P_{3,0} + \mu_m P_{3,1},\\
&(3,0)\ 3\mu_s P_{3,0} = \lambda_s P_{2,0},\\
&(2,1)\ (\mu_m + \lambda_s) P_{2,1} = \lambda_m P_{1,0} + \mu_s P_{3,1},\\
&(3,1)\ (\mu_s + \mu_m) P_{3,1} = \lambda_s P_{2,1} + 2\lambda_m P_{2,0},
\end{aligned}
$$

[0029] 수학식 1은 도 1의 전이도 및 표 1에 기초하여 상태 (3,1)의 거절가능성을 산출하는 수식을 나타낸다. 수학식 1에서, λ_s=0.5, λ_m=1, μ_s=0.2, μ_m-0.1로 주어질 때, (3,1) 상태의 거절 가능성을 개산하면, $P_{0,0}$=0.0103, $P_{1,0}$=0.0258, $P_{2,0}$=0.0323, $P_{3,0}$=0.0269, $P_{2,1}$=0.2585 및 $P_{3,1}$=0.6461로 산출될 수 있다. 상술한 각 상태의 거절 가능성에 기초하여 거절 가능성은 수학식 2와 같이 나타낼 수 있다.

수학식 2

[0030]
$$P_r(R) = \sum_{j=0}^{K} P_{N,j}$$

[0031] 수학식 2에서, 거절 가능성 Pr(R)은 각 상태의 거절 가능성의 합산 형태로 나타내어 진다. 예를 들어, 서비스를 제공하는 가상 머신 슬롯이 3(N=3)이고, 마이그레이션 슬롯 개수가 1(K=1)이라면, 거절 가능성 Pr(R)은 $P_{3,0}+P_{3,1}$=0.0269+0.6461=0.673으로 산출될 수 있다.

[0032] 도 3a는 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치의 제1 케이스의 대기 시간을 설명하기 위한 도면이다.

[0033] 도 3a를 참조하면, 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치는 마이그레이션 과정 동안 충분한 슬롯을 가지고 있지 않으면 대기 시간이 발생한다. 대기 시간 발생은 할당된 마이그레이션 슬롯이 모두 사용중이거나, 원본 서버(10)의 가용 가상 머신 슬롯이 모두 사용 중일 경우 발생하게 된다. 할당된 마이그레이션 슬롯이 모두 사용중인 경우, 상태는 (2K+1,K), (2K+2,K),…,(N,K)가 된다. 그리고, 원본 서버(10)의 가용 가상 머신 슬롯이 모두 사용 중일 경우 상태는 (N,0), (N,1),…,(N,K-1)이 된다.

[0034] 제1 케이스는 동시에 하나의 마이그레이션 슬롯만 허용된 경우로, 총 4명의 사용자(User)가 원본 서버(10)에 접

속하여 있으며, 원본 서버(10)는 가용 가상 머신 슬롯이 5개(N=5)이고, 마이그레이션 슬롯 개수가 1개(K=1) 일 때, 사용자 3(User 3) 및 사용자 4(User 4)가 마이그레이션을 요청한 경우를 나타낸다. 도 3a에서 ts는 서비스 시간, tm은 마이그레이션 시간을 나타내며, S^s는 각 사용자의 서비스 요청 시간을 나타낸다. 사용자 3이 마이그레이션 요청 시간(S_1^m)이후 마이그레이션을 시작하여 마이그레이션 시간(301) 동안 마이그레이션을 수행한다.

[0035] 사용자 4는 마이그레이션 요청 시간(S_2^m)이후 마이그레이션을 요청하지만, 제1 케이스에서는 K=1이기 때문에 마이그레이션을 바로 수행하지 못하고 대기 시간(T, 302)가 발생하게 된다. 또한, 사용자 3이 마이그레이션을 수행하는 도중에 사용자 1의 서비스가 종료되지만, 마이그레이션 슬롯의 개수가 1개이기 때문에, 사용자 4에 대한 마이그레이션이 시작될 수 없다. 사용자 4는 대기 상태를 유지하다 사용자 3의 마이그레이션이 끝난 시점에 사용자 4에 대한 마이그레이션이 수행된다. 따라서, 제1 케이스(N=5,K=1)의 사용자 4의 대기 시간(302)은 수학식 3과 같이 산출된다.

수학식 3

[0036]
$$T_{i,k}^{D} = E[S_1^m] + E[t_m] - E[S_{K+1}]$$

[0037]
$$2K < i \leq N$$

수학식 4

[0038]
$$E[S_i^s] = \frac{i}{\lambda_s}, E[S_i^m] = \frac{i}{\lambda_m}$$

[0039] 수학식 3에서 $T_{i,k}^{D}$는 제1 케이스의 대기 시간을 나타내고, $E[S_1^m]$는 첫 번째 마이그레이션 요청 사용자(사용자 3)의 마이그레이션 요청 시간을 나타내며, $E[t_m]$는 사용자 3의 마이그레이션 요청시간이며, 일반화하여 수학식 3을 도출하여 마이그레이션 요청 사용자 K의 요청 시간을 나타낼 수 있다. 마이그레이션 시간을 나타내고, $E[S_{K+1}]$는 두 번째 마이그레이션 요청 사용자(사용자 4)의 마이그레이션 요청 시간을 나타내며, 일반화를 통해 K+1번째 요청자에 의한 요청 서비스 시간을 나타낼 수 있다. 즉, 제1 케이스의 대기 시간은 사용자 3의 마이그레이션 요청 시간과 마이그레이션 시간에서 사용자 4의 마이그레이션 요청 시간을 제거하여 산출될 수 있다. 수학식 4는 지수 분포를 이용하여 산출된 사용자 i에 대한 평균 서비스 시간 및 마이그레이션 시간을 나타낸다. 지수 분포를 갖는 동일한 두 사건 중에서 어느 것이든지 먼저 사건이 일어날 때까지의 시간 분포도 역시 지수시간분포를 가지며 평균 서비스 시간은 지수 시간의 1/2이 되기 때문에 지수분포에 의해 사용자 i에 대한 평균 서비스 시간 및 마이그레이션 시간은 위와 같다. 수학식 3 및 수학식 4를 정리하면, 수학식 5와 같이 나타난다.

수학식 5

[0040]
$$T_{i,K}^{D} = \frac{1}{\lambda_m} + \frac{1}{\mu_m} - \frac{K+1}{\lambda_m} = \frac{\lambda_m - K\mu_m}{\lambda_m \mu_m}$$

[0041]
$$T_{i,K}^{D} = \begin{cases} \dfrac{\lambda_m - K\mu_m}{\lambda_m \mu_m}, & 0 \le K \le \rho_m, \\ 0, & \rho_m < K \le N \end{cases}$$

$\rho_m = \lambda_m / \mu_m$

[0042] 수학식 5는 제1 케이스일 경우 대기 시간을 산출하기 위한 수학식이다. 제1 케이스의 경우 최종적으로 정리된 수학식 5를 통해 대기 시간을 산출할 수 있다. 수학식 5에서 ρ_m는 서비스 이용률을 나타낸다.

[0043] 도 3b는 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치의 제2 케이스의 대기 시간을 설명하기 위한 도면이다.

[0044] 도 3b를 참조하면, 제2 케이스는 총 4명의 사용자(User)가 원본 서버(10)에 접속하여 있으며, 원본 서버(10)는 가용 가상 머신 슬롯이 5개(N=5)이고, 마이그레이션 슬롯 개수가 2개(K=2)일 때, 사용자 3(User 3) 및 사용자 4(User 4)가 마이그레이션을 요청한 경우를 나타낸다.

[0045] 사용자 3이 마이그레이션 요청 시간(S_1^m)이후, 마이그레이션을 시작하여 마이그레이션 시간(tm) 동안 마이그레이션을 수행한다. 사용자 4는 마이그레이션 요청 시간(S_2^m)이후 마이그레이션을 요청하지만, 총 5개의 가용 가상 머신 슬롯이 5개이고, 사용자가 4명이기 때문에, 마이그레이션을 수행할 수 있는 가상 머신 슬롯은 한 개이다. 하지만, 한 개의 가상 머신 슬롯은 사용자 3이 마이그레이션하고 있기 때문에, 사용자 4는 마이그레이션을 수행하지 못하고 대기 시간을 가지게 된다.

[0046] 303에서는 사용자 3의 마이그레이션이 끝난 이후, 사용자 4의 마이그레이션이 수행된다. 반면에, 304에서는 사용자 3의 마이그레이션이 끝나기 전에, 사용자 1의 서비스가 종료된다. 따라서, 사용자 1이 사용하던 가상 머신 슬롯에 여유가 생겨 사용자 3의 마이그레이션이 종료 여부와 관계없이 사용자 1의 서비스 종료 이후 사용자 4의 마이그레이션을 수행할 수 있다. 이를 통해, 제2 케이스의 대기 시간을 산출할 수 있다.

수학식 6

[0047]
$$\begin{aligned} T_{N,j}^{D} &= \Pr(Y_m \cdot (E[S_1^m] + t_m - E[S_{j+1}^m]) + \Pr(Y_s) \cdot E[S_1^s] + t_s - E[S_{j+1}^m] \\ &= P(t_s \ge t_m + E[S_1^m] + E[S_{j+1}^m] \cdot (\frac{1}{\lambda_s} + \frac{1}{\mu_s} - \frac{j+1}{\lambda m}) \\ &= e^{\mu_s(\lambda_m - \mu_m)/\mu_m\lambda_m}(\frac{\lambda_m - j\mu_m}{\mu_m\lambda_m}) + (1 - e^{-\mu_s(\lambda_m - \mu_m)/\mu_m\lambda_m})(\frac{1}{\lambda_s} + \frac{1}{\mu_s} - \frac{j+1}{\lambda_m}) \end{aligned}$$

[0048] 수학식 6은 제2 케이스의 대기 시간을 산출하기 위한 수학식이다. 수학식 6에서 j는 사용자를 나타낸다. 제1 케이스의 대기 시간을 산출하는 수학식 5 및 제2 케이스의 대기 시간을 산출하는 수학식 6과 거절 가능성을 산출하는 수학식 2를 고려하여 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)의 예측 대기 시간을 산출할 수 있다.

[0049] 제 2 케이스의 대기시간의 경우는 4명의 유저가 요청을 했을 때 현재 총 슬롯은 5개이고 마이그레이션 할 수 있는 슬롯은 2개이다. 요청은 유저에 따른 서비스가 4개이고 마이그레이션 서비스는 2개이다. 요청에 따른 총 필

요한 슬롯의 수는 6개이나 주어진 총 슬롯은 5개이므로 이 사이에 대기시간이 발생하게 된다. 303의 경우, 서비스 4개의 요청이 슬롯에 할당되고 마이그레이션 서비스 요청슬롯은 하나만 시작하게 된다. 마이그레이션 요청하나가 먼저 끝나게 되어 기다렸다가 마이그레이션을 시작하게 된다. 304의 경우는 서비스 4개 요청이 슬롯에 할당되고 마이그레이션 서비스 요청슬롯은 하나만 시작하게 된다. 이 경우에는 사용자 1의 서비스가 먼저 끝나게 되어 끝날때까지 기다리다가 마이그레이션을 시작하게 된다. 제2 케이스의 경우에는 수학식 7과 같이 303(a) 및 304(b)의 경우를 합하여 대기 시간을 산출할 수 있다.

수학식 7

[0050]
$$\underbrace{\Pr(Y_m).(E[S_1^m]+t_m-E[S_{j+1}^m])}_{(a)}$$
$$+\underbrace{\Pr(Y_s).(E[S_1^s]+t_s-E[S_{j+1}^m])}_{(b)}$$

수학식 8

[0051]
$$E[T^D]=\underbrace{\sum_{i=2K+1}^{N}P_{i,K}T_{i,K}^D}_{case\,1}+\underbrace{\sum_{i=0}^{K-1}P_{N,j}T_{N,j}^D}_{case2}$$

[0052] 수학식 8에서, $E[T^D]$ 는 제1 케이스 및 제2 케이스를 모두 포함하는 실시간 마이그레이션 슬롯 할당 장치(100)의 대기 시간을 나타낸다. 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 제1 케이스 및 제2 케이스에 대한 거절 가능성 및 대기 시간을 고려하여 산출된 예측 대기 시간을 최소화하는 최적 K값을 산출한다.

수학식 9

[0053]
$$\min_K \quad \sum_{i=2K+1}^{N}P_{i,K}T_{i,K}^D+\sum_{j=0}^{K-1}P_{N,j}T_{N,j}^D$$
$$\text{s.t.} \quad \sum_{i=0}^{K}P_{N,i}\le Th_R.$$

[0054] 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 수학식 8을 통해 최적의 마이그레이션 슬롯을 산출한다. 즉, 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 거절 가능성 및 대기 시간을 고려한 예측 대기 시간을 최소화 할 수 있는 K값을 산출하여 최적의 마이그레이션 슬롯을 산출한다. Th_R은 거절 가능성이 만족해야 하는 임계값(Threshold)으로, 거절 가능성이 만족해야 하는 임계값 내에서 예측 대기 시간을 최소화하는 K값을 산출한다.

표 3

[0055] K값 결정 알고리즘

K값을 결정하는 알고리즘

Require $N, \lambda_s, \lambda_m, \mu_s, \mu_m, Th_R$

Find Range Of K from Constraint $\sum_{i=0}^{K} P_{N,i} \le Th_R$;

For k∈Range of K do

$$E[T^D] = \underbrace{\sum_{i=2K+1}^{N} P_{i,K} T^D_{i,K}}_{case\,1} + \underbrace{\sum_{i=0}^{K-1} P_{N,j} T^D_{N,j}}_{case\,2}$$

Find ;

if $E[T^D]$ is minimum then

$K^* = k$

end if

end for

return K^*

[0056] 표 3은 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)의 최적 K값을 산출(결정)하는 알고리즘을 나타낸다. 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 표 3에 따른 알고리즘을 통해 마이그레이션 슬롯을 추가하여 대기 시간을 줄이는 방법으로 마이그레이션 슬롯의 수(K)를 찾는다. 이를 통해, 대기 시간을 최소화하면서 거절가능성을 임계값 내에서 만족시킬 수 있다.

[0057] 도 4는 본 발명의 일 실시예에 따른 실시간 마이그레이션 자원 할당 방법을 나타내는 흐름도이다.

[0058] 도 2 내지 도 4를 참조하면, 본 발명의 일 실시예에 따른 실시간 마이그레이션 자원 할당 방법은 먼저, 마이그레이션 과정에서의 거절 가능성을 산출한다(401). 도 2의 서비스 전이도를 통해 서비스 마이그레이션 과정에서의 거절 가능성을 산출할 수 있다. 표 1의 서비스 전이도의 균형 방정식에서 상태 (0,0)을 나타내는 카테고리(category) 1(201)에서 서비스 발생 비율(λ_S)과 (0,0)에서의 거절가능성($P_{0,0}$)의 곱은 평균서비스 시간(μ_S)과 (1,0)에서의 거절가능성의 곱과 같다. 카테고리 1(201)이 나타내는 상태 (0,0)은 서비스를 제공중인 가상 단말 슬롯 및 마이그레이션 하고 있는 슬롯 모두 0인 상태, 즉 어떤 서비스 및 마이그레이션도 수행하고 있지 않은 상태를 나타낸다. 그리고, 상태 (1,0)은 서비스를 제공중인 가상 단말 슬롯이 1개이고 마이그레이션 하고 있는 슬롯은 0인 상태로서, 상태 (0,0)에서 (1,0)으로 변화된 상태는 하나의 서비스가 발생한 것을 나타낸다. 그리고, 카테고리 2의 상태(N,0)에서 $N\lambda_S P_{N,0}$은 $\mu_S P_{N-1,0}$과 동일하다.

[0059] 상태 (1,0)에서 상태 (2,1)로 변화되는 것은 하나의 가상 단말 슬롯을 통해 서비스를 제공하는 중에 마이그레이션이 발생한 것으로, 마이그레이션 또한 서비스이므로 서비스가 1 증가한다. 이를 통해, 거절 가능성은 수학식 2와 같이 나타낼 수 있다. 수학식 2에서, 거절 가능성 Pr(R)은 각 상태의 거절 가능성의 합산 형태로 나타내어진다.

[0060] 다음으로 제1 케이스 및 제2 케이스의 대기 시간을 산출한다(402). 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 마이그레이션 과정 동안 충분한 슬롯을 가지고 있지 않으면 대기 시간이 발생한다. 대기 시간 발생은 할당된 마이그레이션 슬롯이 모두 사용중이거나, 원본 서버(10)의 가용 가상 머신 슬롯이 모두 사용 중일 경우 발생하게 된다.

[0061] 제1 케이스는 한번에 하나의 마이그레이션 슬롯 만을 허용하는 경우이다. 하나의 실시예로서, 제1 케이스는 총 4명의 사용자(User)가 원본 서버(10)에 접속하여 있으며, 원본 서버(10)는 가용 가상 머신 슬롯이 5개(N=5)이고, 마이그레이션 슬롯 개수가 1개(K=1) 일때, 사용자 3(User 3) 및 사용자 4(User 4)가 마이그레이션을 요청한 경우를 나타낸다. 도 3a에서 ts는 서비스 시간, tm은 마이그레이션 시간을 나타내며, S^s는 각 사용자의 서비스 요청 시간을 나타낸다. 사용자 3이 마이그레이션 요청 시간(S_1^m)이후 마이그레이션을 시작하여

마이그레이션 시간(tm) 동안 마이그레이션을 수행한다.

[0062] 사용자 4는 마이그레이션 요청 시간(S_2^m)이후 마이그레이션을 요청하지만, 제1 케이스에서는 K=1이기 때문에 마이그레이션을 바로 수행하지 못하고 대기 시간(T)가 발생하게 된다. 또한, 사용자 3이 마이그레이션을 수행하는 도중에 사용자 1의 서비스가 종료되지만, 마이그레이션 슬롯의 개수가 1개이기 때문에, 사용자 4에 대한 마이그레이션이 시작될 수 없다. 사용자 4는 대기 상태를 유지하다 사용자 3의 마이그레이션이 끝난 시점에 사용자 4에 대한 마이그레이션이 수행된다. 따라서, 제1 케이스의 사용자 4의 대기 시간은 수학식 3과 같이 산출된다. 제1 케이스의 대기 시간은 사용자 3의 마이그레이션 요청 시간과 마이그레이션 시간에서 사용자 4의 마이그레이션 요청 시간을 제거하여 산출될 수 있다. 수학식 4는 지수 분포를 이용하여 산출된 사용자 i에 대한 평균 서비스 시간 및 마이그레이션 시간을 나타낸다. 지수 분포를 갖는 동일한 두 사건 중에서 어느 것이든지 먼저 사건이 일어날 때까지의 시간 분포도 역시 지수시간분포를 가지며 평균 서비스 시간은 지수 시간의 1/2이 되기 때문에 지수분포에 의해 사용자 i에 대한 평균 서비스 시간 및 마이그레이션 시간은 위와 같다. 수학식 3 및 수학식 4를 정리하면, 수학식 5와 같이 나타난다.

[0063] 제2 케이스는 동시에 둘 이상의 마이그레이션 슬롯을 허용하는 경우이다. 하나의 실시예로서, 제2 케이스는 총 4명의 사용자(User)가 원본 서버(10)에 접속하여 있으며, 원본 서버(10)는 가용 가상 머신 슬롯이 5개(N=5)이고, 마이그레이션 슬롯 개수가 2개(K=2)일 때, 사용자 3(User 3) 및 사용자 4(User 4)가 마이그레이션을 요청한 경우를 나타낸다. 사용자 3이 마이그레이션 요청 시간(S_1^m)이후, 마이그레이션을 시작하여 마이그레이션 시간(tm) 동안 마이그레이션을 수행한다. 사용자 4는 마이그레이션 요청 시간(S_2^m)이후 마이그레이션을 요청하지만, 총 5개의 가용 가상 머신 슬롯이 5개이고, 사용자가 4명이기 때문에, 마이그레이션을 수행할 수 있는 가상 머신 슬롯은 한 개이다. 하지만, 한 개의 가상 머신 슬롯은 사용자 3이 마이그레이션하고 있기 때문에, 사용자 4는 마이그레이션을 수행하지 못하고 대기 시간을 가지게 된다. 301에서는 사용자 3의 마이그레이션이 끝난 이후, 사용자 4의 마이그레이션이 수행된다. 반면에, 302에서는 사용자 3의 마이그레이션이 끝나기 전에, 사용자 1의 서비스가 종료된다. 따라서, 사용자 1이 사용하던 가상 머신 슬롯에 여유가 생겨 사용자 3의 마이그레이션이 종료 여부와 관계없이 사용자 1의 서비스 종료 이후 사용자 4의 마이그레이션을 수행할 수 있다. 이를 통해, 제2 케이스의 대기 시간을 산출할 수 있다. 이는 상술한 수학식 6으로 표현될 수 있다.

[0064] 다음으로, 거절 가능성의 임계값을 설정한다(403). 서비스를 마이그레이션 하는 과정에서 마이그레이션 슬롯의 개수가 늘어날수록 대기 시간은 줄어들 수 있으나, 동시에 거절 가능성이 증가하게 된다. 따라서, 본 발명에 따른 실시간 마이그레이션 자원 할당 방법에서는 대기 시간과 거절 가능성 사이의 협상을 통해 최적의 마이그레이션 슬롯의 개수를 산출한다. 따라서, 협상 조건으로 거절 가능성의 임계값을 설정한다. 거절 가능성의 임계값을 낮게 설정한다면 거절 가능성은 낮아지지만 대기 시간이 증가하게 되고, 거절 가능성의 임계값을 높게 설정한다면 거절 가능성은 높아지지만 대기 시간은 감소하게 된다. 따라서, 서비스 품질(QoS)이나, 제공하는 서비스의 종류에 따라 거절 가능성의 임계값을 설정하게 된다.

[0065] 다음으로, 대기 시간 및 거절 가능성을 고려하여 예측 대기 시간을 산출한다(404). 상술한 401 단계 내지 403 단계에서 산출된 제1 케이스 및 제2 케이스의 대기 시간과 거절 가능성을 고려하여 예측 대기 시간을 산출한다. 예측 대기 시간을 산출하는 공식은 상술한 수학식 8로 표현될 수 있다. 수학식 8에서, $E[T^D]$는 제1 케이스 및 제2 케이스를 모두 포함하는 실시간 마이그레이션 슬롯 할당 장치(100)의 대기 시간을 나타낸다. 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 제1 케이스 및 제2 케이스에 대한 거절 가능성 및 대기 시간을 고려하여 산출된 예측 대기 시간을 최소화하는 최적 K값을 산출한다.

[0066] 다음으로, 최적 마이그레이션 슬롯 개수를 결정한다(405). 404 단계에서 산출된 예측 대기 시간은 대기 시간 및 거절 가능성을 고려하여 산출된다. 따라서, 예측 대기 시간의 값을 최소화할 수 있는 K값을 산출한다. 예측 대기 시간의 값을 최소화할 수 있는 K값을 산출하는 과정은 상술한 수학식 9로 표현될 수 있다. 수학식 8을 통해 산출된 K값은 거절 가능성의 임계값 범위 내에서 예측 대기 시간을 최소화하는 마이그레이션 슬롯의 개수를 나타낸다. 최적 마이그레이션 슬롯 개수를 결정되면, 결정된 마이그레이션 슬롯 개수에 기초하여 마이그레이션 슬롯을 할당한다(406). 본 발명에 따른 실시간 마이그레이션 슬롯 할당 장치(100)는 거절 가능성 및 대기 시간을 고려한 예측 대기 시간을 최소화 할 수 있는 K값을 산출하여 최적의 마이그레이션 슬롯을 산출한다. 그리고, 산출된 결과에 따라 가용한 가상 머신 중에서 마이그레이션 슬롯을 할당한다. 할당된 마이그레이션 슬롯을 통해 서비스를 제공중인 가상 머신을 원본 서버(10)에서 타겟 서버(20)로 마이그레이션 할 수 있다.

[0067] 상술한 내용을 포함하는 본 발명은 컴퓨터 프로그램으로 작성이 가능하다. 그리고 상기 프로그램을 구성하는 코드 및 코드 세그먼트는 당분야의 컴퓨터 프로그래머에 의하여 용이하게 추론될 수 있다. 또한, 상기 작성된 프로그램은 컴퓨터가 읽을 수 있는 기록매체 또는 정보저장매체에 저장되고, 컴퓨터에 의하여 판독되고 실행함으로써 본 발명의 방법을 구현할 수 있다. 그리고 상기 기록매체는 컴퓨터가 판독할 수 있는 모든 형태의 기록매체를 포함한다.

[0068] 이상 바람직한 실시예를 들어 본 발명을 상세하게 설명하였으나, 본 발명은 전술한 실시예에 한정되지 않고, 본 발명의 기술적 사상의 범위 내에서 당분야에서 통상의 지식을 가진자에 의하여 여러 가지 변형이 가능하다.

부호의 설명

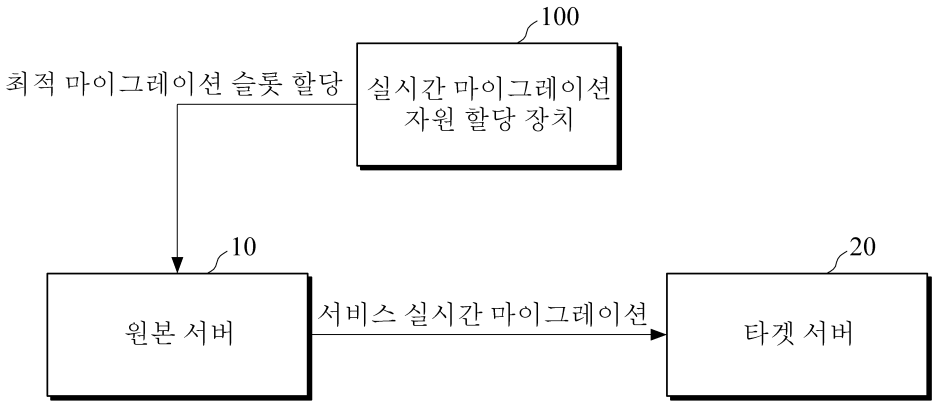
[0069] 10: 원본 서버

20: 타겟 서버

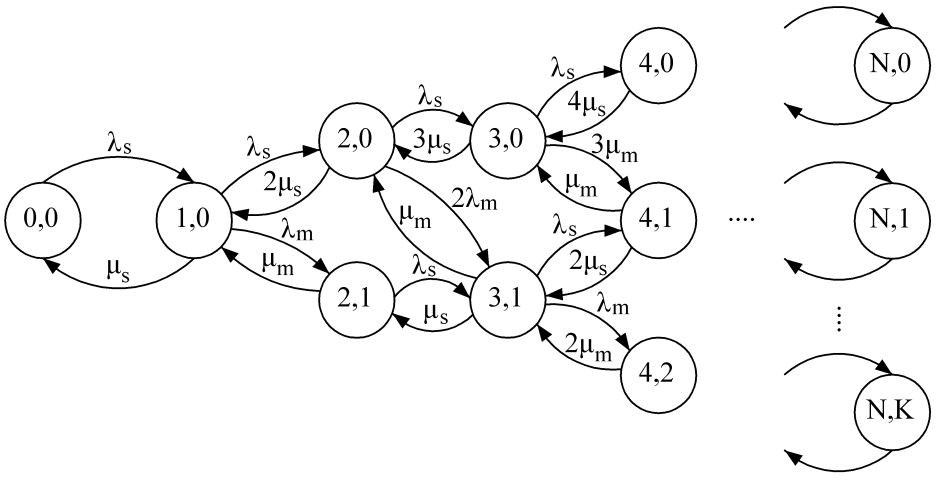
100: 실시간 마이그레이션 슬롯 할당 장치

도면

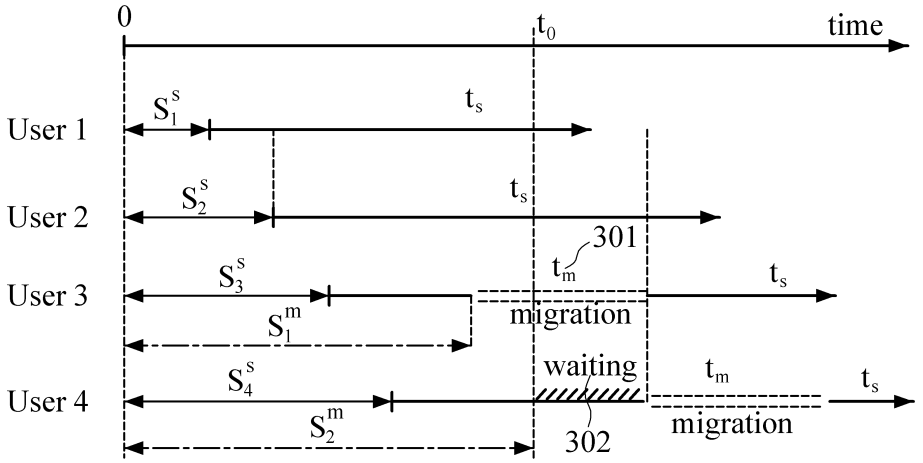
도면1



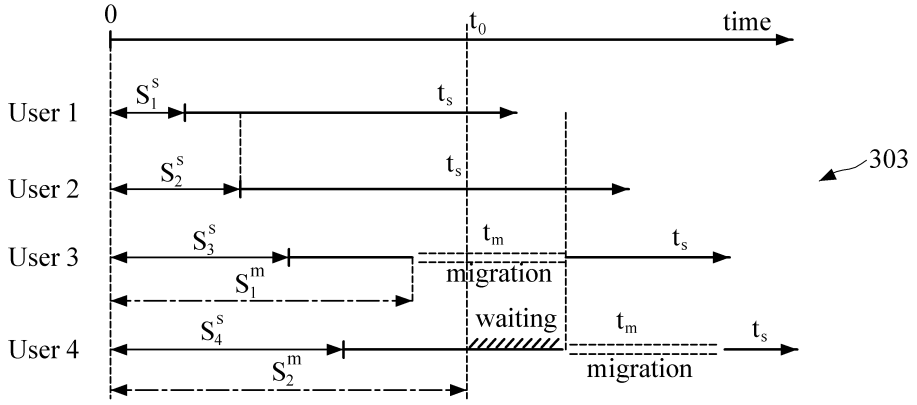
도면2

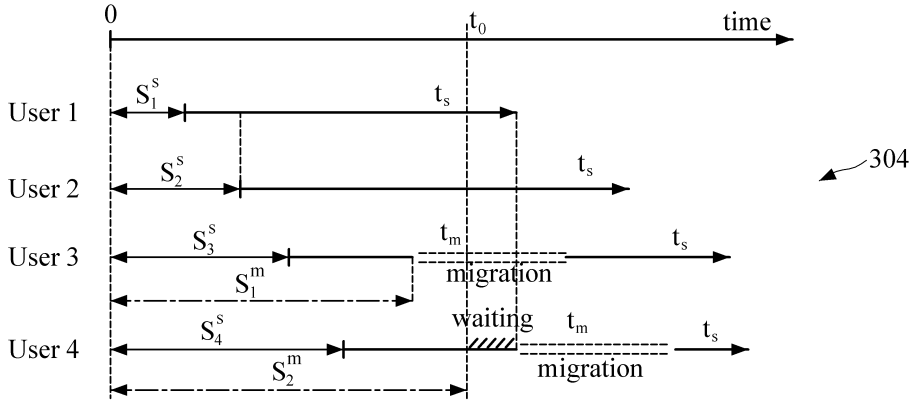


도면3a

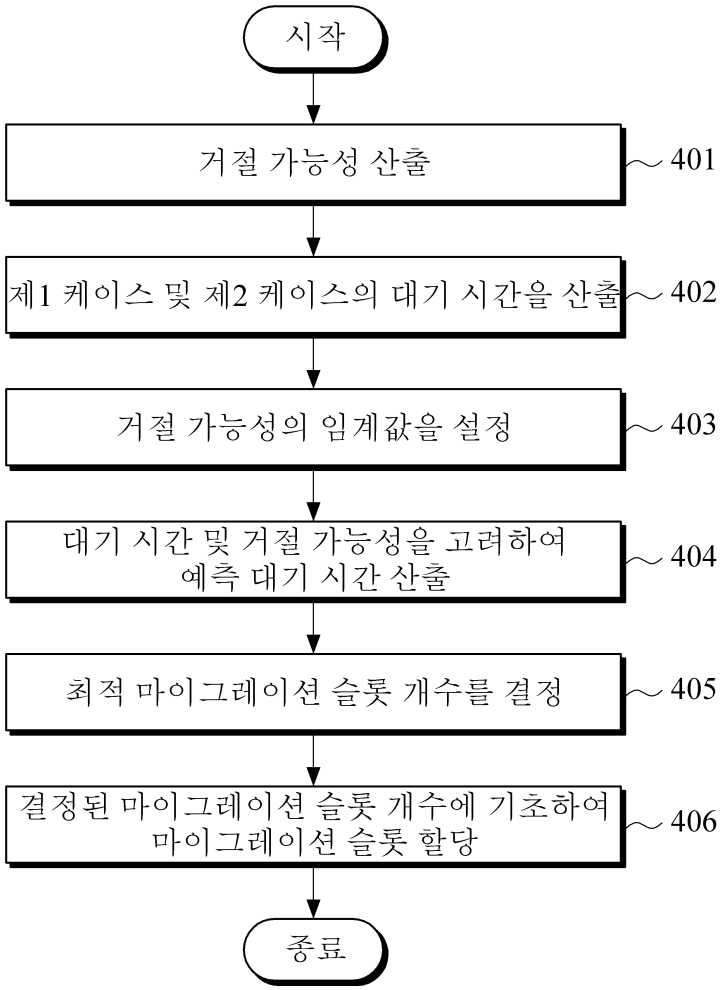

0
t_0
time
User 1
User 2
User 3
User 4
S_1^s
S_2^s
S_3^s
S_4^s
S_1^m
S_2^m
t_s
t_m
301
migration
waiting
302

도면3b


0
t_0
time
User 1
User 2
User 3
User 4
S_1^s
S_2^s
S_3^s
S_4^s
S_1^m
S_2^m
t_s
t_m
migration
waiting
303


0
t_0
time
User 1
User 2
User 3
User 4
S_1^s
S_2^s
S_3^s
S_4^s
S_1^m
S_2^m
t_s
t_m
migration
waiting
304

도면4


시작
거절 가능성 산출
401
제1 케이스 및 제2 케이스의 대기 시간을 산출
402
거절 가능성의 임계값을 설정
403
대기 시간 및 거절 가능성을 고려하여
예측 대기 시간 산출
404
최적 마이그레이션 슬롯 개수를 결정
405
결정된 마이그레이션 슬롯 개수에 기초하여
마이그레이션 슬롯 할당
406
종료