통신 서비스 가용도의 추계적 모델

정회원 함 영 만*, 이 강 원**

Stochastic Model for Telecommunication Service Availability

Young Marn Ham*, Kang Won Lee** *Regular Members*

요 약

본 연구의 주된 목적은 사용자 관점에서 본 통신 시스템 서비스 가용도의 이론적 모델 개발이다. 이를 위하여 호(Call) 도착은 non-homogeneous 포아손 과정의 가정, 그리고 시스템 상태는 CTMC 모델의 가정을 토대로 서비스 가용도의 추계적 모델을 개발하였다. 제시한 모델은 시간에 따라 변하는 호 도착률을 포함하여 사용자 관점에서 본 서비스 신뢰도 모형의 사용자 모델을 효율적으로 나타냈다. 아울러 시스템 자원의 고장 없이도 사용자가 서비스를 받지 못하는 시스템 상태인 운영 고장 상태를 모델에 포함하여 제공자 입장이 아니라 사용자 관점에서 모델을 구축하였다.

Key Words : 서비스 가용도, Non-homogeneous 포아손 과정, Continuous Time Markov Chain

ABSTRACT

The objective of this study is to develop the theoretical model of the telecommunication system service availability from the user perspective. We assume non-homogeneous Poisson process for the call arrival process and continuous time Markov chain for the system state. The proposed model effectively describes the user model of the user-perceived service reliability by including the time-varying call arrival rate. We also include the operational failure state where the user cannot receive any service even though the system is functioning.

I. 서 론

서비스 신뢰도라 불리는 새로운 신뢰도 측도가 시스템과 사용자의 상호 작용(Interaction)을 고려하기 위하여 제안되었다. 기존 신뢰도 측도가 시스템 수준에서 정의되어지는 반면에 서비스 신뢰도는 사용자 관점으로부터 정의된다. 사용자 관점으로부터 서비스 신뢰도를 개괄적으로 살펴보면 다음 (그림 1)과 같다. 이는 사용자 모델, 서비스 모델, 그리고 시스템 모델 3개의 수준으로 구성되어 있다.

먼저 사용자 모델은 서로 다른 다수개의 임무(Task)나 호(Call)의 발생과 같은 사용자의 행위를 포함한다. 위 그림에서 t_1, t_2, ... , t_m은 task의 도착 시점을 나타낸다. 시스템 모델은 전통적 신뢰도 모델에 해당되는 부분으로 하드웨어나 소프트웨어로 구성된 시스템 자원 (위 그림의 R_1, R_2, ..., Rr)의 고장, 수리 그리고 유지 보수 등을 포함한다. 서비스 모델은 시스템 모델과 임의로 도착하는 임무가 요구하는 시스템 자원 (예로 처음 도착한 임무는 R_1과 R_2를 요구하고 두 번째 도착한 임무는 R_1과 R_3 요구)과 같은 임무 요구 사항들로부터 유도된다. 시스템에서 임무 수행은 이 서비스 모델에 의해 표현된다.

* 서울과학기술대학교 IT정책전문대학원 박사과정(산업정보시스템 전공)
** 서울과학기술대학교 IT정책전문대학원 산업정보시스템공학과 교수(교신저자 kwlee@seoultech.ac.kr)
논문번호 : KICS2011-08-372, 접수일자 : 2011년 8월 24일, 최종논문접수일자 : 2011년 12월 20일

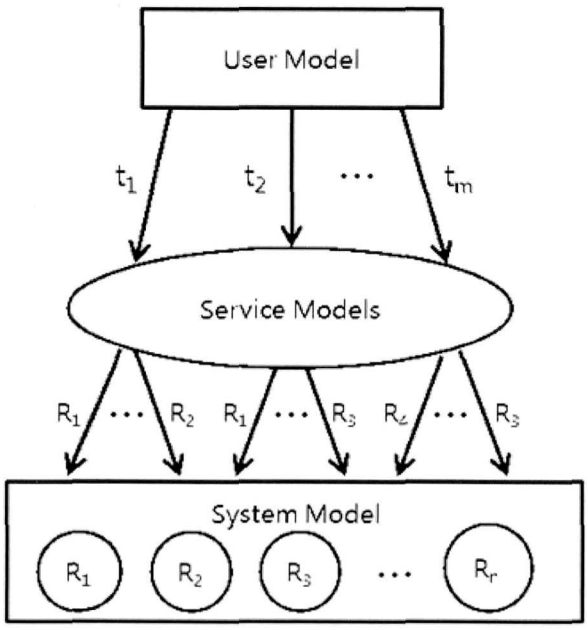

그림 1. 사용자 관점에서의 서비스 신뢰도 모형
Fig. 1. User perceived service reliability model

전통적인 시스템 신뢰도 측도에 비해 서비스 신뢰도는 시스템 자원 신뢰도뿐만 아니라 시스템 이용 빈도와 같은 사용자 이용 형태, 그리고 사용자가 시스템에게 요구하는 일의 특성 등을 복합적으로 고려해야 한다. 사용자는 임의의 서로 다른 시간에 서로 다른 특성을 갖는 다수개의 임무를 발생시킨다. 이때 특정 임무 기간 동안 도착한 모든 임무를 성공적으로 처리할 확률을 나타낼 수 있는 새로운 서비스 신뢰도 측도가 필요하다. 모든 임무가 성공적으로 처리되기 위해서는 먼저 하드웨어나 소프트웨어로 구성된 시스템 자원이 임무 도착 시점에 고장 없이 작동해야 한다. 이와 아울러 도착한 임무가 원하는 서비스 특성 등을 고려하여 시스템이 사용자가 원하는 서비스를 제공할 수 있어야 한다. 통신 분야는 사용자 관점에서 보는 서비스 신뢰도 개념을 필요로 하는 많은 영역들 중의 하나다. 몇 가지 예를 살펴보면 다음과 같다.

1) 현재의 셀에서 다른 셀로 이동하는 무선 단말은 핸드오프(Handoff) 과정을 필요로 한다. 따라서 모든 셀에는 일정 기간 동안에 다수개의 핸드오프 호가 발생한다. 무선망에서 핸드오프 호가 성공적으로 처리되기 위해서는 핸드오프 호가 진입코자 하는 목표 셀 내에 빈 채널이 존재해야 한다. 그렇지 않으면 호는 차단된다. 비록 모든 망 구성요소들이 정상 작동한다고 하더라도 목표 셀 내에 빈 채널이 존재하지 않으면 사용자는 망으로부터 원하는 서비스를 받을 수 없다.

2) VoIP에서 SIP(Session Initiation Protocol) proxy 서버는 호의 수락 여부를 결정하기 위하여 동시 호 (Simultaneous Call)의 최대치와 프로세서 용량을 사용한다. 트래픽 양이 증가하게 되면 많은 양의 호 설정 요구가 있다. 이때 이 숫자가 SIP proxy 서버가 처리할 수 있는 수를 초과하게 되면, 이 임계치를 초과하는 호 시도는 거부된다. VoIP 사용자가 원하는 서비스를 받지 못하는 이유는 SIP 메시지의 교환 중 발생하는 망 구성요소들의 고장일 수 있다. 그러나 보다 결정적 요인은 어떠한 망 요소의 고장 없이 망 정체(Network Congestion)로 인해 사용자가 원하는 서비스를 받지 못하는 경우이다.

3) RFID 시스템에서 리더(Reader)는 먼저 추적(Interrogation) 신호를 송신한다. 만약 태그(Tag)가 리더의 추적 신호 영역 내에 존재하면 태그는 활성화되고 리더와 통신한다. 이때 다수개의 태그가 동시에 리더에게 데이터 전송을 시도하면 충돌(Collision)이 발생한다. 모든 시스템 구성요인들이 고장 없이 정확하게 작동한다고 하더라도 리더는 태그에 있는 정보를 받을 수 없다.

위의 예들은 시스템 자원의 고장 없이도 사용자가 원하는 서비스를 받지 못하는 경우를 보여주고 있는데, 이는 사용자 관점에서 서비스 신뢰도 개념 도입의 중요성을 나타내고 있다.

최근 들어 서비스 신뢰도 측도의 도입 필요성을 강조한 여러 연구들이 진행되었다[1][2][3]. 특히 참고 문헌 [4]는 통신 분야에서 서비스 신뢰도 개념 도입의 필요성을 주장했다. 참고 문헌 [5]는 웹서비스 신뢰도를 제공자 관점과 사용자 관점에서 살펴보았고, 참고 문헌 [6]에서는 QoS(Quality of Service) 관점에서 VoIP 서비스 신뢰도 개념을 도입하였다. 참고 문헌 [7]은 IP backbone 망의 서비스 신뢰도 개념을 제안하였다. 여러 선행 연구에서 서비스 신뢰도에 대한 개념 도입 및 특정 분야에서의 서비스 신뢰도 개념 도입 등 일반적인 연구는 산발적으로 진행 되어온 반면, 서비스 신뢰도에 대한 이론적 모델링 연구는 아직 본격적으로 이루어 지지 않고 있다.

본 연구의 주된 목적은 서비스 신뢰도에 대한 이론적 모델 개발에 있다. 이를 위하여 먼저 서비스 신뢰도 모델에 포함되어야 할 주요 구성 요인을 살펴보았다.

1.1. 임의의 임무 도착 과정

시스템은 주어진 임무 기간 동안 임의로 도착하는 다수개의 임무를 처리해야 한다. 위의 예에서 임무 도착은 인접 셀로 가는 핸드오프 호일 수 있고 SIP proxy 서버에 도착하는 호 시도일 수 있으며 RFID 리더에게 데이터 전송을 시도하는 RFID 태그일 수 있다.

시스템은 사용자가 서비스를 요구할 때만 on 상태에 있으면 된다. 사용자가 시스템을 요구하지 않을 때 시스템의 off 상태는 사용자 관점에서 본 서비스 신뢰도 값에 전혀 영향을 주지 않는다. 다음 (그림 2)에서 t_1과 t_3가 임무 도착시점을 나타낸다면 사용자 관점에서 시스템은 항상 가용하다. 반면에 t_2와 t_4가 임무 도착 시점이라면 시스템은 항상 가용하지 않다. 즉 같은 시스템 자원 하에서 임무 도착 시점에 따라 시스템 가용도 값은 전혀 다른 값을 나타낸다.

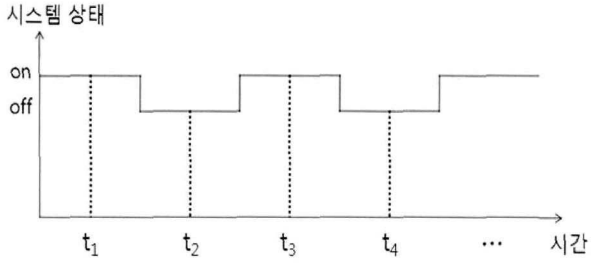

그림 2. 임무 도착 시점에 따른 시스템 가용도
Fig. 2. System availability at task arrival time

1.2. 시스템 상태

전통적인 신뢰도 모델에서 시스템은 on 상태나 off 상태 두 가지 중 하나다. 시스템이 on 상태에 있으면 작동 중이고 off 상태는 작동 불능으로 수리 상태에 있음을 나타낸다. Off 상태는 하드웨어나 소프트웨어로 구성된 시스템 자원의 고장에 기인한다. 사용자 관점에서 보는 서비스 신뢰도 모델에서는 또 다른 고장 상태가 있을 수 있다. 이 상태에서는 모든 시스템 구성 요인들이 고장 없이 정확하게 작동하더라도 사용자가 시스템으로부터 어떠한 서비스도 받지 못한다. 본 논문에서는 이러한 고장을 운영 고장(Operational Failure)으로 명 하였다. 이는 사용자가 원하는 서비스를 시스템 운영상의 문제들로 인해 제공 받을 수 없는 경우를 나타낸다. 예로 다수개의 태그들이 동시에 데이터 전송을 시도하면 RFID 리더는 어떠한 서비스도 받을 수 없다. 그리고 VoIP의 proxy 서버도 과도한 트래픽 정체 하에서는 새로운 호 시도에게 서비스를 제공할 수 없다. Tortorella[4]는 통신 시스템의 고장 형태의 예를 하드웨어나 소프트웨어의 시스템 자원 고장(breakdown 고장)과 성능(performance) 고장(본 연구의 운영 고장)으로 나누어 살펴보았다. <표 1>에서 service accessibility는 호 설정 단계, service continuity는 호 유지 단계, 그리고 service release는 호 해제 단계에 해당된다.

이제 서비스 신뢰도 모델에서 시스템 상태는 크게 3가지로 분류할 수 있다: 작동 중인 on 상태, 시스템 자원 고장을 나타내는 off-1 상태, 그리고 운영 고장을 나타내는 off-2 상태로, 시스템은 시스템 자원 고장이나 운영 고장이 발생할 때까지 임의의 시간 T_{on} 동안 작동 상태에 존재한다. 시스템 자원 고장이 발생하면 시스템은 임의의 시간 $T_{off\text{-}1}$ 동안 off-1 상태에 머문다. Off-1 상태에서의 체류시간은 고장난 요소의 수리시간에 의해 결정된다. 시스템 운영 고장이 발생하면 시스템은 임의의 시간 $T_{off\text{-}2}$ 동안 off-2 상태에 머문다. Off-2 상태에서의 체류시간은 운영 고장 상태에서 벗어나기 위한 시스템 설계의 효율성에 의해 결정된다. 예로 SIP proxy 서버의 트래픽 정체로 인해 새로운 호 시도가 거부되면 트래픽 정체에서 벗어나기 위한 정체 제어가

표 1. 통신 서비스의 고장 형태(Failure Mode) 일례
Table 1. Example of communication service failure modes

	Breakdown 고장	Performance 고장
Service Accessibility	- No ready signal (e.g., dial tone) - No response from far end - Connection to wrong far end	- Excessive ready signal delay - Excessive far end answer delay - Excessive delay loading URL
Service Continuity	- Call cut off - Page permanently stalls during loading	- Poor audio or video quality - Missed deadline for data transfer
Service Release	- Can't hang up	- Billing error

필요하다. 따라서 호 차단 상태인 off-2 상태에서 체류 시간은 얼마나 효율적인 정체 제어 알고리즘을 사용하느냐에 달려있다. 어떤 경우에는 운영 고장 발생률이 임무 도착률과 밀접한 관계가 있을 수 있다. 예로 호 시도가 많으면 많을수록 트래픽 정체는 더욱 심해진다. 이 경우 임무 도착률에 따라 시스템 상태가 변하게 되는데, 이에 대한 수리적 접근 방법은 매우 힘들다. 따라서 본 연구에서는 일단 임무 도착률과 운영 고장 발생률은 서로 독립적이라 가정하였다.

1.3. 측도(Measure)

본 연구에서는 새로운 측도로 서비스 가용도(Service Availability)를 제안하였다. 기존 신뢰도 분석에서 널리 사용되는 측도 중 가용도(Availability)에 해당된다. 가용도는 임의의 시점에 시스템이 on 상태에 있을 확률로 정의된다[8]. 그런데 임무 기간 동안 임의 개수의 임무 도착이 있을 경우에는 임무 도착 시점 마다 시스템이 on 상태에 있어야 한다. 따라서 본 연구에서 제안하는 서비스 가용도는 임무 도착 시점마다 시스템이 항상 on 상태에 있을 확률로 정의되는데, 이는 시스템이 요구되어질 때마다 서비스를 제공할 수 있는 능력을 나타낸다. 예로 (그림 2)에서 임무 도착 시간이 t_1과 t_3라면 서비스 가용도는 1이고, t_1과 t_2라면 0이다.

서론에 이어 2장에서는 모델 개발을 위한 가정과 임무 도착과정, 그리고 CTMC(Continuous Time Markov Chain) 모델을 제시하였다. 3장은 서비스 가용도의 수리적 모델 개발과 수치 예제 들을 제안하였다. 4장에서는 본 논문의 결론을 다루었다.

Ⅱ. 시스템 모델

2.1 가정

1) 시스템에서 발생하는 고장은 시스템 자원 고장 (System Resource Failure)과 운영 고장 (Operational Failure) 두 가지로 분류하였다.

- 시스템 자원 고장; 시스템을 구성하는 하드웨어나 소프트웨어 고장을 나타내며, 이 고장 발생 시 시스템은 사용자가 원하는 서비스를 제공 받을 수 없다.

- 운영 고장; 사용자가 시스템 이용 시 발생하는 운영상의 문제로 인해 사용자가 원하는 서비스를 제공 받을 수 없는 시스템 상태를 나타낸다.

2) 시스템은 주어진 임무기간 동안 도착하는 임의 개수의 임무를 처리해야 한다. 임무의 도착 과정은 평균값 함수(Mean Value Function) $M(t)$를 갖는 non-homogeneous 포아손 과정을 따른다.

3) 매 임무 도착 시점에 시스템은 세 가지 상태 중 하나에 있다; on 상태, off-1 상태, off-2 상태로. On 상태는 시스템의 정상 상태를 나타내며 off-1 상태는 시스템 자원 고장 상태, off-2 상태는 운영 고장 상태를 각각 나타낸다.

4) 시스템 초기에 시스템은 on 상태에 있다. 고장이 발생할 때까지 임의의 시간 T_{on} 동안 작동 상태에 있다. 시스템 자원 고장이 발생하면 임의의 시간 $T_{off\text{-}1}$, 운영 고장이 발생하면 임의의 시간 $T_{off\text{-}2}$ 동안 고장 상태에 머문다. 임의의 시간 동안은 on 상태에, 또 다른 임의의 시간 동안은 off-1 상태나 off-2 상태에 머무르는 사이클을 반복한다.

5) 확률 변수 T_{on}은 전 사이클에 걸쳐 동일한 분포를 갖는다. 이는 $T_{off\text{-}1}$나 $T_{off\text{-}2}$도 마찬가지다. T_{on}의 분포는 임무기간 동안 시스템 자원 고장율과 운영 고장률에 의해 결정된다. $T_{off\text{-}1}$ 분포는 고장난 시스템 자원의 수리율에 의해 결정되며, $T_{off\text{-}2}$ 분포는 운영 고장을 극복하는 시스템의 능력에 의해 결정된다.

6) 임무 기간은 고정된 상수값 T를 갖는다.

7) 임무 도착률과 운영 고장 발생률은 서로 독립적이라고 가정한다.

2.2 Notation

$SA(t_1,t_2,...,t_k)$ 임무 도착시간이 $t_1, t_2, \ldots, t_k$로 주어졌을 때 서비스 가용도

$SA(T)$ 임무 기간 T 동안 서비스 가용도

$f(t_1,t_2,...,t_{N(T)}|N(T)=k)$ $N(T)=k$가 주어졌을 때 임무 도착 시간 $t_1,t_2,...,t_{N(T)}$의 joint pdf

$M(T)$ non-homogeneous 포아손 과정의 평균값 함수

$N(T)$ 임무기간 T 동안 도착하는 임무의 개수

T 임무기간

$z(t)$ 시간 t에서 시스템 상태를 나타내는 지표(Indicator) 변수 (0: on 상태, 1: off-1 상태, 2: off-2 상태)

λ_1 시스템 자원 고장률

λ_2 운영고장 발생률

μ_1 시스템 자원 고장 수리율

μ_2 운영 고장 복구율
$m(t)$ 시간 t에서 임무 도착률
$P_{0i}(t)$ 시간 0에서 시스템 상태가 0일 때 시간 t에 시스템이 상태 i에 있을 확률

2.3 임무 도착 과정

서비스 가용도 값은 임무 도착률 뿐만 아니라 도착률 형태에 따라 서로 다른 값을 갖는다. 예로 임무 기간 초기에 시스템이 on 상태로 시작하는 (그림 3)과 같은 시스템 상태를 가정하자.

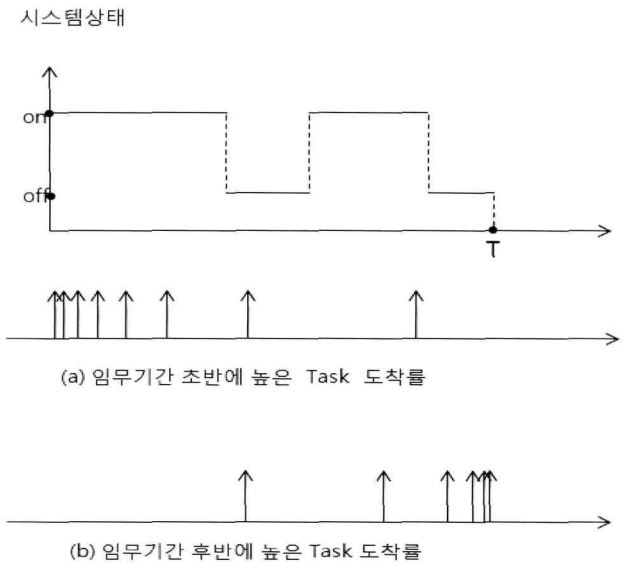

그림 3. 시스템 상태와 임무 도착률
Fig. 3. System state and task arrival rate

이때 (그림 3.a)와 같이 시스템이 on 상태에 있는 임무기간 초반에 임무 도착률이 높으면 비교적 높은 서비스 가용도 값을 갖는다. 반대로 (그림 3.b)와 같이 임무 기간 후반에 임무 도착이 몰리면 앞의 경우에 비해 상대적으로 낮은 서비스 가용도 값을 갖는다. 따라서 본 연구에서는 시간에 따른 임무 도착률의 변화를 고려하기 위하여 임무 도착을 다음과 같은 non-homogeneous 포아손 과정으로 가정하였다. 특히 통신 시스템의 경우 최 번시(Peak Time) 등을 고려하기 위해서 시간에 따른 임무 도착률의 변화를 반드시 고려해야 한다.

$$\Pr[N(T)=k]=\frac{e^{-M(T)}\cdot M(T)^k}{k!}$$

위 식에서 $M(T)$는 non-homogeneous 포아손 과정의 평균값 함수를 나타낸다.

2.4 CTMC(Continuous Time Markov Chain) 모델

2.1 가정에서 언급했듯이 시스템 상태를 on, off-1, off-2 상태 3가지로 분류하였다. 다음 (그림 4)는 시스템의 CTMC 모델을 나타낸다.

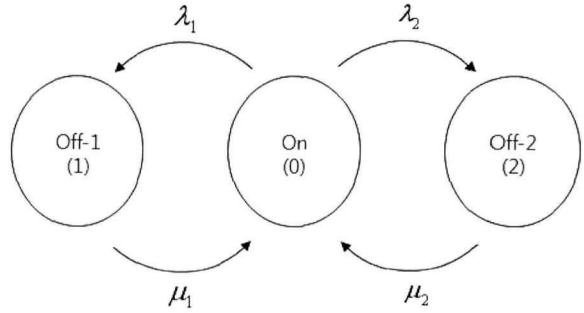

그림 4. CMTC 모델
Fig. 4. CMTC Model

시간 0에 시스템은 on 상태에 있다. 이후 발생률이 각각 λ_1과 λ_2인 시스템 자원 고장이나 운영 고장이 발생할 때까지 on 상태에 머무른다. 시스템 자원 고장이 발생하면 시스템은 off-1 상태에 체류한다. 자원 수리율을 u_1으로 가정하면 off-1 상태에 체류 시간은 평균 $1/u_1$인 지수분포를 따른다. 운영 고장이 발생하면 시스템은 off-2 상태에 빠지는데, 복구율을 u_2로 가정하면 off-2 상태에서 체류시간은 평균 $1/u_2$인 지수 분포를 따른다.

Ⅲ. 서비스 가용도의 추계적 모형

3.1 정의

서비스 가용도는 사용자가 시스템을 요구할 때마다 on 상태에 있을 확률로 정의된다. 이는 주어진 임무 기간 동안 임의로 도착하는 사용자에게 원하는 서비스를 제공할 수 있는 시스템의 능력을 나타낸다.

본 절에서는 먼저 임무 도착 시간이 고정된 상수로 주어질 때 서비스 가용도를 유도한 후, 이를 토대로 임무 도착 과정이 non-homogeneous 포아손 과정을 따를 때의 서비스 가용도를 유도하였다.

3.2 임무 도착 시간이 고정된 상수로 주어질 경우

일정 임무 기간 T 동안 k 개의 임무가 도착하고 그 시각이 $t_1, t_2, \ldots, t_k$로 주어 졌다고 가정하자. 이때 k 개의 도착 시간에 시스템이 on 상태에 있을 확률은 다음과 같이 주어진다.

$$
\begin{aligned}
&SA(t_1, t_2, \ldots, t_k) \\
&= \Pr\{Z(t_i)=0 : i=1,\ldots,k\} \\
&= \Pr\{Z(t_1)=0\} \cdot \Pr\{Z(t_2)=0 | Z(t_1)=0\} \\
&\quad \cdots \Pr\{Z(t_k)=0 | Z(t_{k-1})=0\} \\
&\quad (Markov\ property) \\
&= \Pr\{Z(t_1)=0\} \cdot \Pr\{Z(t_2-t_1)=0 | Z(0)=0\} \\
&\quad \cdots \Pr\{Z(t_k-t_{k-1})=1 | Z(0)=0\} \\
&\quad (time-\text{homogeneous property}) \\
&= P_{00}(t_1) \cdot P_{00}(t_2-t_1) \cdot \cdots \cdot P_{00}(t_k-t_{k-1}) \\
&= \prod_{i=1}^{k} P_{00}(t_i-t_{i-1}),\ \ t_0=0
\end{aligned}
\tag{1}
$$

$P_{00}(t)$를 구하기 위한 과정은 다음과 같다. 먼저 상태 전이 행렬(state transition matrix) $\underline{V}$ 는 아래와 같이 주어진다.

$$
\underline{V} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}
\begin{array}{c}
\begin{array}{ccc} 0 & 1 & 2 \end{array} \\
\begin{bmatrix} -(\lambda_1+\lambda_2) & \lambda_1 & \lambda_2 \\ \mu_1 & -\mu_1 & 0 \\ \mu_2 & 0 & -\mu_2 \end{bmatrix}
\end{array}
$$

한편 전방 미분 방정식(forward derivative equation)은 다음과 같다.

$$
\underline{P}'(t) = \underline{P}(t) \cdot \underline{V} \tag{2}
$$

여기서

$$
\underline{P}(t) = \begin{bmatrix} P_{00}(t) & P_{01}(t) & P_{02}(t) \\ P_{10}(t) & P_{11}(t) & P_{12}(t) \\ P_{20}(t) & P_{21}(t) & P_{22}(t) \end{bmatrix}
$$

위 전방 미분 방정식을 토대로 $P_{00}(t)$를 구하기 위한 미분 방정식을 구하면 다음과 같다.

$$
\begin{aligned}
P'_{00}(t) &= -P_{00}(t) \cdot (\lambda_1+\lambda_2) + P_{01}(t) \cdot u_1 + P_{02}(t) \cdot u_2 \\
P'_{01}(t) &= P_{00}(t) \cdot \lambda_1 - P_{01}(t) \cdot u_1 \\
P'_{02}(t) &= P_{00}(t) \cdot \lambda_2 - P_{02}(t) \cdot u_2
\end{aligned}
$$

이제 $P_{0i}(t), (i=0,1,2)$의 라플라스 변환(Laplace Transform)을 취하면

$$
\begin{aligned}
-1+\theta L_0(\theta) &= -L_0(\theta) \cdot (\lambda_1+\lambda_2) + L_1(\theta) \cdot u_1 + L_2(\theta) \cdot \mu_2 \\
\theta L_1(\theta) &= L_0(\theta) \cdot \lambda_1 - L_1(\theta) \cdot \mu_1,\ \ L_1(\theta) = \frac{\lambda_1}{\theta+u_1} \cdot L_0(\theta) \\
\theta L_2(\theta) &= L_0(\theta) \cdot \lambda_2 - L_2(\theta) \cdot \mu_2,\ \ L_2(\theta) = \frac{\lambda_2}{\theta+u_2} \cdot L_0(\theta)
\end{aligned}
$$

위 식을 토대로 $L_0(\theta)$, $L_1(\theta)$, $L_2(\theta)$를 구하면

$$
L_0(\theta) = \frac{(\theta+u_1)(\theta+u_1)}{\theta\{\theta^2+\theta(\lambda_1+\lambda_2+u_1+u_2)+\lambda_1u_1+\lambda_2u_1+u_1u_2\}}
$$

$$
L_1(\theta) = \frac{\lambda_1(\theta+u_2)}{\theta\{\theta^2+\theta(\lambda_1+\lambda_2+u_1+u_2)+\lambda_1u_1+\lambda_2u_1+u_1u_2\}}
$$

$$
L_2(\theta) = \frac{\lambda_2(\theta+u_1)}{\theta\{\theta^2+\theta(\lambda_1+\lambda_2+u_1+u_2)+\lambda_1u_1+\lambda_2u_1+u_1u_2\}}
$$

이제 $P_{00}(t)$는 위 식 $L_0(\theta)$를 도태로 구할 수 있다.

예로 $\lambda_1=\lambda_2=1$, $u_1=u_2=2$ 라 가정하면

$$
\begin{aligned}
L_0(\theta) &= \frac{(\theta+2)^2}{\theta\{\theta^2+6\theta+8\}} \\
&= \frac{(\theta+2)^2}{\theta(\theta+4)(\theta+2)} \\
&= \frac{\theta+2}{\theta(\theta+4)}
\end{aligned}
$$

그러므로

$$
P_{00}(t) = \frac{1}{2} + \frac{1}{2} \cdot e^{-4t}
$$

로 주어진다.

3.3 임무 도착 시각이 확률 변수일 경우

2.1 가정의 2)에서 언급한 대로 임무 도착 과정이 non-homogeneous 포아손 과정을 따를 경우 서비스 가용도는 다음의 과정을 거쳐 유도할 수 있다.

- 임무 수행 기간 T 동안에 임무 도착 개수 $N(T)$가 k개로 주어졌을 때 도착 시간의 조건부 분포(Conditional Distribution)에 관해서 $SA(t_1, t_2, \ldots, t_{N(T)})$의 평균값을 구한다.
- 이후에 non-homogeneous 포아손 분포를 따르는 $N(T)$에 관해서 평균값을 구한다.

즉

$$SA(T) = E\{E\{SA(t_1, t_2, \ldots, t_{N(T)}) : N(T) > 0\}\} + \Pr\{N(T) = 0\} \cdot 1$$

$$= \sum_{k=1}^{\infty} E\{SA(t_1, t_2, \ldots, t_{N(T)} : N(T) = k\} \cdot \exp[-M(T)] \cdot \frac{M(T)^k}{k!} + \exp[-M(T)]; \quad (3)$$

$$E\{SA(t_1, t_2, \ldots, t_{N(T)} : N(T) = k\} = \int \cdots \int_{\Omega} SA(t_1, t_2, \ldots, t_k) \cdot f(t_1, t_2, \ldots, t_{N(T)} : N(T) = k) dt_1 \cdots dt_k,$$

$$\Omega \equiv 0 \le t_1 < t_2 < \cdots < t_k \le T \quad (4)$$

여기서

$f(t_1, t_2, \ldots, t_{N(T)} : N(T) = k)$ 는 $N(T) = k$가 주어졌을 때 임무 도착시간인 $t_1, t_2, \ldots, t_{N(T)}$의 joint pdf 인데 다음 식으로 주어진다[9].

$$f(t_1, t_2, \ldots, t_{N(T)} : N(T) = k) = k! \cdot \prod_{i=1}^{k} \frac{m(t_i)}{M(T)^k}, \Omega \quad (5)$$

3.4 Numerical Examples

3.4.1 파라메타 값

서비스 가용도의 수치 예제를 들기 위하여 각 파라메타들을 다음과 같이 가정하였다.

- 임무 도착률: 임무 도착률은 (그림 5)와 같이 다음의 세 가지 형태로 가정 하였다.
  - $m_1(t)$ = 0.05t
  - $m_2(t)$ = 0.25
  - $m_3(t)$ = 0.5-0.05t

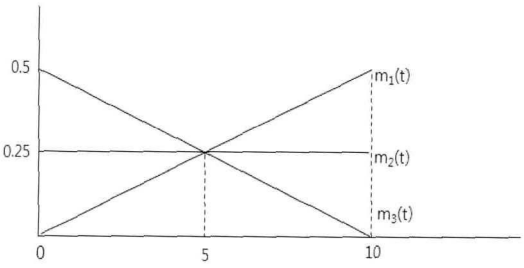

그림 5. 임무 도착률
Fig. 5. Task arrival rate

- 임무 기간은 T = 10 시간으로 가정하였다. 이는 위의 세 가지 형태 임무 도착률 하에서는 평균 2.5개의 임무 도착이 발생한다.
- 시스템 자원 고장 발생률은 λ_1 = 0.2/시간 과 운영 고장 발생률은 λ_2 = 0.3/시간 으로 가정하였다.
- 고장난 시스템 자원 수리율은 μ_1 = 5/시간 과 운영 고장에서 복구율은 μ_2 = 2/시간 으로 가정 하였다.
- 시스템 초기에 시스템은 on 상태에 있다.

3.4.2. 결과 및 분석

이제 위의 세 가지 도착률 형태 하에서 서비스 가용도를 구하면 다음 <표 2>와 같다.

표 2. 세 가지 도착률 하에서 서비스 가용도
Table 2. Service availability under the three types of task arrival rate

임무 도착률	서비스 가용도
m1(t) = 0.05t	0.563
m2(t) = 0.25	0.641
m3(t) = 0.5-0.05t	0.746

먼저 $m_2(t)$ = 0.25는 임무 기간인 T = 10시간 동안 평균 2.5개의 임무가 도착하는 포아손 과정을 나타낸다. 이때 서비스 가용도 값은 0.641이다.

$m_3(t)$ = 0.5-0.05t 하에서는 임무 기간 초기에 임무 도착률이 높고, 뒤로 갈수록 도착률이 낮아진다. 그런데 시스템 상태는 임무 초기에 on 상태에 있기 때문에 임무 들이 시스템이 on 상태에 있을 때 도착할 가능성이 높다(그림 3.a 참조). 따라서 서비스 가용도 값은 포아손 분포를 따르는 $m_2(t)$ = 0.25 하에서 보다 높은 값을 갖게 된다. 실제로 <표 2>에서 보듯이 서비스 가용도 값은 0.746으로 0.641보다 높은 값을 갖는다.

$m_1(t)$ = 0.05t는 $m_3(t)$ = 0.5-0.05t와 반대로 임무 기간 초기에 임무 도착률이 낮고, 뒤로 갈수록 도착률이 높아진다. 그런데 시스템 상태는 임무 초기에 on 상태에 있기 때문에 임무 들이 시스템이 on 상태에 있을 때 도착할 가능성이 $m_2(t)$나 $m_3(t)$에 비해 비교적 낮다(그림 3.b 참조). 따라서 서비스 가용도 값은 $m_2(t)$나 $m_3(t)$ 하에서 보다 낮은 값을 갖게 된다. 실제로 <표 2>에서 보듯이 서비스 가용도 값은 0.563으로 가장 낮다.

위의 예에서 볼 수 있듯이 일정 기간 동안 평균 임무 도착 개수가 2.5 개로 같더라도 도착률 형태에 따라 서비스 가용도는 0.641, 0.746 그리고 0.563으로 서로 다른 값을 갖는다.

한편 임무 도착을 고려하지 않은 (그림 3) CTMC 모델의 극한 확률(Limiting Probability) $\lim_{t\to\infty} P_{oi}(t) = \pi i$ 는 다음 식과 같이 유도된다.

$$(\pi_0,\ \pi_1,\ \pi_2)\begin{bmatrix} -(\lambda_1+\lambda_2) & \lambda_1 & \lambda_2 \\ \mu_1 & -\mu_1 & 0 \\ \mu_2 & 0 & -\mu_2 \end{bmatrix} = (0,\ 0,\ 0)$$

위 식을 토대로 구하면

π_0 = 0.84

π_1 = 0.034

π_2 = 0.126

본 연구에서 제안한 측도인 서비스 가용도는 임의 개수의 임무 도착 시점마다 시스템이 항상 on 상태에 있을 확률로 정의되는데, 이는 전통적인 가용도 측도에 비견된다. 기존의 가용도는 사용자의 이용 빈도를 고려하지 않고, 임의의 시점에 시스템이 on 상태에 있을 확률을 나타낸다. 사용자 특성인 임무 도착률을 고려하지 않고 시스템 상태만을 고려하여 유도한 CTMC 모델의 극한 확률 π_0 = 0.84 가 기존의 가용도 값인데 <표 1>에 제시한 임무 도착률에 따른 서비스 가용도와 많은 차이가 남을 볼 수 있다. 또 <표 1> 내에서도 도착률 형태에 따라 서비스 가용도에 많은 차이가 있음을 알 수 있다. 이는 서비스 제공자 입장에서가 아니라 사용자 입장에서 가용도 측도를 구하기 위해서는 사용자 이용 특성을 반드시 고려해야 한다는 사실을 보여준다.

서론에서 언급했듯이 임무 도착률과 운영 고장 발생율은 서로 밀접한 관계가 있다. 예로 통신 시스템에서 발생할 수 있는 전형적인 운영 고장의 예는 트래픽 정체로 인해 사용자가 원하는 서비스를 제공받지 못하는 경우이다. 이때 트래픽 정체는 호 도착률과 직접적인 관계가 있기 때문에 정체에 따른 운영 고장의 수리적 모델을 유도하기 위해서는 호 도착률에 따라 운영 고장 발생율이 변해 나가는 모델이 필요하다. 본 연구에서는 수리적 모형을 유도하기 위하여 임무 도착률과 운영 고장 발생률은 서로 독립적이라고 가정하였는데, 만약 운영 고장 발생률 λ_2가 임무 도착률 $m(t)$에 따라 변한다고 하면 이에 대한 일반적인 수리적 분석은 매우 어렵다. 다만 이를 해결하기 위한 하나의 특수한 예를 들면 다음과 같다. 시간 t_i에 도착한 임무의 수행 시간이 δ 라 할 때 다음 임무가 $t_i + \delta$ 내에 도착하면 이를 운영 고장으로 가정하자. 따라서 시간 t_{i+1}에 서비스가 가용하기 위해서는 $t_i + \delta < t_{i+1}$ 이어야 하고, 서비스 가용도의 식 (4)의 적분 구간은 $\{t_i + \delta < t_{i+1}, i = 1, 2, \ldots, k\}$로 바뀌어야 한다. δ = 0.2로 가정하면 위의 예제 3가지 임무 도착률 $m_1(t)$, $m_2(t)$, 그리고 $m_3(t)$ 하에서 서비스 가용도는 각각 0.44, 0.52, 그리고 0.61로 나타났다. 위 예는 하나의 특수한 경우로 이에 대한 일반적인 수리 분석은 추후 연구 과제로 남겨둔다.

Ⅳ. 결 론

본 연구의 주된 목적은 통신 시스템의 사용자 관점에서 본 서비스 가용도의 이론적 모델 개발이다. 이를 위하여 임무 도착은 non-homogeneous 포아손 과정의 가정, 그리고 시스템 상태는 CTMC 모델의 가정을 토대로 서비스 가용도의 추계적 모델을 개발하였다. 제시한 모델은 시간에 따라 변하는 임무 도착률을 포함하여 (그림 1) 사용자 관점에서 본 서비스 신뢰도 모형의 사용자 모델을 효율적으로 나타냈다. 아울러 시스템 자원의 고장 없이도 사용자가 서비스를 받지 못하는 시스템 상태인 운영 고장 상태를 모델에 포함하여 제공자 입장이 아니라 사용자 관점에서 모델을 구축하였다.

본 연구에서 개발한 추계적 모델은 다음의 측면에서 계속 보완될 여지를 남겨두고 있다.

첫째, 제시된 모델은 (그림 1) 사용자 관점에서 본 서비스 신뢰도 모형 중 서비스 모델을 효율적으로 나타내지 못한다. 서로 다른 시간에 도착한 사용자는 일의 특성이 서로 상이할 수 있고, 이로 인해 요구하는 시스템 자원이 서로 다를 수 있다. 예로 처음 도착한 임무는 시스템 자원 R_1과 R_2를 요구하고, 두 번째 도착한 임무는 R_1과 R_3를 요구할 수 있다. 각 임무 도착 시점에 시스템이 on 상태에 있을 확률은 요구하는 시스템의 자원에 따라 서로 다르기 때문에, 이를 효율적으로 나타낼 수 있는 새로운 모델이 필요하다.

둘째, 임무 도착률과 운영 고장 발생률은 서로 밀접한 관계가 있을 수 있다. 통신 시스템에서 정체(Congestion)는 운영 고장의 주요 요인 중 하나다. 그런데 많은 호 시도 (즉 높은 임무 도착률)는 트래

픽 정체 상태를 야기한다. 따라서 임무 도착률에 따라 운영 고장 발생률이 변하게 된다. 이를 효율적으로 나타낼 수 있는 수리적 모형은 매우 어려운 문제로 이를 효율적으로 다룰 수 있는 새로운 모델이 필요하다.

참 고 문 헌

[1] D. Wang and K.S. Trivedi, "Modeling User-Perceived Reliability Based on User-behavior Graphs," WSPC, April 2009.

[2] M. Tortorella, "Service Reliability Theory and Engineering, I: Foundations," Quality Technology and Quantitative Management, Vol.2, No. 1, 2005, pp.1-16.

[3] W. Xie, H. Sun, Y. Cao, and K. S. Trivedi, "Modeling of User Perceived Webserver Availability, IEEE International Conference on Communication, May 2003.

[4] M. Tortorella, "Service Reliability Theory and Engineering, II: Models and Examples," Quality Technology and Quantitative Management, Vol.2, No. 1, 2005, pp.17-37.

[5] M. Kaaniche, K. Kanoun, and M. Martinello, "A User-Perceived Evaluation of a Web Based Travel Agency," International Conference on Dependable Systems and Networks, June 2003, pp. 709-718.

[6] C. R. Johnson, Yakov Kogan, Y. Levy, F. Saheban, and P. Tarapore, "VoIP: A Service Provider's Perspective," IEEE Magazine, July 2004, pp. 48-54.

[7] R. Keralapura and S. Bhattacharyya, "Service Availability: A New Approach to Characterize IP Backbone Topologies, 12th IEEE International Workshop on Quality of Service, June 2004, pp. 232-241.

[8] E. E. Lewis, Introduction to Reliability Engineering: John Wiley&Sons, 1987.

[9] E. Parzan, Stochastic Processes: Holden Day.

함 영 만 (Young Marn Ham) 정회원

1994년 2월 서울산업대학교 전자공학과(공학사)

1996년 8월 한양대학교 공학대학원 전자통신공학전공(공학석사)

2012년 현재 서울과학기술대학교 IT정책전문대학원 박사과정(산업정보시스템전공)

2012년 현재 영인 IT 기술사 사무소 대표

정보통신기술사, 전자응용기술사, 전자기기기능장, 대한민국 무선통신 명장

<관심분야> 정보통신, 무선통신, 방송통신

이 강 원 (Kang Won Lee) 정회원

1980년 서울대학교 공과대학 산업공학(공학사)

1982년 서울대학교 대학원(공학석사)

1985년 Kansas State Univ. U.S.A 공학박사(산업공학)

1985년 한국 전자통신연구원 TDX 개발단

2012년 현재 서울과학기술대학교 산업정보시스템공학과 교수

<관심분야> 정보통신, 품질 및 신뢰성, O.R.