응답시간기반 웹 서비스 가용성 측정방안

박 상 근† · 최 덕 재††

요 약

E-Business의 고객과 서비스 제공자들은 사업을 성공으로 이끄는 최선의 방법으로 SLA(Service Level Agreement)를 제공해야 하며, SLA에 포함되어야할 기술적 파라미터로 서비스 가용성을 최우선으로 선정하였다. 이러한 결과들은 웹을 기반으로 한 서비스 제공자들에게 웹 서비스 가용성의 관리가 매우 중요함을 말해준다. 응용서비스 가용성은 사용자의 기대 성능을 유지하면서 제공되는 서비스의 비율로 정의된다. 하지만, 현재까지 웹 서비스 가용성을 측정하는 도구는 단순하게 웹 서비스의 중단 여부만을 판단하기 때문에 사용자의 기대 성능에 대한 요구를 반영하지 못하고 있다. 본 논문에서는 응답시간에 기반한 웹 서비스의 성능저하의 경우를 고려하여 가용성을 측정하고, 불가용을 초래하는 요소가 무엇이며, 그 원인이 무엇인지 밝혀내는 방안을 제시한다. 연구의 결과로 측정도구 WebSerAvail를 개발하였다.

Response Time-based Web Service Availability Measurement Method

Sang Kun Park† · Deok Jai Choi††

ABSTRACT

E-Business providers and customers have chosen the service availability as the most important technical parameter which should be included in their SLA to succeed in their business. This means that web service availability management is crucial to the web-based service providers. Application availability is originally defined as a measure of the fraction of time during a defined period when the service provided is deemed to be better than user expectation of service performance. But, because most web service availability measurement tools simply monitor disconnected state, they do not satisfy user's expectation of extended availability concept. In this paper, We propose the web service availability measurement method which supports extended availability concept. It takes account of performance degradation of web service based on response time, and determines what is the cause of unavailability of the service. We developed a measurement tool, WebSerAvail.

키워드 : 가용성(Availability), 웹 서비스(Web Service), 응답시간(Response Time), SLA, WebSerAvail

1. 서 론

네트워크 인프라의 QoS 제공과 인터넷을 기반으로 한 전자상거래의 활성화로 인하여 다양한 인터넷 ASP(Application Service Provider)들이 등장하고 있다. ASP는 응용서비스를 개발하고 설치, 관리하는 기능을 수행하고 있지만, 사용자가 원하는 서비스의 요구를 만족시키기 위해서는 NSP(Network Service Provider), HSP(Hosting Service Provider), MSP(Management Service Provider) 등의 Outsourcing 서비스 제공자와의 관계가 반드시 필요하다. 현재의 인터넷 환경은 IT 인력 부족 해결, 저가의 고품질 자원 보장, 빠른 서비스 제공, 위험 부담 공유 등의 Outsourcing이 주는 혜택과 웹기반의 중소형 e-Business 업체의 대거 등장 등으로 웹기반 응용서비스에 대한 요구가 증가하고 있다.

ASP 사업자가 사업에 성공하기 위해서는 사용자가 만족할 수 있는 서비스 품질을 유지해야 하며, 더불어 이들 응용서비스들에 대한 품질을 평가하는 방법들이 제시되어야 한다. 특별히 이들 응용서비스의 기반 서비스가 되는 웹 서비스의 성능에 대한 평가는 중요한 이슈가 될 것이다. 인터넷 망은 서비스 질을 보장할만한 신뢰성을 제공하지 못하기 때문에 웹 서버의 성능이 우수하다 하더라도, 복잡한 네트워크 환경이나 계층화된 서비스 등의 여러 원인으로 인해 고객은 일관된 서비스 품질을 받기가 어려운 실정이다.

Outsourcing에 관련된 파트너들 사이의 관계를 분명하게 정의하고, 제공하는 서비스의 신뢰성을 보장하기 위하여 이들 상호간의 서비스 수준 계약(SLA : Service Level Agreement)이 필요하다. SLA는 서비스 제공자(SP : Service Pro-

† 준 회 원 : 전남대학교 대학원 전산학
†† 종신회원 : 전남대학교 전산학과 교수
논문접수 : 2002년 8월 23일, 심사완료 : 2002년 10월 31일

vider)와 그 서비스를 이용하는 고객(Customer)사이의 책임과 의무를 정의한다. 여기에는 관련 서비스 제품, 측정범위와 방식, 고장이나 성능 및 결과에 대한보고 방식, 서비스에 대한 질적 표준 등이 포함된다.

ASPIC(ASP Industry Consortium)에 의하면, 서비스 제공자와 고객 각각 55%와 59%가 e-Business를 성공으로 이끄는 최선의 방법으로 SLA 제공을 들고 있으며, SLA에 포함되어야할 기술적 파라미터로는 사용자 응답자의 78%가 응용서비스 가용성(Application Availability)을 최우선으로 선정하였다[12]. 이러한 결과들은 웹을 기반으로 한 ASP들에게 웹 서비스 가용성(WSA : Web Service Availability)의 측정과 관리가 중요함을 말해준다.

지금까지 사용자 또는 클라이언트 입장의 웹 성능을 측정하기 위하여 다양한 연구와 테스트들이 진행되고 있으며, 더불어 웹 서비스 가용성에 관심이 집중되고 있다. 최근의 응용 서비스 가용성은 사용자의 기대 성능을 유지하면서 서비스에 접근할 수 있는 비율로 정의되면서, 응용계층 응답시간까지 포함한 확장된 의미로 해석되고 있다.

현재까지 웹 서비스 가용성을 측정하는 도구들은 가용성 개념자체를 장애에 의해 서비스가 중단되지 않은 상황으로 한정하고, HTTP 요청에 대하여 서버로부터 정상적인 응답이 오는지의 여부만을 단순하게 측정하고 있다. 여기에는 2가지 문제가 있다. 첫째로, 불가용성(Unavailability)을 초래하는 요소가 무엇이며, 그 원인이 고객측에 있는지, 제공자측에 있는지를 밝히지 못하기 때문에 책임규명의 어려움이 있다. 둘째로, 사용자의 입장은 단지 장애의 경우만이 아니라 웹 서비스의 지나친 성능 저하의 경우에도 가용할 수 없는 것으로 간주한다는 것을 고려하지 못하고 있다.

본 논문에서는 대표적인 응용 서비스인 웹 서비스에 대하여 위 두 가지 문제점을 고려한 가용성 측정 방안을 제시한다. 장애의 관점뿐만 아니라 응답시간에 기반한 웹 페이지 전송 성능의 관점에서 웹 서비스의 가용성을 측정하며, 가용성에 영향을 미친 요소들을 응용 서비스 수준에서 정의하고, 이를 측정 분석하여 그 원인을 진단한다. 이를 위해서 측정 구조를 모델링하고, 웹 서비스에 연관된 요소들로 구성된 Web Quality Factor(WQF)를 정의하였다. 분석 알고리즘을 바탕으로 SLA에 명시된 임계치를 초과하는 불가용성에 가장 크게 영향을준 요소를 판별할 수 있는 근거를 제시하였다. 연구의 결과로 측정도구 WebSerAvail를 개발하였다.

측정도구 WebSerAvail를 이용하는 서비스 제공자는 사용자에게 서비스의 성능저하나 서비스 중단에 대한 원인을 제공함으로써 서비스에 대한 신뢰성을 높일 수 있게 될 것이다. 또한 고객과 서비스 제공자들 사이에 서비스 성능저하나 장애로 인해 발생한 문제에 대한 명확한 책임규명을 가능하게 하여 서비스 관리에 대한 투명성을 확보할 수 있게 될 것이다. 특별히 웹 서비스 제공자에게는 더 많은 사용자를 확보하고, 운영을 위한 비용 절감을 가능하도록 하여 다른 사업자들에 비해 경쟁력을 갖춘 서비스 제공자로 남는데 기여할 것이다.

2. 관련 연구

2.1 웹 성능 측정

지금까지 웹 성능을 향상시키기 위하여 TCP, HTTP 등의 관련 프로토콜 수정이나 클라이언트, 서버, Proxy의 로그, 벤치마크, 사이트의 재 디자인과 같은 다양한 연구들이 진행되어 왔으며, 이러한 연구들은 사용자가 느끼는 웹 성능을 향상시키는데 많은 도움을 제공하였다. A. Habib은 웹 페이지 다운로딩 시간을 DNS Time, Connection Setup Time, Time to Get Frist Byte, Downloading Time 등으로 구별하여 다운로딩 시간의 지연이 커진 원인을 DNS, 서버, 링크 및 라우터 등으로 구별하여 분석하였다[1]. B. Krishnamurthy는 클라이언트가 700개 이상의 상용 웹 서버를 방문하여 측정한 End-to-End 지연에 대하여 영향을 미치는 요소들을 HTTP 프로토콜, 캐싱, Multi-sever Content, Request 크기 등의 환경에서 분석하였다[2]. J. Charzinski는 지연과 전송율을 이용하여 Fun Factor를 정의하고, HTTP 성능과의 관계를 분석하였다[3]. J. Mogul과 V. Padmanabhan은 네트워크 관점에서 웹 트랜젝션에 대한 사용자가 느끼는 지연에 대한 이슈들을 다루었다[4, 5]. Nina Bhatti는 사용자 기대와 웹 작업의 영향에 관한 테스트를 통하여 지연에 대한 사용자 Tolerance를 추정하여 사용자 요구사항에 기반한 웹 서버의 설계방안을 다루었다[6]. Mimkka Koletsou는 Medusa Proxy 툴을 이용하여 사용자의 요청을 받아 다양한 서비스 형태에 대한 사용자 입장의 지연을 분석하였다[7].

이들 연구들은 사용자가 느끼는 성능을 측정하기 위하여 클라이언트 측에서의 지연 결과에 초점을 맞추고 있으며, 서비스에 관련된 요소들을 정의하여 지연이 발생한 원인에 대하여서 분석하고 있다. 하지만, 웹 서비스의 가용성에 영향을 미치는 요소가 클라이언트에 있는 지를 판별할 수 있는 근거를 제공하지 못하고 있다. 그러므로, DNS 질의를 보내기 이전에 소켓을 열어 서버에 최초의 TCP 패킷을 전송하기까지의 과정에서 클라이언트가 미치는 영향을 규명하여야 할 필요가 있다. 또한, 지금까지의 연구는 HTTP 메시지를 주고받는 과정을 단순히 네트워크의 영향으로 전제하면서 클라이언트나 서버가 미치는 영향을 고려하지 못하고 있다. 즉, HTTP 메시지를 주고받는 과정은 TCP 패킷을 생성하고 적절한 응답을 보내기 위하여 클라이언트와

서버가 어느 정도 영향을 미치고 있기 때문에 이를 반영할 수 있는 방안이 필요하다.

2.2. 가용성 측정

일반적인 가용성은 전체 서비스 제공시간에 대한 서비스 가용시간의 백분율을 나타낸다. B. Chandra와 M. Kalyanakrishnan은 종래의 이러한 개념을 바탕으로 인터넷 불가용성을 초래하는 네트워크 인프라의 원인을 분석하려는 연구를 수행하였다[8, 10]. 특히 B. Chandra는 WAN환경에서 네트워크의 오류가 서비스 가용성에 어떻게 영향을 미치는지를 고장위치, 고장기간, 고장율의 핵심 파라미터를 이용한 분석모델을 통하여 제시하였다[8]. Wei Xie는 Markov ReGenerative Process(MRGP)에 기반한 모델을 이용하여 인터넷 불가용성을 초래하는 원인과 이에 대한 온라인 사용자의 행위를 분석하여 사용자 관점에서의 웹 서비스의 불가용성을 측정하였다[11].

지금까지 연구에서 다루어진 가용성 개념은 응용 서비스의 관점을 반영하지 못하고 있다. 최근의 응용서비스 가용성 개념은 사용자의 기대 성능을 유지하면서 서비스에 접근할 수 있는 비율로 기존의 가용성의 개념에 응용수준의 응답시간까지 포함된 의미로 해석되고 있다. TMF(TeleManagement Forum)에서도 QoS 임계치를 초과하지 않는 성능을 유지하는 서비스 접근 비율로서 해석하고 있으며, SAP(Service Access Point)에 대하여 주어지는 SDF(Service Degradation Factor)의 비율을 적용하여 서비스 가용성을 정의하였다[9]. 본 논문에서는 최근의 응용서비스 수준에서의 가용성 개념을 적용하여 웹 서비스의 가용성을 측정하는 방안을 제시하고자 한다.

3. 웹 서비스 가용성 측정 모델

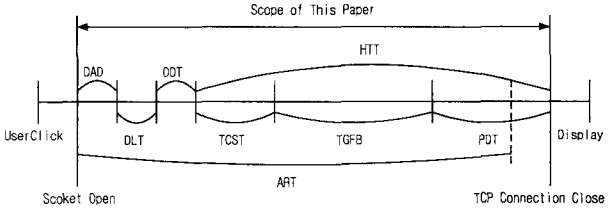
사용자 관점에서 웹 페이지를 다운로딩하는데 걸리는 시간은 사용자가 클라이언트에 원하는 URL을 요청한 시간에서 서버로부터 페이지 전송이 완료되어 클라이언트가 페이지 Display를 완료한 시간까지를 의미한다. 하지만, 대표적인 클라이언트 응용인 브라우저의 경우 사용자와의 인터페이스 처리나 HTML Parser 등의 Display에 대한 처리가 매우 빠른 우수한 성능을 갖고 있다. 이러한 성능을 유지하는 클라이언트 시뮬레이터를 구현하여 웹 서비스 가용성을 측정하는 것은 매우 복잡한 작업이 될 것이다. 그러므로 본 연구에서는 측정범위를 클라이언트가 HTTP 요청 메시지를 보내기 위해 소켓을 연 시간으로부터 클라이언트가 TCP 연결을 종료하는 시간까지로 한정하였다.

최근의 웹 서비스는 클라이언트와 서버사이에 여러개의 TCP 연결설정을 통하여 여러 서버에 의해 이루어지는 다양한 형태를 갖지만, 본 측정 모델에서는 가장 기본적인 웹 서비스인 단일 서버에 하나의 TCP 연결을 설정하여 서비스를 제공받는 서비스 형태를 기반으로 하였다. 이러한 관점에서 웹 페이지를 다운로딩하는데 관련된 구성요소들을 정의하고, 이들 요소들이 장애나 전체 응답시간에 어느 정도 영향을 미치는 지를 분석하기 위한 가용성 측정 모델을 제시한다. 웹 서비스 가용성은 주기적인 웹 페이지 다운로딩을 수행하는 과정에서 장애의 발생으로 다운로딩을 할 수 없는 경우의 수와 응답시간이 임계치를 초과하여 성능저하가 발생한 경우의 수를 계산하여 측정된다.

3.1 구성요소 정의

아래의 (그림 1)에서와 같이 웹 페이지를 다운로딩하는데 걸리는 시간을 클라이언트가 소켓을 열어 DNS를 거쳐 TCP 연결을 설정하고, 페이지를 전송, 연결을 종료하는 순차적인 과정에서 관여되는 다양한 요소들로 정의한다.

본 논문의 측정범위에서 사용자가 느끼는 페이지 다운로딩에 걸리는 시간을 측정하기 위해 ART(Application Response Time)를 정의한다. ART는 클라이언트에서 측정되는 페이지 다운로딩에 걸리는 시간으로, 서버 연결을 위해 TCP 소켓을 연 시간에서 서버로부터 마지막 페이지 데이터를 받은 시간까지의 시간차로 정의된다. 클라이언트가 서버로부터 마지막 데이터를 받은 시간을 인식하는 순간은 서버로부터 TCP가 연결 종료 패킷을 받아 클라이언트가 연결 종료 메시지를 보내는 사이에 있다. 서비스 과정 중에 장애로 인한 예외상황이 발생하면 ART는 측정되지 않게 되며, 웹 서비스의 구성요소에 의해 지연이 발생하면 ART는 높게 측정될 것이다. ART가 장애를 나타내거나, ART가 임계치를 초과하는 경우에 불가용성으로 판정된다.



(그림 1) 웹 서비스의 구성요소

기존의 많은 연구자들은 소켓을 열고 TCP 패킷을 전송하기 전까지의 고장에 대하여 진단을 단순한 DNS 문제로 다루고 있다. 하지만, 소켓을 여는 시점과 DNS질의 메시지를 보내는 시점, 그리고 TCP 전송의 시점 사이에 시간차가 발생한다. 본 논문에서는 이러한 클라이언트의 소켓 처리 시간을 고려하여 아래와 같은 요소들을 정의한다.

- DLT(DNS Lookup Time) : 클라이언트가 서버 도메인

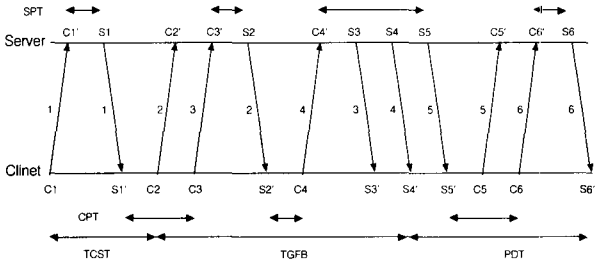
네임에 대하여 DNS Lookup에 걸리는 시간

- DAD(Distance between ART starting point and DLT starting point) : ART의 시작 시간에서 DLT 시작 시간까지의 시간차
- HTT(Http Transfer Time) : HTTP 통신을 위한 TCP 연결 설정에서 연결 종료까지 시간차
- DDT(Distance between DLT ending point and HTT starting point) : DLT 종료 시간에서 HTT 시작 시간까지의 시간차

DAD와 DDT는 클라이언트가 서버에게 TCP 연결 설정을 시도하기 전까지의 응답시간에 미치는 영향을 분석하기 위하여 정의되었다. 이는 TCP 패킷을 전송하기 전에 발생하는 장애가 DNS에 의한 것인지 클라이언트에 의한 것인지를 판단하는 증거를 제공한다. DNS Resolver는 DNS 질의에 대한 결과를 캐쉬하지는 않는다. 하지만, 하나의 프로세스가 이미 알고 있는 IP에 대하여 질의하는 경우에는 도메인 네임이 아닌 해당 IP로 질의하기 때문에 측정하고자 하는 정확한 DLT를 측정하지 못하는 경우가 발생한다. 그러므로 ART를 매번 새로운 프로세스가 측정하도록 만들어 DNS 질의 메시지를 보내도록 해야 한다.

HTT는 기존의 연구와 유사하게 TCP와 HTTP의 프로토콜 수준에서 세 부분으로 구별하여 정의한다[1]. 즉, 클라이언트와 서버간의 TCP 연결 설정 시간을 나타내는 TCST (TCP Connection Setup Time), TCST 이후 클라이언트의 요청에 대하여 서버로부터 처음 페이지 데이터를 받은 시간을 나타내는 TGFB(Time to Get the First Byte), TGFB 이후로 TCP 연결 종료까지의 시간을 나타내는 PDT(Page Downloading Time)로 HTT를 구분할 수 있다. ART와 HTT의 정의에 따라 ART는 HTT를 포함하지는 않는다. 즉, ART는 클라이언트의 종료에 대한 서버의 응답을 기다리는 시간을 포함하지 않기 때문에 PDT를 포함하지 않는다.

본 논문에서는 HTT에 미치는 클라이언트, 서버, 네트워크의 영향력을 분석하기 위한 새로운 요소들을 정의한다. (그림 2)의 예에서 클라이언트와 서버사이에 HTT를 구성하는 TCP 메시지를 살펴보면, 처음 3-way handshaking을 위한 3개의 패킷을 주고받으며, 양방향 TCP 연결 종료를 위하여 마지막에 4개의 패킷을 주고받는다. 연결이 설정된 후, 클라이언트는 C_3 시간에 HTTP 요청 메시지의 첫 바이트를 전송하여 S_2' 시간에 서버로부터 응답을 받는다. 클라이언트는 C_4 시간에 HTTP 요청 메시지의 나머지 바이트를 전송한다. 서버는 S_3 시간에 클라이언트에게 이에 대한 응답을 보내고 S_4 시간에 페이지의 데이터를 전송하기 시작한다. 나머지 페이지 데이터의 전송이 완료되면(이 부분은 그림에서 생략), 서버는 S_5 시간에 클라이언트에게 종료 패킷을 전송한다.



(그림 2) HTT에 대한 CPT, SPT의 구성 예

(그림 2)와 같이 HTT 측정값은 메시지를 생성하고 응답하는 클라이언트, 서버의 영향과 전송을 담당하는 네트워크의 영향으로 구분하여 정의될 수 있다. 즉, 서버로부터의 연결 설정에 대한 응답 패킷을 받은 S_1' 시간부터 HTTP 요청 메시지를 보내는 C_3 시간까지는 네트워크나 서버의 영향이 없는 순수한 클라이언트 처리 시간이다. 또한 클라이언트로부터의 연결 설정에 대한 요청 패킷을 받은 C_1' 시간부터 이에 대한 응답 패킷을 보내는 S_1 시간까지는 네트워크나 클라이언트의 영향이 없는 순수한 서버 처리 시간이다. 본 논문에서는 클라이언트와 서버의 영향을 분석하기 위해 클라이언트 처리 시간과 서버 처리 시간을 아래와 같이 정의하였다.

- CPT(Client Processing Time) : 서버로부터 요청이나 응답을 받아 클라이언트가 메시지를 생성하여 전송하기까지의 소요된 시간들의 합
- SPT(Server Processing Time) : 클라이언트로부터 요청이나 응답을 받아 서버가 메시지를 생성하여 전송하기까지의 소요된 시간들의 합

위와 같은 정의에 따라 (그림 2)에서 CPT는 $(C_3 - S_1') + (C_4 - S_2') + (C_6 - S_5')$값이 되며, SPT는 $(S_1 - C_1') + (S_2 - C_3') + (S_5 - C_4') + (S_6 - C_6')$값이 된다. 이러한 정의로부터 네트워크의 영향을 나타내는 요소 NLT(Network Latency Time)를 다음과 같이 정의할 수 있다.

$$NLT = HTT - (CPT + SPT) \quad (1)$$

위의 정의에 따라 HTT는 CPT + NLT + SPT값이 된다. HTT에 포함되는 이들 3 요소들은 HTT 전체가 ART에 포함되는 것이 아니기 때문에 ART와의 관계에서 나타나는 의미를 정확하게 논의할 수는 없다. 즉, HTT가 ART를 초과하는 높은 값으로 측정되는 경우, HTT의 분석이 ART의 성능 저하에 대하여 100% 진단할 수는 있는 자료가 될 수는 없다는 것이다. 하지만, ART는 TCST와 TGFB를 포함하기 때문에 이들 요소에 의해 ART가 영향을 크게 받은

경우, 그 영향을 파악하는 것은 의미가 있다. 만약 ART의 성능 저하에 대하여 TCST와 TGFB가 매우 높은 영향을 끼쳤다면, HTT의 분석을 통하여 CPT, NLT, SPT중 어느 요소가 HTT에 영향을 끼쳤는지 분석할 수 있다. NLT가 매우 높게 측정되었다면, 네트워크의 부하가 높은 경우인데, 이를 보조하기 위한 자료로 Ping 측정치를 이용하는 것이 의미가 있을 것이다. 테스트에서 Ping은 서버의 영향을 없애기 위하여 바로 이전 노드까지 측정하며, Ping 측정치와 NLT의 상관관계 통하여 NLT의 정확성을 분석하였다.

마지막으로 사용자가 느끼는 ART에 대하여 직접적으로 영향을 미치는 요소들을 측정하여 이들 요소들의 측정치의 합이 ART를 얼마나 반영하고 있는지를 분석하고자 한다. (그림 1)과 같이 본 논문에서 제시한 모델에서 ART에 직접적으로 영향을 미치는 요소들은 DAD, DLT, DDT, TCST, TGFB이며, 사용자가 느끼는 웹 품질을 의미하는 WQF(Web Quality Factor)를 아래와 같이 정의하였다.

$$WQF = DAD + DLT + DDT + TCST + TGFB \quad (2)$$

모델에 따라 WQF는 ART보다 작거나 같으며, 이들 사이에는 매우 높은 상관관계가 있음을 입증하여야 할 것이다. 모델의 정확성을 입증하기 위하여 WQF의 신뢰계수 α를 다음과 같이 정의한다.

$$\alpha = WQF/ART \quad (3)$$

$$WQF >= \alpha \times ART \quad (4)$$

α는 0과 1사이의 값을 갖게되며, α값이 1에 가까울수록 WQF는 ART를 잘 반영함을 의미한다. α값은 모든 측정치에 대하여 식 (4)를 만족하는 최대값으로 결정되며, 1에 가까울수록 WQF를 구성하는 요소들의 분석에 대한 그 의미가 크게 된다. 또한, 우리의 관심사는 모든 측정치보다는, 주어진 임계치보다 성능이 저하된 특별한 경우에 있으므로 이러한 경우에 식 (4)를 만족하는 최대값 α를 분석하여 모델의 정확성을 입증하여야 한다. 예를 들어, α = 0.9일 때 모든 측정치에 대하여 만족하였다면, WQF >= (0.9)×ART 임을 나타내며, ART값이 임계치를 초과하는 모든 경우들에 대하여 WQF >= (0.99)×ART를 만족하였다면, 0.9와 0.99가 모델의 정확성을 나타내는 척도가 된다.

결국, 모델의 정확성을 위하여 본 논문에서 정의하는 WQF가 측정결과 어느 정도 1에 가까운 α를 갖는지를 분석해야하며, 필요하다면 더 높은 α를 얻을 수 있는 WQF를 새롭게 정의해야 할 것이다.

테스트 환경에서 ART가 임계치를 초과하는 경우의 클라이언트, DNS, 네트워크, 서버가 어떠한 패턴으로 영향을 미쳤는지를 분석하고자 한다. 본 논문에서는 DAD, DDT, CPT를 클라이언트 그룹요소(GC : Group Component)로 구분하고, DLT를 DNS 그룹요소로, NLT를 네트워크 그룹요소로, SPT를 서버 구룹요소로 구분하여 클라이언트, DNS, 네트워크, 서버의 영향력을 분석한다. ART가 임계치를 초과한 경우, 가장 크게 영향을 미친 그룹요소의 측정치를 MAX(GC)라 하면, ART에 가장 크게 영향을 미치는 그룹요소의 영향력을 나타내는 β를 식 (5)와 같이 정의할 수 있다.

$$\beta = MAX(GC)/ART \quad (5)$$

$$MAX(GC) >= \beta \times ART \quad (6)$$

β은 0과 1사이의 값을 갖게되며, β값은 식 (6)을 만족하는 최대값으로 결정된다. 1에 가까울수록 성능 저하의 원인이 클라이언트, DNS, 서버, 네트워크 중 하나의 그룹요소에 크게 영향을 받는 특성을 나타낸다고 말할 수 있다. 예를 들어, β = 0.9로 결정된다면, 테스트 환경의 특성이 ART 값이 임계치를 초과하는 모든 경우, 단일 그룹요소에 의해 0.9 이상의 영향력으로 성능저하를 초래한다는 의미이다.

임계치는 고객과 서비스 제공자 사이의 SLA 협약에서 결정된다. 서비스 제공자는 특정 고객에 대하여 제공할 임계치에 대한 정확한 기준이 없기 때문에 가용성에 대한 분석을 먼저 수행해야 할 것이다. 예를 들어, 특정 고객이 2초 이하의 응답시간을 갖는 99% 이상의 가용성을 요구할 경우, 서비스 제공자는 이를 제공할 수 있는지 없는지를 먼저 분석해야 한다.

3.2 측정구조

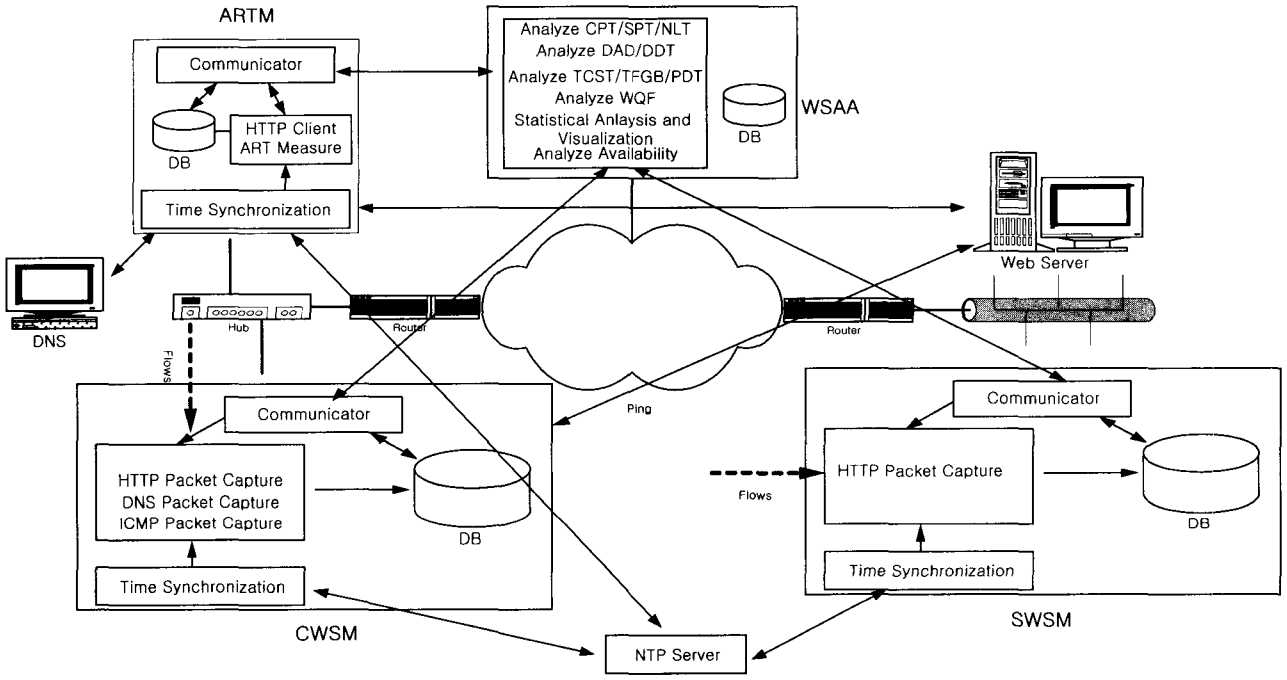
WebSerAvail은 Client-Server 기반의 웹 서비스의 가용성을 측정하기 위하여 ARTM(ART Monitor), CWSM(Client-side Web Service Monitor), SWSM(Server-side Web Service Monitor), WSAA(Web Service Availability Analyzer)로 구성된 측정구조를 갖는다. ARTM는 클라이언트의 주요기능인 HTTP 클라이언트의 기능을 수행하면서, ART를 측정한다. CWSM은 Client-Server 사이의 HTTP Flow를 수집하는 기능과 웹 서버에 대한 DNS 패킷 수집 기능, ICMP 패킷을 생성하고 수집하는 기능을 수행한다. SWSM에서는 Client-Server 사이의 HTTP Flow를 수집하는 기능을 수행하게 된다. Ping과 함께 네트워크 지연에 대한 의미를 정확하게 해석하고자 3개의 모니터들은 시간 동기를 수행하는 기능을 포함한다. WSAA는 ARTM, CWSM, SWSM으로부터 정보를 수집하여 웹 서비스 가용성을 분석하는 기능을 수행한다. 각각의 기능들을 요약하면 아래와 같다.

- ARTM
  - Simple HTTP 클라이언트 기능
  - ART Monitoring 기능

- 시간 동기화 기능
- CWSM
  - DNS 패킷 수집 기능
  - TCP/HTTP 패킷 수집 기능
  - ICMP 패킷 생성/수집 기능
  - 시간 동기화 기능
- SWSM
  - TCP/HTTP 패킷 수집 기능
  - 시간 동기화 기능
- WSAA
  - 원시 데이터 수집/필터 기능
  - DAD/DLT/DDT 분석 기능
  - TCST/TGFB/PDT 분석 기능
  - WQF 분석 기능
  - CPT/SPT/NLT 분석 기능
  - 통계분석 및 가시화 기능
  - 웹 서비스 가용성 분석 기능

아래의 (그림 3)은 WebSerAvail의 측정구조를 보여준다. ARTM와 CWSM은 클라이언트 측+에서 기능을 수행하며, SWSM은 서버 측에서 그 기능을 수행한다. WSAA는 일반적으로 서비스 제공자 측에 위치하며, MSP와 같은 Outsourcing 사업자의 경우는 MSP 내부에 둘 수 있을 것이다. 그러나, 논리적으로 WSAA가 기능을 수행하는 위치는 어느 곳이든 상관이 없다. WSAA는 각 모니터에게 원시 데이터를 측정하도록 요청한다. ARTM은 WSAA로부터 측정 대상이 되는 서버의 도메인 네임과 URL을 받아 HTTP 클라이언트 기능을 수행한다. 수행하는 과정에서 소켓을 열어, URL로부터 모든 정보를 받을 때 까지의 시간(ART)을 측정한다. 이러한 ARTM에서의 측정은 일정 주기로 반복하여 이루어진다. CWSM은 ARTM에서 보내는 DNS 패킷, TCP/HTTP 패킷을 수집하고, WSAA로부터 얻은 목적지에 대하여 Ping 측정을 수행한다. SWSM은 ARTM에서 보내는 TCP/HTTP 패킷을 수집한다. 세 모니터는 시간 동기화를 위하여 NTP 통신을 수행한다. 테스트 기간이 완료되면, WSAA는 각 모니터로부터 원시 데이터를 수집하고 분석을 수행한다.

아래의 <표 1>에서는 각 모니터들로부터 받은 원시 데이터를 바탕으로 WSAA에서 가용성 분석을 위해 생성하는 Base 데이터의 예를 보여준다. 측정 순차를 나타내는 Index와 수집 시간을 나타내는 Checktime은 생략되었다. Conntry는 TCP가 연결을 설정하려는 시도 회수를 나타내며, Con-



(그림 3) WebSerAvail의 측정구조

<표 1> WSAA에서의 Base 데이터 예

(ms 또는 개수)

art	dlt	htt	tcst	tgfb	pdt	spt	conntry	connsetup	request	reply	status	dad	ddt	dcs	ping	wqf	cpt	nlt
2751	5	3738	920	1890	928	5	196	196	1177	1176	200	2	2	-2825	908	2819	26	3707

nsetup는 실제 TCP 연결 설정 회수를 나타낸다. 두 요소는 정상적인 경우, 이전 Index로부터 각각 1회씩 증가하게 된다. 하지만, TCP 연결이 제대로 설정되지 않는 경우에는 Conntry 값이 2 이상 증가하거나, 이전 Index의 값과 비교하여 증감이 없게 될 것이다. Status는 HTTP 응답 메시지의 상태 코드 값을 의미한다. Status 값은 200이 정상이지만, 서버의 다양한 정책에 의하여 300 시리즈의 값들이 나타날 수 있다. DCS는 클라이언트가 TCP 연결 설정을 시작한 시간과 서버가 이 패킷을 받은 시간의 차를 나타낸다. 클라이언트와 서버의 시간 동기가 정확하게 이루어진다면, 네트워크 지연을 나타내지만, 실제 NTP를 사용하여 시간 동기를 하는 경우에는 이전 Index의 DCS와 값을 비교하여 네트워크 지연에 대한 참고자료로 사용될 수 있다.

3.3 분석 알고리즘

본 절에서는 WSAA가 가용성 분석 기능을 수행하기 위한 알고리즘을 살펴본다. 웹 서비스의 가용성을 장애 측면과 성능 측면에서 분석하는 것이 알고리즘의 핵심이다. 웹 서비스 가용성은 식 (7)과 같이 전체 측정회수에 대하여 장애 회수와 임계치보다 큰 응답시간을 갖는 성능 저하 회수의 합을 제외한 가용회수의 비율로 계산된다.

$$WSA = (\text{측정회수} - (\text{장애회수} + \text{성능저하회수})) / \text{측정회수} \times 100(\%) \quad (7)$$

3.3.1 장애 분석

장애는 웹 서비스가 정상적으로 동작되지 않는 경우로 클라이언트, DNS, 네트워크, 서버의 에러에 의해 발생할 수 있다. 단, 본 논문에서는 모니터의 비정상적인 동작으로 발생하는 장애는 논외로 한다. 장애의 발생여부는 ARTM이 측정하는 ART 값에 따라 결정된다. 웹 서비스에 관련된 요소들의 에러에 의하여 ARTM은 OS로부터 소켓 예외상황을 받게 되고, ARTM은 이에 대하여 장애를 나타내는 적절한 값을 기록하게 된다. 장애의 원인을 진단하는 알고리즘은 HTTP를 위한 TCP 패킷의 전송여부, DNS 질의/응답 패킷 송수신여부, 서버의 TCP 패킷 송수신여부, ICMP 패킷의 종류, TCP 재전송여부 등에 의해서 결정된다. 클라이언트 에러와 DNS 에러의 경우는 TCP 패킷이 전혀 발생하지 않으며, 네트워크의 에러나 서버의 에러인 경우에는 TCP 패킷 전송이 일어나게 된다.

TCP 패킷이 전송이 안된 경우를 살펴보자. DNS 질의 패킷이 전송되지 않았다면, 클라이언트의 에러가 발생한 것이다. (그림 1)에서 이 경우는 DAD 요소 값이 이상치로 기록될 것이다. DNS 질의 패킷이 전송되고, DNS 응답 패킷이 수신되었다면, DNS에는 문제가 없는 경우이며, 마찬가지로 클라이언트의 에러가 발생한 것이다. 이 경우 DDT 요소 값이 이상치로 기록될 것이다. DNS 에러가 발생한 경우는 DNS 응답 패킷이 오지 않게 되며, DLT 요소 값이 이상치로 기록될 것이다.

TCP 패킷이 전송된 경우를 살펴보면, 서버가 클라이언트로부터의 TCP 요청 패킷을 수신하지 못하였다면, ICMP 응답 패킷을 분석하여 서버의 에러인지 네트워크의 에러인지 판단할 수 있다. TCP 요청 패킷을 수신한 후, 응답 패킷을 보내지 않았다면, 서버의 에러가 발생한 경우이다. 서버가 TCP 응답 패킷 송신 후, 클라이언트에서 TCP 연결 설정을 위한 재전송이 이루어졌다면, 네트워크 에러가 발생한 경우이며, 재전송이 이루어지지 않았다면, 서버의 에러가 발생한 경우이다.

3.3.2 응답시간에 의한 성능 저하 분석

모든 측정치에 대하여 ART에 대한 그룹요소별 영향력을 계산하며, 응답시간이 주어진 임계치를 넘는 성능 저하의 경우에 가장 크게 영향을 미치는 그룹요소를 찾고 성능 저하 회수에 포함시킨다. 분석과정에서 WQF의 신뢰계수 α와 ART에 미치는 그룹요소의 영향력 β를 계산한다.

4. 테스트 환경 및 결과

4.1 테스트 환경

본 논문에서 이루어진 테스트는 전남대학교 네트워크 내의 168.131.161.20에 ARTM과 CWSM를, 168.131.85.84에 WSAA를 두었고, KOREN 망 내의 203.255.253.34, 203.255.253.37에 각각 서버와 SWSM를 배치하였다. 서버는 상용되지 않는 테스트용으로 만들었으며, 테스트 트래픽이외의 다른 부하는 없도록 하였다. 전남대학교 캠퍼스 망은 Gigabit 이더넷 백본과 T3 억세스 라인으로 구성되었으며, 서버 망은 ATM으로 구성되었다. 각 시스템의 사양을 살펴보면 <표 2>와 같다.

〈표 2〉 시스템 사양

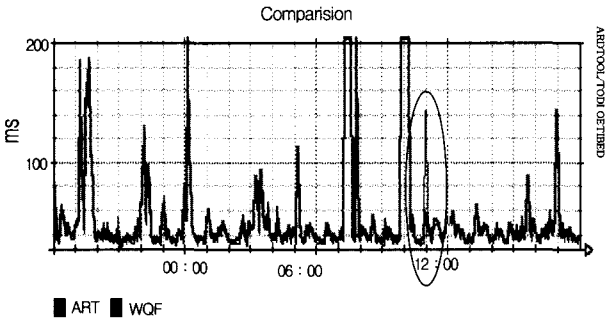
시스템	운영체제	CPU/Memory	IP
WSAA	Linux	PantiumIII-500, 256M	168.131.85.84
ARTM and CWSM	Linux	PantiumIV-1.7G, 256M	168.131.161.20
Web Server	Linux	PantiumIII-500, 512M	203.255.253.34
SWSM	Linux	PantiumIII-500, 256M	203.255.253.37

ARTM과 서버 사이의 홉수는 10이며, 이들 경로는 전남대학교 캠퍼스 망, KREN, KOREN로 구성되어 있다. 전남대학교 망 IP는 168.131/16으로 엑세스 라우터로서 168.131.18.3을 사용하고 있다.

테스트는 2002년 7/23 16 : 00부터 7/24 16 : 00까지 24시간 동안 5분 간격으로 수행하였다. 그러므로 전체 측정회수는 (12회/시간)×24시간 = 288회이다.

4.2 구성요소 관련성 검사

가장 먼저 ART와 WQF와의 상관관계를 분석한다. (그림 4)를 살펴보면 100ms 이하의 수치에서 나타나는 두 값의 그래프의 오차가 거의 없음을 알 수 있다. 타원으로 표시된 11시 부근에서 WQF가 ART를 초과하고 있음을 보여주는데, 이는 CWSM이 오류를 범한 경우이다. 이 경우는 ART = 60인데 비하여, WQF = 159로 나타나 α = 2.65로 분석되었다. 이는 ART가 적은 경우에 CWSM의 측정오류가 크게 반영된 예로, ART가 임계치를 넘는 큰 값을 갖는 경우에는 측정오류는 매우 미미하게 반영될 것이다.



(그림 4) ART와 WQF의 상관관계

ART와 WQF의 상관관계는 회귀분석을 이용한 F-검정에 의한 p-값이 0.0001로 매우 유의하며, 결정계수는 0.9996으로 회귀모형에 대한 설명력이 매우 높게 분석되었다. 구해진 회귀모형에서 β_0 = 2.193, β_1 = 0.995의 값으로

$$ART = (2.193) + (0.995) \times WQF \quad (8)$$

을 만족한다. (그림 5)는 ART와 WQF의 1차 회귀분석 모형이다. ART가 크게 측정되는 경우는 WQF 측정치도 ART에 매우 근접한 수치로 크게 측정되고 있는데, 이는 그 원인을 분석하는 방법으로 WQF을 구성하는 요소들을 분석하는 것이 타당함을 보여준다.

NLT와 WQF의 상관관계는 회귀분석을 이용한 F-검정에 의한 p-값이 0.0001로 매우 유의하며, 결정계수는 0.9927으로 회귀모형에 대한 설명력이 매우 높게 분석되었다. 구해진 회귀모형에서 β_0 = -23.07954, β_1 = 1.32484의 값으로

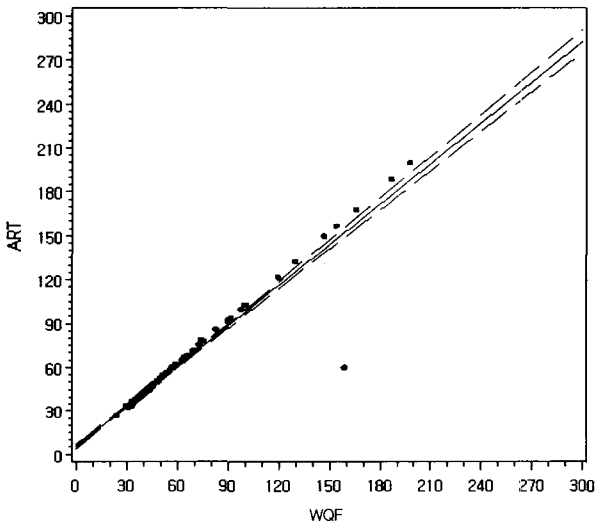
$$NLT = (-23.07954) + (1.32484) \times WQF \quad (9)$$

을 만족한다. 식 (9)의 의미는 현재의 WQF에 영향을 주는 요소가 네트워크 지연임을 의미한다.

NLT와 Ping의 상관관계는 회귀분석을 이용한 F-검정에 의한 p-값이 0.0001로 매우 유의하며, 결정계수는 0.9975으로 회귀모형에 대한 설명력이 매우 높게 분석되었다. 구해진 회귀모형에서 β_0 = -0.24899, β_1 = 4.02822의 값으로

$$NLT = (-0.24899) + (4.02822) \times Ping \quad (10)$$

을 만족한다.



(그림 5) ART와 WQF의 1차 회귀분석 모형

4.3 웹 서비스 가용성 분석

웹 서비스 가용성의 분석에 의하여 아래의 <표 3>과 같은 결과를 얻을 수 있었다. 장애가 발생한 경우는 클라이언트 측의 오류에 의한 것이었으며, 5회는 네트워크의 부하에

〈표 3〉 웹 서비스 가용성 분석표

Threshold	WSA	Performance						Fault						MC	UC
		C	D	N	S	CM	SM	C	D	N	S	CM	SM		
250ms	97.222%	0	1	5	0	1	0	1	0	0	0	0	0	288	8
500ms	97.222%	0	1	5	0	1	0	1	0	0	0	0	0	288	8
1000ms	97.916%	0	0	4	0	1	0	1	0	0	0	0	0	288	6
2000ms	97.916%	0	0	4	0	1	0	1	0	0	0	0	0	288	6
4000ms	99.652%	0	0	0	0	0	0	1	0	0	0	0	0	288	1

의하여, DNS와 CWSM 모니터에 의하여 각각 1회씩 임계치를 초과하였음을 분석하였다. C는 클라이언트를 의미하며, D는 DNS, N은 네트워크, S는 서버, CM은 CWSM, SM은 SWSM을 나타낸다. MC는 Monitoring Count를 UC는 Unavailable Count를 나타낸다.

기존의 측정방식처럼 장애에 의존하여 가용성을 계산한다면 모든 경우에 99.652%((288 - 1)/288×100)의 높은 가용성을 제공하는 것처럼 보인다. 하지만, WebSerAvail이 제시하는 <표 3>에 의해서 실제 사용자에게는 250ms의 임계치에서 97.222%의 가용성이 제공되었음을 알 수 있다. 또한, 클라이언트의 에러와 DNS에 의해 발생한 서비스 불가용성을 분별가능하여 서비스 제공자 측면에서 그 책임여부를 판단할 수 있는 근거를 제시하고 있다.

5. 결론 및 향후연구방향

본 논문에서는 전통적인 SLA의 가장 중요한 요소인 가용성과 응답시간을 통합한 새로운 개념의 응용서비스 가용성을 웹 서비스 관점에서 측정하는 방안을 제시하였다. 구현한 WebSerAvail 툴을 이용하여 기존의 장애 관점뿐만 아니라 성능까지 포함하여 그 원인을 진단하기 때문에 사용자 및 클라이언트 관점에서 요구되는 정확한 웹 서비스 가용성을 산출할 수 있게 되었다.

Test 결과, WQF는 ART가 임계치를 초과한 경우 α = 0.994를 만족하여 매우 높은 신뢰도를 나타내었다. 이를 통하여 본 논문에서 설정한 측정모델과 정의한 WQF가 ART의 특성을 매우 잘 반영하고 있음을 확실할 수 있었다. 또한, β = 0.922를 만족하여 ART가 임계치를 초과하는 이유가 단일 그룹요소에 의한 것이었음을 알 수 있었다. 그리고 이들 요소가 ART에 미치는 영향은 92% 이상이었다. 288회의 측정동안 기존의 측정법에 의하면 99.652%의 높은 가용성을 제공하는 것처럼 보이지만, 실제 사용자에게 250ms의 임계치에서 97.222%의 가용성이 제공되었음을 분석하였다. 1회의 클라이언트의 에러와 1회의 DNS에 의한 서비스 불가용성을 분별하여 서비스 제공자 측면에서 그 책임이 없음을 판단할 수 있는 근거를 제시하였다. 최소값과 백분위수 95 사이에서 측정된 ART에 대한 점유비율은 클라이언트 : 9~26%, DNS : 5~20%, 서버 : 2~15%, 네트워크 : 45~82%로 분석되었다.

측정도구 WebSerAvail를 이용하는 서비스 제공자는 사용자에게 서비스의 성능저하나 서비스 중단에 대한 원인을 제공함으로써 서비스에 대한 신뢰성을 높일 수 있게 될 것이다. 또한 고객을 포함한 서비스 제공자들 사이에 서비스 성능저하나 장애로 인해 발생한 문제에 대한 명확한 책임규명을 가능하게 하여 서비스 관리에 대한 투명성을 확보할 수 있게 될 것이다. 특별히 웹 서비스 제공자에게는 더 많은 사용자를 확보하고, 운영을 위한 비용의 절감을 이룩하도록 하여 다른 사업자들에 비해 경쟁력을 갖춘 서비스 제공자로 남는데 기여할 것이다.

향후에는 다양한 환경에서의 테스트를 통하여 본 논문의 결과를 검증하고, 더 높은 α 값을 갖는 WQF를 찾는 노력이 필요할 것이다. 또한 일반 상용 서버에 대한 가용성을 측정하기 위하여 SWSM를 대신할 수 있는 SPT 예측기능에 대한 연구가 필요하다.

참 고 문 헌

[1] A. Habib and M. Abrams, "Analysis of Sources of Latency in Downloading Web Pages," In Proc. Webnet, San Antonio, USA, 2000.

[2] B. Krishnamurthy, CE. Wills, "Anlayzing factors that influence end-to-end Web Performance," Proceedings of WWW-9 conference, May, 2000.

[3] J. Charzinski, "Measured HTTP performance and fun factors," In Teletraffic Engineering in the Internet Era (J. Moriera de Souza, N. L. S. Fonseca and E. A. de Souza e Silva eds.), pp.1063-1074. 2001.

[4] J. Mogul, "The case for persistent-connection HTTP," Tchnical Report WRL 95/5, DEC Western Research Laboratory, 1995.

[5] V. Padmanabhan and J. Mogul, "Improving HTTP latency," Computer Networks and ISDN Systems, Dec., 1995.

[6] Nina Bhatti, Anna Boutch, Allan Kuchinsky, "Integrating user-perceived quality into Web server design," 2000.

[7] Mimkka Koletsou, Geoffrey M. Voelker, "The Medusa Proxy : A Tool for exploring user-perceived web performance," In Proceedings of the Sixth International Workshop on Web Caching and Content Distribution, June, 2001.

[8] B. Chandra, M. Dahlin, L. Gao, A. Nayate, "End-to-end WAN Service Availability," USITS, 2001.

[9] NMF 701, "Performance Reporting Definitions Document," 1997.

[10] M. Kalyanakrishnan, R. K. lyer, and J. U. Patel, "Relaiability of Internet hosts : A case study from the end user's perspective," Computer Networks, Vol.31, 1999.

[11] Wei Xie, Hairong Sun, Yonghuan Cao, Kishor S. Trivedi, "Modeling of Online Service Availability Perceived by Web Users," 2001.

[12] Application Service Provider Industry Consortium, "White Paper on Sevice Level Agreements," 2000.

박 상 근

e-mail : psk@tyranno.chonnam.ac.kr
1996년 전남대학교 전산학 학사
1999년 전남대학교 전산통계학 석사
2000년~현재 전남대학교 전산학 박사과정
관심분야 : TMN, 인터넷 서비스 관리, SLA

최 덕 재

e-mail : dchoi@chonnam.ac.kr
1982년 서울대학교 컴퓨터공학 학사
1984년 한국과학기술원 전산학 석사
1995년 Univ.of Missouri-Kansas 컴퓨터 망 박사
1984년~1996년 전남대학교 전자계산소 운영부장
1997년~1998년 전자통신연구원 통신시스템 연구단 위탁 연구원
1996년~2000년 전남대학교 전산학과 부교수
2001년~현재 전남대학교 전산학과 교수
관심분야 : 네트워크 관리, 서비스 관리, TMN, 차세대 인터넷