멀티미디어 서비스를 위한 QoS 지원 자원관리 모델 연구

궁 상 환† · 윤 석 환† · 황 승 구†

요 약

이제까지는 통신망을 중심으로 한 서비스 품질(Quality of Services:QoS)이나 혹은 컴퓨터 시스템 등 단말 시스템의 내부적인 QoS가 주로 다루어져 왔으나, 최근 멀티미디어 응용이 출현하고 이를 지원하기 위한 QoS가 중요시 됨에 따라 종단간 일관성있는 QoS의 제공이 중요시 되고 있는 실정이다.

본 논문에서는 이질적인 환경에서 대규모 분산시스템을 지원하기 위한 QoS 기반 자원관리 구조를 제안하고 있는 데, 이 구조에서의 중요한 두가지 개념은 추상화(Abstractions)와 알고리즘이다. 추상화는 모든 시스템을 응용 관점, 자원 관점, 시스템 관점에서 모델화 하는 것을 말하며, 알고리즘은 이 세가지 모델들의 추상화를 통해서 획득한 정보를 활용하여 시스템 자원을 관리하는 데에 이용하는 것을 말한다. 아울러 본 논문에서는 자원관리 체계에 의한 자원관리의 흐름에 대해 설명하며, 특히 동적으로 변화하는 QoS 제공환경을 최적으로 활용하기 위한 응용 이득함수(Benefit Function)를 소개한다. 끝으로 멀티미디어의 대표적 응용인 텔레세미나를 이용하여 자원관리 구조의 활용 및 효과를 살펴 보기로 한다.

A Study on QoS Based Resources Management Model for Supporting Multimedia Services

Sanghwan Kung† · Seokhwan Yoon† · Seungku Hwang†

ABSTRACT

A lot of talks on the Quality of Services(QoS) so far were limited only either for the network itself or for the end system, But, recently it is becoming more important to have the seamless end-to-end QoS support to meet the networked multimedia service requirements.

This paper proposes the QoS based resource management architecture which supports the large scale distributed systems under heterogeneous environment. The architecture introduces a couple of key concepts such as abstractions and algorithms. Abstractions capture the three perspectivest found in all systems:application, resource, and system, on the other hand algorithms manage system resources based on information in the three abstractions. We also explain the on-line resource management flow as well as the benefit function which specifies the level of benefit accrued to the application user as a function of the level of service provided by the system. At the end of the paper, the QoS based resource management applied to the tele-seminar application is described to show its practical utilization.

1. 배 경

멀티미디어 응용은 일반적으로 시스템에서 지원하여야 하는 일련의 서비스 품질(Quality of Service:QoS) 요구사항을 가지고 있으며, 이는 시간성(temporal), 정확도(accuracy), 정밀도(precision)의 개념이다. 정밀

† 정 회 원:한국전자통신연구소 컴퓨터연구단 멀티미디어연구부
논문접수:1997년 3월 21일, 심사완료:1997년 11월 26일

도는 프레임 크기와 오디오 샘플 크기를 의미하며, 정확도는 훼손되는 비트 량을 의미하고, 시간성은 오디오 샘플과 프레임의 지터 및 종단간 지연시간을 의미한다.

멀티미디어 응용에 대해 QoS를 지원하는 방법은 3가지로 분류할 수 있다. 첫째는 정해진 수준의 서비스 품질을 실현하기 위하여 가능한 모든 자원을 활용하는 방안(best effort QoS)이 있으며, 이 경우에는 QoS 수준을 시스템 부하의 함수로서 표현한다. 둘째는 보장된 수준의 QoS를 응용에게 지원해야 하는 경우로서, 이 때에는 QoS 값이 시스템 부하 수준에 관계없이 미리 정해진 범위 이내의 값으로 고정되어 있다. 셋째는 예측할 수 없는 외부 환경에 대한 경우로서, 이 때에는 외부의 충격을 최소화하기 위하여 해당 응용의 QoS를 점진적으로 감쇄하는 방식이 바람직하다.

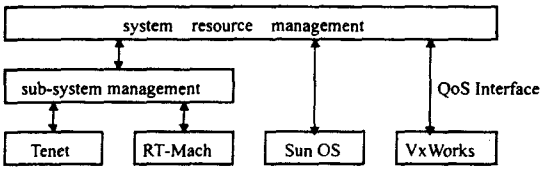
이들에 대한 공통된 어려움은 대부분의 통신 네트워크나 분산 시스템에서 있을 수 있는 높은 수준의 자원 공유 문제이다. 이에 대해서는 일반적인 운용 조건하에서 컴퓨팅이나 저장 및 통신 시스템 자원들에게 필요한 QoS 보장을 해 주고, 예측할 수 없는 외부 환경에 대해 점진적 QoS 감쇄가 가능하도록 하는 자원관리 체계의 도입이 필요하다. 이에 관한 몇몇 연구들이 진행되기는 하였으나 주로 개별적인 자원 관리에 관한 내용들이었으며[1, 2, 9, 11], 본 논문에서는 종단간의 자원관리는 물론 응용과 하부 구조를 포함한 총체적 관리를 위한 종합적 모델을 제시하고자 한다. 또한 다른 연구들이 QoS 요구사항이나 시스템 목표만을 고려하는 데 비해[7, 10, 11, 12, 15, 16], 본 논문에서는 응용의 QoS 요구사항과 시스템 목표의 최적화를 동시에 도모하고 있다.

제안하는 자원관리 체계는 다음의 내용으로 구성된다. 사용자의 응용 호출에 의해 자원관리 체계가 호출되면, 이는 해당 응용의 QoS 요구사항을 요청한다. 승인 단계에서, 자원관리 체계는 시스템 자원을 이 응용에 할당하도록 시도하고, 이 자원이 공유된 다른 응용과 동시에 QoS 요구사항을 만족시키는 방법으로 스케줄 한다. 만약 주어진 응용의 QoS 요구사항을 만족시킬 수 없다면, 시스템과 응용 사이의 협상(negotiation)을 통하여 시스템이 가능한 한 많은 QoS 요구사항을 만족시킬 수 있도록 한다. 일단 승인 단계가 통과되면, 자원관리 체계는 응용 QoS 요구사항이 시스템에 의하여 만족되어 지는가를 확인하기 위해 시스템을 모니터한다. QoS 요구사항을 계속해서 만족시키기 위하여 장애 관리(fault management)가 요구될 수도 있다. 만약 모니터링 과정에서 자원관리 체계가 시스템 장애를 발견하면, 자원관리 체계는 가능한 한 많은 장애를 극복하기 위하여 컴퓨팅 및 통신 자원을 재분배하고 응용들을 재스케쥴 하려고 할 것이다. 만약 장애가 심각하여 QoS 요구사항을 만족시키기 위한 재조정 과정이 성공하지 못하게 되면, 자원관리 체계는 각 응용에 대한 충격을 최소화하기 위하여 응용 QoS를 서서히 감쇄시키는 효과적 감쇄 알고리즘을 호출해야 한다. 만지막으로, 응용이 끝났을 때에(장애가 발생하지 않았더라도), 남아있는 응용들의 QoS를 최대화하기 위하여 자원관리 체계는 다른 응용들에게 자원을 재분배하고 재스케쥴 한다.

2. QoS기반 자원관리 체계

멀티미디어 응용에 대한 효율적인 QoS 지원을 위해서는 이질적인 시스템 자원을 조정하고 관리하는 계층적이며 적응형인 종단간 종합 자원관리 체계가 요구된다. 대부분의 원격 멀티미디어 응용은 통신 및 컴퓨팅 자원을 가지고 수행되기 때문에, 여러가지 자원의 관리를 포함하는 응용 수준의 일관성 있는 종단간 QoS 지원이 필요하다. 네트워크 상태는 응용의 시작과 종료, QoS 변경 요청, 자원 이용 실패 및 복구 등 시간의 변화에 대하여 빠르게 변화하기 때문에 적응형 QoS 관리가 필요하다. 적응형은 영상 전송시 QoS의 하나 파라미터인 패킷 손실을 측정해서 이 값이 임계값(threshold) 이하가 되면 응용에서의 영상 전송 프레임 수를 감소시켜 네트워크에 적응하도록 하는 것이다. 계층적이라 함은 서브 시스템 자원의 계층화를 통하여 시스템을 보다 더 큰 시스템으로 모델링하는 것을 의미한다. 컴퓨팅, 통신 및 저장 자원들이 하부층을 형성하고, 하나의 자원관리 체계에 의해 통제되는 자원들의 집합은 하나의 서브 시스템을 형성하며, 하나의 자원관리 체계에 의해 통제되는 서브 시스템들의 집합은 상위 레벨의 서브 시스템을 형성한다. 이러한 구조 형성은 완전한 시스템이 정의될 때까지 반복적으로 계속된다. 즉, 개인용 컴퓨터, 워크스테이션 및 서버 등이 FDDI, Ethernet, 기타 정비 등과의

연결을 통하여 컴퓨팅 자원관리 체계를 이루게 되며, 궁극적으로는 종단간 관리 체계를 갖추게 되는 것이다. (그림 1)에서는 이러한 개념을 보여 준다.



(그림 1) 시스템 모델의 예제
(Fig. 1) System modeling example

본 논문에서는 이를 위한 두 가지의 핵심적인 요소로 추상화(abstractions)와 알고리즘(algorithms)을 제안한다. 추상화는 모든 시스템에서 볼 수 있는 세 가지의 관점-응용, 자원, 시스템-을 수용하며, 알고리즘은 이러한 세가지 추상화에서 발견되는 정보를 토대로 시스템 자원을 관리하는 것을 말한다.

2.1 QoS 기반 자원관리 추상화

하나의 분산 시스템을 서술하기 위해서는 다음과 같이 세가지 모델의 추상화 방식이 필요하다.

- 응요 모델:각 응용의 실행 특성과 QoS 요구사항 서술.
- 자원 모델:각 자원의 부하 특성과 국부적 자원관리 정책 서술.
- 시스템 모델:시스템을 구성하는 자원들의 상호작용, 시스템 토폴로지, 종단간 자원관리 정책 서술.

응용은 다양한 실행 특성과 QoS 요구사항을 갖게 되는 데, 본 논문에서는 이 두 가지 중요한 속성을 모델링하기 위한 계층적 추상화 방식으로 Logical Realization of Service(LRoS)를 제안한다. LRoS는 방향성 그래프로 나타내어지는 데, 그래프 노드는 Logical Service(LS)와 Logical Units of Work(LUoW)에 대응하며, 그래프 가지는 응용 내에서의 제어 및 데이터의 흐름을 지정한다. LUoW는 특정한 자원 혹은 한 서브시스템에서 응용 요소의 정확한 실행 특성을 지정하며, LS는 해당 응용이 시스템으로부터 필요로 하는 서비스를 지정한다.

서비스 사용자는 시스템으로부터 어느 정도의 QoS 수준을 요구한다. 본 논문에서는 이를 이득 함수 추상화를 이용하여 모델링한다. 제안하는 이득 함수는 시스템이 일정 수준의 QoS 값을 제시할 때 사용자가 받게 되는 이득을 나타내는 다차원 그래프이다. 이득 함수는 응용과 시스템 사이의 효과적 감쇄를 촉진하는 데에 특히 유용하다. 만약 자원이 실패하거나 보다 높은 우선순위를 가진 응용에 사용된다면, 시스템은 낮은 우선순위를 가진 응용의 QoS 수준을 계속해서 만족시키기가 어려울 것이다. 그러나 복수 개의 응용과 각 응용에 대한 복수개의 QoS 메트릭스가 존재하기 때문에, 시스템에서는 일반적으로 어떤 응용 또는 어떤 QoS 메트릭스를 먼저 감쇄 시킬지에 대한 지능적 의사결정을 위한 충분한 정보를 가지고 있지 않다. 이득 함수는 바로 이러한 정보를 시스템이 응용 이득에 대한 충격을 최소화 하는 방법으로 QoS 파라미터를 감쇄시킬 수 있도록 지정하는 것이다.

본 논문에서는 여러 시스템들로 이루어진 거대한 시스템을 서브시스템 자원의 계층적 구조를 이용하여 모델화한다. 즉, 컴퓨팅, 통신 및 저장 자원은 하위층을 형성하며, 통합된 하나의 자원관리 체계에 의해 관리되는 자원들은 하나의 서브시스템을 형성한다. 이러한 계층적 구조는 완전한 시스템이 정의될 때까지 계속된다.

본 모델은 분산 환경에서의 다양한 이질성을 극복하는 일반적인 모델에 초점을 둔다. 일질적 환경의 예는 다양한 통신망이나 단말 시스템, 서로 다른 O/S 환경 뿐 아니라 하나의 단말 시스템 내에서도 디스크나 CPU 등 다양한 자원들을 총체적으로 다루게 된다. 예를 들어 Ethernet에 연결된 여러 대의 PC 시스템들과 FDDI에 연결된 SUN WS 시스템들을 통합해서 어떠한 응용 서비스를 제공하고자 할 때, Ethernet의 시스템들과 FDDI의 시스템들은 각각이 서브 시스템으로 자원관리에 의해 관리되도록 통신망과 단말의 이질성을 극복하면서 총체적인 QoS를 감시하고 제어할 수 있도록 모델링이 가능한 것이다.

2.2 QoS 기반 자원관리 알고리즘

응용 QoS 및 시스템 목표에 기반한 자원관리는 새로운 개념의 알고리즘을 요구하는 데, 이에는 다음의 세가지 중요 기능이 포함되어야 한다.

- 승인 제어/협상 : 자원관리자는 응용 QoS 요구사항을 만족시키는 방법으로 시스템 자원을 각 응용에 배분하며, 동일 시스템 자원에 대한 타 응용도 고려하여 응용을 스케쥴 한다. 만약 초기의 응용 QoS 요구사항이 만족되지 않는다면, 자원관리자는 시스템이 가능한 한 많은 QoS 요구사항을 만족시킬 수 있도록 하기 위하여 시스템과 응용 사이의 협상을 촉진한다.
- 모니터링 : 자원관리자는 보안 침입, 자원의 장애 및 복구, 응용의 시작과 종료에 기인한 자원의 유용성의 변화를 감지하기 위하여 시스템 자원을 모니터링한다.
- QoS 적응 : 자원관리자는 모니터링 정보를 토대로 필요에 따라 자원을 재분배하고 응용들을 재스케쥴한다. 자원관리자가 모니터링 중 특정 자원 장애를 발견하면, 자원관리자는 가능한 한 장애를 최소화 하기 위하여 자원들을 재분배하고 응용들을 재스케쥴 한다. 만약, 장애가 치명적이어서 QoS 요구사항을 만족시키는 것이 불가능하면, 자원관리자는 응용에 대한 영향을 최소화하는 방향으로 응용 QoS를 효과적으로 감쇄할 수 있는 알고리즘을 호출한다.

이러한 기능들은 자원 및 서브시스템 계층에서 제공되어야 한다. 서브시스템 관리 체계가 영역내에 있는 모든 하부 서브시스템 및 자원에 대하여 종단간 승인 제어/협상, 모니터링 및 QoS 적응을 제공해야 하는 한편, 자원관리 체계는 단일 자원에 대한 승인 제어/협상, 모니터링 및 QoS 적응을 제공해야 한다. 이러한 알고리즘은 분산성이 있고 신축적이며 온-라인 영역에서 사용될 수 있도록 설계되어야 한다.

3. QoS 기반 자원관리의 구조

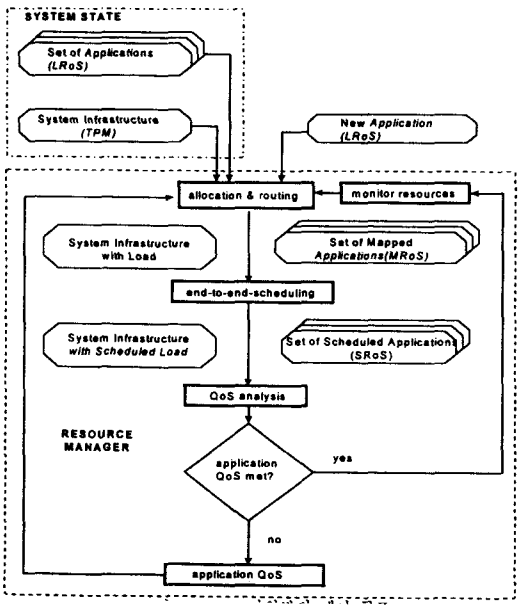
승인 제어/협상 및 모니터링, QoS 적응을 설명하는 제어 구조를 (그림 2)에 도시하였다. 특정 응용에 대한 사용자의 요구는 자원관리 체계를 동작시키고, 자원관리 체계는 서비스의 논리적 실현(Logical Realization of Services : LRoS) 추상화를 이용하여 응용의 자원 이용 특성 및 QoS 요구사항에 관한 정보를 저장한다. 그 다음, 자원관리의 분배 및 라우팅 알고리즘은 주어진 QoS 요구사항을 만족하고 시스템의 목적함수를 최적화하기 위하여 새로운 응용에 대해 어떤 자원을 예약할 것인가를 정한다. 이러한 결정은 시스템 토폴로지 및 자원 특성, 새로운 응용의 자원 사용 특성과 최초 응용의 요청에 대해 이루어진 예약 결정 등에 근거해서 내려진다. 또한 각 자원의 상태를 알려 주는 모니터링 알고리즘이 제공하는 정보에 의하여 이 결정은 영향을 받게 된다.

분배 및 라우팅 알고리즘은 LRoS를 서비스의 사상화된 실현(Mapped Realization of Service : 으로 변환하고 부하를 각 자원에 지저아고 이를 목표 플랫폼 모델(Target Platform Model : TPM)에서 갖고 있는 정보에 추가한다. 응용을 논리적 방식으로 혹은 목표 플랫폼에 독립적인 방식으로 기술하는 LRoS와는 달리, MRoS는 각 작업의 논리적 실현 단위(Logical Units of Works : LUoW)가 실행하는 정확한 자원을 기술한다. MRoS는 또한 분산 LUoW들 사이에서 통신하기 위하여 사용되는 정확한 입출력 경로를 제시한다. 예를 들어, 영상회의 응용이 이미 필요한 통신망 대역폭 등의 QoS를 할당 받아 동작하다가 전자 칠판 등의 데이타 회의 응용을 원하게 될 때에, 데이타 회의 응용은 필요한 대역폭 등의 QoS를 LRoS의 파라미터로서 요구하게 되고, 할당 및 라우팅 기능은 현재 통신망의 기용도를 모니터링한 결과를 토대로 요구한 QoS의 범위 이내에서 할당해 주고, 새로운 현실적인 QoS는 MRoS의 파라미터로 기억된다. 이 때, 시스템에서 새로운 대역폭의 QoS를 통신망의 여러 경로를 통해 제공할 수 있을 경우 최적의 경로와 제공 가능한 QoS가 MRoS의 값이 된다.

자원관리자는 그 후에 종단간 스케쥴링 알고리즘을 이용하여 각 응용을 스케쥴하는 데, 스케쥴링 알고리즘은 시스템의 행위에 대한 예측이 가능하게 하면서, 응용 QoS 요구사항을 만족시키도록 한다. 스케쥴링 알고리즘은 각 응용에 대한 스케쥴링 속성을 기술함으로서 MRoS와 TPM을 변환한다. 또한 각 응용이 실행되어야 할 시점과 우선순위 수준을 정한다. 이러한 정보는 서비스의 스케쥴에 따른 실현(Scheduled Realization of Service : SRoS) 추상화를 이용하여 표현한다. MRoS 및 TPM에서 스케쥴링 알고리즘에 의해 정해진 스케쥴링 속성과 우선순위를 고려하여 서비스의 실현 단위를 수행되는 순서에 따라 정리한 것을 SRoS라고 한다.

자원관리자는 이어서 응용의 QoS 요구사항이 시스템의 의하여 만족되고 있는지를 확인하는 QoS 분석 알고리즘을 이용한다. 만족되지 않는다면, 시스템적 방식으로 응용의 QoS 요구사항을 감쇄시키고, 할당 및 라우팅 알고리즘을 재호출한다. 감쇄는 LRoS의 부분으로서 응용 사용자에 의하여 지정된 이득 함수 추상화 방식에 따라 수행된다. 이득 함수는 시스템이 제공하는 QoS 지원 수준의 함수로서 사용자에게 발생된 이득을 지정한다. 이득 함수의 기울기는 응용에 대한 이득의 손실을 최소화하기 위하여 어떤 QoS 메트릭스를 먼저 감쇄할 것인지를 결정할 때 자원관리자가 직접 이용할 수 있다.

자원관리자는, QoS 분석 알고리즘에 의하여 응용의 QoS가 만족된다고 판단할 때, 모니터링 모드에 들어간다. 자원관리자는 어떤 장애가 발생하는지를 파악하기 위하여 시스템 자원을 모니터링하며, 자원의 상태가 변하거나 응용의 QoS 요구사항이 변하면 모니터링 알고리즘에서 정한 대로 분배 및 라우팅 알고리즘을 호출한다.



(그림 2) On-line 자원관리 제어 구조
(Fig. 2) On-line resource management overview

3.1 응용의 모델링

자원관리는 기본적으로 응용의 QoS 요구사항과 자원의 작업 부하 특성에 의해 영향을 받는다. 작업 부하 특성은 응용을 위하여 각 자원에서 어떤 일이 수행되어야 하는가를 기술한다. 응용의 QoS 요구사항과 자원의 작업 부하 특성을 모델링 하기 위한 여러 기법이 연구되어 왔다. 네트워크 응용을 모델링 하는 일반적 파라미터 집합과 트래픽 모델이 제시되었는데, 이 네트워크 트래픽 모델은 응용 수준의 풍부한 QoS 요구사항을 수용하지 못하며, 응용 수준의 자원 작업 특성을 충분히 수용할 수 있을 정도로 정교하지도 않다[1, 5]. 최근에 제시된 모델들은 CPU에서 수행되는 응용에 촛점을 맞추고 있다[2, 3, 6, 8]. 그러나 대부분은 응용의 동기화, 피드백, 멀티캐스팅 속성 등은 고려하지 않는다.

본 논문에서는 이질적 자원 형태에서 수행되는 종단 응용간을 모델링 할 수 있는 보다 더 다양한 추상화 방식을 제안한다. LRoS 추상화는 보다 풍부한 QoS 메트릭스와 결정적인 응용 작업 부하 특성을 포착하기 위한 정교한 모델을 제시한다.

3.1.1 LRoS 추상화(Logical Realization of Service Abstraction)

본 논문에서는 모든 응용을 시스템에 의하여 사용자에게 제공되는 서비스라고 가정한다. 사용자는 전형적으로 시스템이 QoS 요구사항을 만족시키면서 서비스를 수행해 주기를 원한다. 서비스의 구조는 LRoS 추상화를 이용하여 정의되는 데, LRoS는 이 서비스가 논리적인 측면에서 어떻게 수행되는 가를 설명한다. LRoS의 구체적인 예로서는 멀티미디어의 압축 서비스나 미디어간 동기 서비스 등을 들 수 있다. 하나의 서비스는 여러가지 방법으로 수행될 수 있기 때문에 하나의 서비스에 대한 복수 개의 LRoS가 존재할 수 있다. 또한 하나의 LRoS는 자신을 위한 추가적인 서비스를 호출하는 방식으로 반복적으로 정의될 수 있다. 하나의 LRoS는 다음의 5가지 속성으로 구성되며, 이를 테이블로 정리하면 〈Table〉과 같다.

- 인식(identification): 이름과 형태의 속성은 LRoS가 제공하는 서비스에 대한 일반적인 서술을 전달한다.
- 작업(work): 워크의 논리적 단위(LUoW)와 논리적 서비스(LS)는 응용이 시스템에 부과하는 작업 부하에 관한 정보를 전달한다.

- 흐름(flow): 제어와 데이타 가지는 작업이 수행되어야 하는 순서에 관한 정보를 전달한다. 즉, OR 구문은 자원관리 체계에서 선택할 수 있는 서로 다른 흐름 경로를 지정한다.
- 사용자 이득(user benefit): 시스템이 제공하는 QoS의 함수로서 응용 사용자가 받게되는 이득을 전달한다.
- 파라미터의 수납(inport/export parameters): LRoS에 제공하거나 LRoS로부터 제공받는 데이타를 의미한다.

〈표 1〉 LRoS의 속성(k : LRoS의 ID)
〈Table 1〉 Attributes of LRoS (k : ID of LRoS)

identification	LN^k	name of the service
	LI^k	identifier for the service iplementation/instance
work	LS^k	set of logical services
	$LUoW^k$	set of logical units of work
flow	LE^k	set of logical data feed-forward edges specifying precedence-constraints between logical work modules
	LF^k	set of logical data feedback edges specifying feedback constraints between logical work modules
	LC^k	set of logical control feed-forward edges specifying precedence constraints between logical work modules
	LO^k	set of logical OR constructs specifying a set of flow paths, any one of which will accomplish the application goals
user benefit	LB^k	one or more benefit functions specifying QoS range, requirements and application user preferences
import/export	LM^k	data that the LRoS imports
parameters	LX^k	data that the LRoS exports

LUoW와 LS 사이의 제어 및 데이타 흐름은 논리적 가지를 이용하여 지정된다. LRoS 추상화는 전방향성 및 후방향성의 데이타/제어 가지를 허용한다. 또한 이득 함수를 이용하여 시스템이 제공할 수 있는 QoS 지원 수준에 대한 함수로서 사용자가 받게 되는 이득을 지정한다.

3.1.2 이득 함수(Benefit Function)

서비스 사용자는 시스템으로부터 QoS 요구사항을 요구한다. 특정한 응용을 위한 QoS 요구사항은 QoS 메트릭스 집합으로 정량화 될 수 있다. 종단간 지연이나, 지터, 성능, 데이터 프레임의 크기, 정확도 등은 일반적인 QoS 메트릭스이다. 그러나 시스템에서는 모든 QoS 요구사항을 만족시킬 수 없기 때문에, 시스템과 사용자 사이의 협상이 필요하다. 이러한 협상에는 일반적으로 사용자에게 관심이 있는 메트릭스와 비교하여 시스템이 제공할 수 있는 서비스 수준 사이에서 조정이 이루어진다.

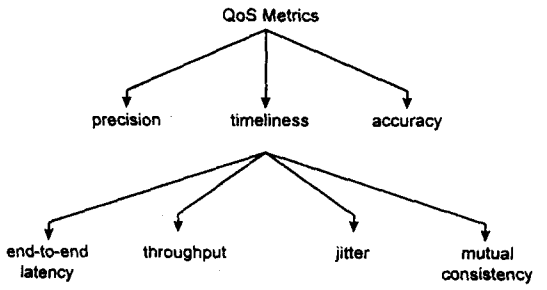
여러 개의 메트릭스가 사용자에게 관심이 있을 수 있기 때문에, 어떤 조정이 사용자에게 최적인가 하는 문제는 반드시 분명한 것은 아니다. 하나의 이득 함수는 시스템에 의하여 제공되는 서비스 수준의 함수로서 응용 사용자에게 부가되는 이득의 수준을 지정하는 다차원 함수이다. 사용자의 관심이 되는 여러 메트릭스의 함수인 이득 함수의 기울기는 시스템이 가능한 범위에서, 응용에 주는 이득을 최대화할 수 있는 QoS 집합을 찾기 위하여 이득공간을 항해하는데 이용될 수 있다. 이득 함수의 차원은 다음의 메트릭스를 갖는 형태로 제공된다.

- *latency*(i): 첫번째 소스 워크 모듈이 실행을 시작할 때와 논리적 워크 모듈 LWM_i^k이 한번 실행을 완성할 때 이전까지의 사이에서 허용되는 최대 시간 간격
- *throughput*(i): 논리적 워크 모듈 LWM_i^k이 수행되는 비율
- *accuracy*(i): 논리적 워크 모듈 LWM_i^k에 요구되는 출력 데이타 구조의 최소 정확도
- *precision*(i): 논리적 워크 모듈 LWM_i^k에 대한 입력 데이타 구조의 최소 정밀도
- *consistency*(i, a, j, b, n): 논리적 워크 모듈 LWM_i^k의 a번째 경우의 완성 시간과 LWM_j^k의 b번째 경우의 완성 시간 사이의 최대 시간 간격
- *cost*(i): 사용자가 허용할 수 있는 비용

즉, 응용은 서비스가 진행되는 동안 각각의 축이 하나의 QoS 파라미터가 된다고 볼 때, 관련된 QoS 값들의 집합으로 구성되는 위치에 놓이게 된다. 이때, 응용은 이 위치에서 어느 축으로 기울여야 할 것인가를 결정해야 하는 데, 이 결정 과정을 기술하는

함수가 바로 이득 함수가 된다.

하나의 이득 함수의 축은 응용에 따라 좌우되는 데, 이 축들은 (그림 3)의 QoS 메트릭스로 구성될 수 있다. 정확성과 정밀성 메트릭스는 응용을 위한 출력 데이터의 정확도와 정밀도 요구사항을 지정한다. 어떤 데이터의 정확도와 정밀도는 응용에 의존적이기 때문에, 데이터 프레임의 크기를 응용에 적합한 정밀도와 정확도로 변환하기 위해 특정한 함수를 LRoS의 범위 이내에서 지정하여야 한다. 대부분의 시간성 메트릭스에 대한 정의는 분명하게 다루어진다. 일관성 메트릭스는 스트림간의 지터를 지정하기 위하여 이용된다. 예를 들면 오디오나 비디오 동기화에 대한 사용자 요구사항을 지정하는 데에 이용될 수 있다.



(그림 3) 자원관리에 이용되는 QoS 메트릭스

(Fig. 3) QoS metrics applied to proposed resource management

또한, 각각의 메트릭스는 사용자가 허용할 수 있는 변동의 범위가 얼마만큼인가를 나타내는 대응 변동 메트릭스를 가질 수 있다. 예를 들면, 사용자가 비디오 그림이 흑백 및 칼라사이에서 오가는 비디오를 원하지 않는다라는 사실을 정밀도 변동에서 표시해야 할 수도 있다.

3.2 시스템 기반 구조의 모델링

자원 구조에 대한 정확한 모델은 효과적인 자원관리를 위하여 매우 중요하다고는 할지라도, QoS 분석을 위해 시스템의 행위를 모델링 하기 위한 포괄적인 추상화 방식은 잘 알려져 있지 않다. 모델링의 필요가 있는 주된 행위로는 QoS 기반의 자원 동시성 행위가 있는 데, 이는 일련의 업무가 자원에 어떻게 작용하는가를 설명한다. 자원 동시성 행위는 기본적으로 자원의 형태(CPU, network, bus, disk), 자원 스케쥴링 및 장애 처리 정책, 업무 할당에 의하여 영향을 받는다.

이 문제에 관한 연구가 다소 있기는 했지만 주로 개별 자원에 관한 것이었다[9, 11]. 본 논문에서 제안하는 TPM 추상화 방식은 대규모 시스템의 동시성 행위를 모델링 하기 위한 것으로서, 이는 각 응용에 대한 시스템 자원의 분배 문제와 가능한 스케쥴링 방법을 QoS 자원관리자가 결정하는 데에 필요하다.

TPM 추상화는 자원 기반 구조의 동시성과 기능 및 토폴로지 특성을 반영하며, 자원의 집합과 자원 사이의 연결성을 지정하는 가지들의 집합으로 구성된다. 여기서 자원은 하나의 단위 처리(atomic processing), 통신 또는 저장 장치가 될 수 있다. 프로세서, 버스, 디스크, 네트워크 등은 전형적인 자원의 예이다. 각 자원의 스케쥴링 모델에 의해 서술되는 데, 스케쥴링 모델은 자원에서 수행하는 일련의 작업에 근거하여 자원의 동시성 행위를 지정한다.

3.3 할당 및 라우팅

자원의 토폴로지 및 속성을 토대로 자원관리기는 할당 및 라우팅 알고리즘을 이용하여 응용과 자원을 연결시킨다. 모니터링 알고리즘은 또한 모든 자원의 조건을 제시함으로서 할당 및 라우팅 알고리즘에 정보를 제공해 준다. 할당 라우팅에 필요한 의사결정은 자원의 조건이 변화할 때 동적으로 변화한다.

라우팅 알고리즘은 고정된 시작점과 종단점 사이의 최적 입출력 루트를 발견하기 위하여 네트워크 설계에서 이용되어 왔다. 고정된 시작점과 여러 종단점 사이의 멀티캐스트 라우팅에 대해서는 최근에 몇몇 연구가 있었다[7, 16]. 특히 Widyono는 멀티캐스트 라우팅 문제의 부분으로서 QoS 요구사항을 시험하였다[15]. 기존의 라우팅에 관한 연구는 응용 QoS 요구사항을 무시하고 시스템 목표만을 고려하였거나, 혹은 시스템 목표를 무시하고 응용의 QoS 요구사항만을 고려하였다[10, 11, 12]. 그러나 네트워크 기반의 멀티미디어 응용을 지원하기 위해서는 이 두가지 측면을 동시에 고려하는 라우팅 알고리즘이 요구되며, 선정되는 루트는 응용 QoS 요구사항과 시스템 목표의 최적화 즉, 부하 균등 조정(load balancing)에 근거해서 정해져야 한다. 라우팅 알고리즘은 또한 이미 루트가 결

정된 응용이 자원 실패 또는 응용의 시작 및 종료에 기인해서 다시 루트를 결정하는 상향 라우팅(up-routing) 기능을 필요로 한다.

할당 및 라우팅은 기본적으로 스케쥴링 정책에 의해 지시되기 때문에, 한 스케쥴링 정책에 의한 할당 및 라우팅 결정이 다른 스케쥴링 정책에 기인한 QoS 요구사항을 만족시키지는 못한다. 그러므로 할당 및 라우팅 알고리즘은 이질적 네트워크에 존재하는 여러가지의 서로 다른 스케쥴링 정책을 수용해야 한다.

3.4 종단간 스케쥴링

효율적인 자원 이용을 위해서는 하나의 시스템 자원이 다수 응용에 의해 공유되어 이용되도록 다중화하는 것이 필요하다. 스케쥴링 알고리즘은 특정 응용에 할당된 자원들 각각에 대하여 적절한 실행의 시점을 정하고, 다른 응용에 관계된 시간 간격동안 응용의 우선순위를 정한다. 응용의 QoS 요구사항을 만족시키는 것은 스케쥴링의 기본적인 목표이다.

여러 자원에 대하여 수행되는 응용으로 구성되는 시스템은, 응용에 시간적 요구사항을 상실하는 폭주(congestion) 현상을 야기하기 쉽다[7]. 이러한 문제는 이질적 시스템 자원의 형태하에서 수행되는 응용의 경우에 특히 심각하다. 시간적 요구사항을 보장하기 위하여 필요한 핵심적인 스케쥴링 요구사항은 이 폭주를 통제하는 것이다. 폭주는 또한 해결책을 예방하게 할 것인지, 회피하게 할 것인지, 혹은 단순히 감지하게 할 것인지 등에 따라 여러 형태로 처리될 수 있다. 즉, 하나의 방법은 폭주가 언제, 어디서 발생되고 회복되는 가를 감지만 하는 것이다. 또한 현재의 상태에 따라 폭주가 발생할 것이라고 시스템이 감지할 때, 응용의 수행율을 감소 시킴으로서 폭주를 피할 수도 있는 데, 비율 기반(rate-based) 및 신용 기반 흐름 제어(credit-based flow control)는 이 분류에 속한다.

또 다른 방법은 폭주는 결코 발생할 수 없다고 스케쥴링 정책을 미리 지정하는 것이다. 이 정책은 미리 정의된 응용 속성에 기초를 두는 데, 응용 수준과 시스템 수준 중 어느 수준에서 수행되는가에 따라 기법이 세분화된다.

3.5 QoS 분석

QoS 분석은 응용의 QoS 요구사항이 시스템에 만족되고 있는 가를 확인시켜 준다. 시뮬레이션 또는 분석적 기법이 QoS 분석에 이용될 수 있는 데, 이들 기법은 시뮬레이션 엔진의 긴 실행 시간 때문에 아주 작은 문제에만 적용할 수 있지, 현실적인 문제 해결에는 적합하지 못하다.

대분의 QoS 분석 기법들은 오로지 응용의 시간적 요구사항이 만족되는 것을 확인 하는 데에 촛점이 맞춰져 왔다. 한편, 해석적 모델에 기반한 대부분의 시간 분석 기법들은 단일 자원에서 수행되는 업무를 위하여 적용되었다. 일부의 종단간 시간 분석 방법들은 네트워크 영역에 한정되었을 뿐, 분산된 이질적 종단간 시스템의 분석에 필요한 시간 분석상의 많은 이슈들 즉, 동기화나 피드백, 복수 수행을 등을 설명하는 것과 같은 것들은 연구되지 않았다[12, 13]. S. Chetterjee et al.은 시간 분석을 위해 비율 제어 기반의 스케쥴링 정책을 수행하는 대규모 분산 시스템에 이용될 수 있는 분할정복(divide-and-conquer) 방법을 제안한 바 있다[4].

3.6 응용 QoS 요구사항의 감쇄

이상적인 환경에서 자원관리자는 응용의 이득이 최대화 되도록 하기 위하여 각 응용에 특정 수준의 QoS를 제공해야 한다. 그러나 너무나 많은 응용의 요구나 혹은 자원의 장애, 불충분한 자원 등은 이러한 이상적인 현상을 현실화 시키지 못한다. 이런 경우 대부분의 자원관리자는 단순히 QoS 요구사항을 무시하거나 혹은 응용을 제거한다.

본 논문에서는, 응용에 대한 충격을 최소화 하기 위하여 QoS를 선택적으로, 또한 점차적으로 감소시키는 방안으로서 추상화 방식과 알고리즘을 제안하는 것이다.

또한, 자원의 부하가 줄어들 때 탐색 알고리즘은 사용자에 대한 이득을 최대화 하기 위하여 어떤 QoS 매트릭스가 먼저 증가되어야 하는가도 결정하게 된다.

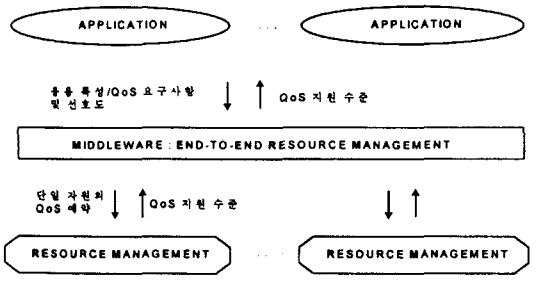
3.7 시스템 자원의 모니터링

시스템 자원을 모니터링 하는 것은 각 자원이 올바르게 동작하는 것을 확인하기 위하여 필요하다. 일시적인 장애가 발생했던 자원은 스케쥴러가 응용의 정밀도와 정확도 요구사항을 만족시키기 위하여 특정 응용의 재실행을 요구할 수도 있다. 지속적인 장애가

발생했던 자원은 그 자원을 이용하여 응용을 재분배하고 다시 루트를 설정하기 위한 할당 및 라우팅 알고리즘을 필요로 한다.

4. 원격세미나 응용 시나리오

종단간 자원관리는 (그림 4)에서와 같이 개별적인 컴퓨팅 및 통신망 자원의 관리와 연계되어 종합적인 QoS가 각각의 응용에게 제공된다. 각각의 응용은 미들웨어인 종단간 자원관리자에게 원하는 QoS 요구사항 및 선호도를 통보하고 종단간 자원관리자가 각각의 자원별 자원관리자에게 예약된 QoS들을 종합하여 해당 응용에게 최종적인 QoS 지원 수준을 통보해 주게 된다. 물론 이 과정은 QoS 협상 과정의 수행 내용으로 응용이 동작하는 동안에도 QoS 감시 및 조정 과정이 응용과 자원관리자간에 계속해서 이루어 지게 된다.



(그림 4) 종단간 QoS 지원 구조
(Fig. 4) End-to-End QoS support infrastructure

본 장에서는 앞서 제안한 모델을 대표적인 멀티미디어 응용 서비스라고 할 수 있는 원격 세미나에 적용하여 보기로 한다. 원격세미나는 분산된 사용자간 다양한 단말과 통신망들간을 연결하여 오디오 및 비디오 전송을 기반으로 한 영상회의와 사용자간 대화의 보조 수단으로 이용되는 전자칠판(whiteboard) 등의 소프트웨어 도구들을 활용한다. 자원관리 관점에서의 원격 세미나 응용의 모델링은 응용을 구성하는 다양한 소프트웨어 및 하드웨어의 요소들을 정의하고, 이들의 관계를 명시하는 데서 출발한다. 전체 시스템은 아주 작은 구성요소까지 세분화되어 세부적인 QoS가 관리될 수 있도록 있으며, 아주 커다란 몇 개의 요소로만 구분될 수도 있다. 만약 작은 많은 구성요소들을 가지게 될 경우 QoS 자체가 오버헤드가 될 가능성이 큰 반면, 커다란 구성요소들로 설계할 경우 종단간 QoS 관리가 효율적으로 달성될 수 없다는 단점이 있으므로 이들간의 트레이드 오프가 설계시 반영되어야 한다.

모델링의 전초적 단계로 원격 세미나의 흐름을 살펴 보는 것이 중요한 데, 이것은 이 흐름을 통해 필요한 구성요소나 그들 간의 관계를 자원관리 관점에서 명확히 할 수 있도록 해게 된다. 원격 세미나는 오디오와 비디오를 카메라와 마이크를 통해 읽어 들이는 것으로 시작하여 여기서 얻어진 아날로그 신호는 디지털 신호로 변환하는 과정에서 샘플링되고 이 디지털 신호는 다시 전송의 효율을 높이기 위해 압축 과정(오디오;ADPCM, 비디오;MPEG/JPEG)을 거치게 된다. 압축된 정보는 통신망을 통해 전송되기 전 미디어별 및 미디어간 동기화 과정을 거치게 되는 데 이때 전자칠판을 사용하면서 기록한 정보도 발표자의 모습이나 음성과 함께 동기화 되어야 하므로 같이 동기화 될 수 있도록 한다. 수신측에서는 통신망을 통한 오디오와 비디오 정보를 다시 압축된 정보에서 원래의 디지털 정보로 복원하는 과정을 거치며, 오디오의 경우 발표자의 출력이 다시 마이크로 입력되는 에코우를 삭제하기 위한 에코우 제거 과정을 수행하게 된다. 오디오 및 비디오, 전자칠판의 동기화를 목적으로 삽입된 시간소인 등의 동기화 정보는 수신측에서 플레이시에 다시 역으로 이용되어 모든 움직임과 소리가 일치하게 된다.

자원관리 관점에서의 원격 세미나의 모데링은 우선 이러한 응용의 전체적인 흐름과 내부적인 구성요소들, 그리고 소요 자원들을 토대로 LRoS와 LUoW들을 종단간 QoS 지원을 극대화 할 수 있도록 설계된다. 본 논문에서는 세부적인 설계 과정을 설명한다기보다는 기본 개념을 원격 세미나를 구성하는 몇 가지 중요한 구성요소들을 중심으로 제시하는 모델의 기본 개념을 설명해 보고자 한다.

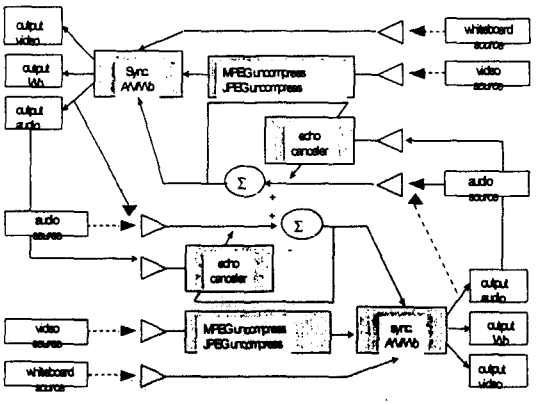
4.1 원격 세미나 응용 예의 모델링

하나의 응용은 시스템이 사용자에게 제공하는 서비스로 해석될 수 있다. 이 서비스는 자신을 위하여 시스템에 추가적인 서비스를 요구할 수도 있다. 그러

므로 각 서비스는 미리 정해진 순서대로 수행되는 추가 서비스의 집합으로 표현될 수 있다. 이러한 계층적 정의는 최하위의 서비스가 정의될 때까지 계속해서 적용된다. 다음의 예에서, 사용자가 원격 세미나 서비스를 요청하면 원격 세미나 응용은 (그림 5)에서와 같이 "echo canceler", "synhronize audio/video/whiteboard", "MPEG/JPEG uncompress" 서비스를 차례대로 요청한다.

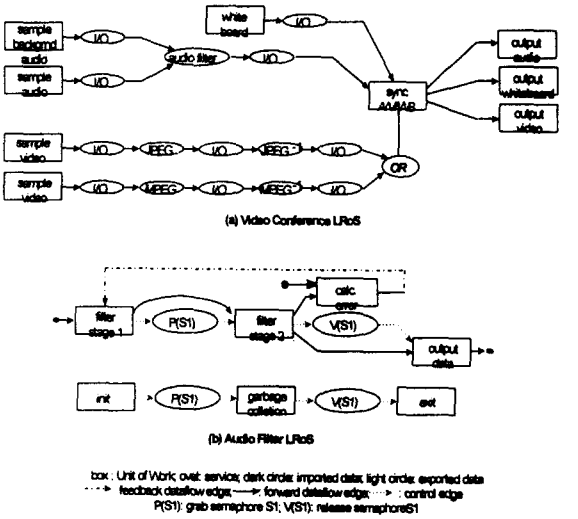
(그림 6a)에서는 원격 세미나 응용을 위한 LRoS를 보여 준다. 이 예제에서 LRoS는 오디오 데이터를 샘플링하고, 배경 잡음을 필터링하고, 여과된 오디오 신호를 비디오 데이터 및 화이트보드 데이터에 동기화하고, 동기화된 오디오-비디오-화이트보드 신호를 출력한다.

최상위 레벨에서, 대부분의 기능들은 모듈화된 서비스들을 이용하여 표현되는 데, 이것을 통해 LRoS가 복수의 플랫폼에 쉽게 이식이 가능하게 된다. 자원관리자는 최상위 LRoS에서 지정한 각 서비스를 제공하기 위하여 자신의 LRoS 데이터베이스를 탐색함으로서 적합한 LRoS를 찾게 된다. 부가적으로, 서비스와 LRoS 사이의 이러한 연결이 실행 시간 동안에 수행되기 때문에, 자원관리자는 시스템의 현재 상태에 기반한 최적의 LRoS를 선택하게 된다.



(그림 5) 텔레 세미나 응용 예
(Fig. 5) Tele-seminar application example

이러한 사항은 비디오의 경우를 이용하여 설명할 수 있다. 원격 세미나 응용은 JPEG 및 MPEG에 의한 압축을 둘 다 지원할 수 있다. 일반적으로 MPEG은 JPEG보다 더 많은 양의 비디오 데이터를 압축하지만 실행 시간은 더 오래 걸린다. 그러므로 컴퓨팅 자원은 충분하고 통신 자원이 부족하다는 사실을 자원관리자가 감지한 경우에는 MPEG 알고리즘을 이용하지만, 역으로 통신 자원이 충분하고 컴퓨팅 자원이 부족할 경우에는 JPEG 알고리즘이 이용된다.



(그림 6) 오디오-비디오 LRoS
(Fig. 6) Audio-video LRoS

한편, (그림 6)는 오디오 및 비디오 처리에 필요한 LRoS들을 도시한다. (그림 6b)의 경우 오디오 필터 서비스에 대한 LRoS들을 보여 주는 데, 이 예에서 LRoS는 5개의 LUoW로 구성되며, 시스템으로 부터 4개의 서비스를 요구하고 있다. 여기에서, semaphore S1 grab, 재설정(reset)과 같은 서비스들은 유닉스 서버의 기능을 이용한다. LRoS는 8개의 제어 신호 가지와 6개의 데이터 플로우 전방향(feed-forward) 가지, 하나의 데이터 플로우 피드백 가지를 포함하고 있다. "filter stage1" 작업 단위가 종료되면 "filter stage2"로 어떤 데이터를 보내게 된다. 부가적으로, "grab semaphore S1" 서비스에 제어 신호를 보낸다. "filter stage2" 작업 단위는 filter stage1으로부터 오는 데이터와 P(S1)으로 부터 오는 제어 신호를 받기 전에는 시작할 수 없다. 또한 "filter stage2" LUoW는 자신의 데이터를 "calc error"와 "output data" LUoWs 모두에 멀티-캐스트한다. I/O 서비스는 지정되지 않았기 때문에

LUoW들과 논리적 서비스는 모두 같은 자원에서 수행되어야 한다.

Garbage collection 알고리즘은 오디오 필터 서비스와 비동기적인 방식으로 수행한다. 데이터의 일관성을 보장하기 위하여, "garbage collection" 또는 "filter stage2"가 수행되기 이전에 재상용되지 않도록 semaphore S1을 이용한다.

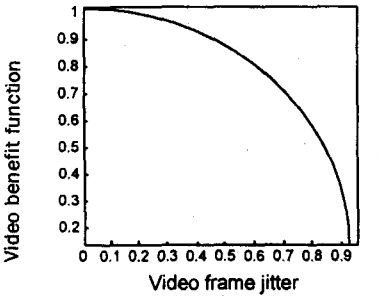
사용자는 또한 응용을 위한 QoS 요구 사항을 구체화 하여야 한다. 특정 응용이 QoS 요구 사항이나 요구 사항 각각의 상대적 중요성과 협의되어지도록 지원하기 위한 좋은 방안으로 이득 함수의 추상화가 이용된다.

(그림 7)은 원격 세미나 응용의 비디오 흐름에 대한 간단한 이득 함수를 보여 주며, 여기서 (a)는 1차원 이득 함수를 나타내고, (b)는 보다 복잡한 다차원의 이득 함수를 나타낸다.

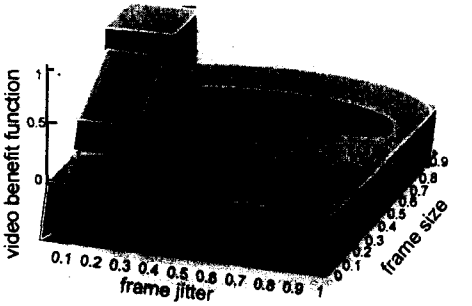
예를 들어 프레임 지터가 고정되었다고 할 때, 프레임 크기를 감소시키면 처음에는 이득의 손실이 없다가 그 후 일정한 점까지는 이득의 선형 감소가 발생하며, 이후에는 지터에 관계없이 이득이 제로 값을 갖는다. 프레임 크기가 고정되었다고 할 때, 지터를 증가시키면 처음에는 일정한 이득이 발생하지만, 어떤 점 이후부터는 이득이 로그 함수적으로 감소하기 시작한다. 지터 및 프레임 크기 메트릭스 뿐만 아니라 이득 함수 역시 이 두개의 그래프에서 정규화되었다. 오디오 및 화이트보드 요소에 대해서도 비슷한 이득 함수가 지정될 수 있을 것이다.

이득 함수는 승인 협상 단계나 일시적 과부하 상태일 때 중요하다. 이상적으로는 사용자의 이득을 최대화 해 주는 QoS를 시스템이 제공해야 하지만, 그러나 이것이 불가능할 경우, 응용 사용자 이득의 손실을 최소화 하기 위해서는 어떤 QoS 메트릭스가 어떤 순서로 감쇄되어야 하는지를 정하기 위하여 자원 관리자는 이득 함수 표면의 기울기를 이용해야 한다. (그림 7)을 이용한 예제는 이 점을 보여 준다. 이상적인 상태에서 시스템은 제로의 프레임 지터 값과 최대의 프레임 크기 값을 제공할 것이다. 그러나 만약 이것이 불가능하다면, 시스템의 우선 프레임 크기를 감소시키려고 할 것이며, 그 다음은 비디오 프레임 지터를 증가시키려고 할 것이다. 유사한 감쇄 접근 방법이 일시적인 과부하 동안 여러 QoS 메트릭스에 대하여 적용될 수 있다.

한편, 이득 함수는 응용 사용자의 이득을 위해 자원 관리자가 이용할 수도 있으며, 각각의 LRoS나 LUoW 자체 내에서 이득 함수를 가지고 스스로 판단하고 동작할 수도 있다. 예를 들어 통신망을 통하여 비디오를 전송하는 LUoW의 경우, 전송하는 비디오를 수신 단말에서 패킷 손실율이나 지연, 지터 등과 같은 QoS 측정을 하여 피드백하게 되면 전송측에서는 이 이 QoS 정보를 토대로 전송율을 조절하는 이득 함수를 동작시킬 수 있다. 이러한 과정에서 LUoW는 수집된 QoS의 파라미터 값이 변화되고 있다고 하더라도 실제 사용자 이득에 영향을 주지 않는 범위의 변화의 경우 어떠한 제어 동작을 생략할 수가 있는 것이고, 또한 여러 파라미터들 중에서 전송율 조정 등을 위해 먼저 고려해야 할 파라미터들을 선택적으로 고려하여 자원 관리 의사결정에 도움을 줄 수가 있는 것이다.



(a) Video conference LRoS



(b) Audio filter LRoS

(그림 7) 비디오 이득 함수
(Fig. 7) Video benefit function

4.2 하드웨어/소프웨어 기반 구조의 모델링

대규모 시스템을 모델링 하기 위해서는 계층적 서브 시스템 추상화를 이용한다. 하나의 서브 시스템은 자원의 집합, 자원의 연결성을 지정하는 가지들의 집합(서브 시스템 토폴리지)과 자원관리 체계로 구성된다. 하나의 서브 시스템은 서브 시스템 모델을 이용하여 표현된다.

자원은 서브 시스템 모델 혹은 디바이스 모델을 이용하여 모형화 된다. 이러한 반복적 모델링을 통하여 우리는 이질적 자원관리 체계에 의해 제어되는 컴퓨팅, 저장장치, 그리고 통신 장비로 구성되는 대규모 시스템을 모델링 할 수 있다.

장치 모델의 속성은 장치명, 장치 형태, 장稅 지향적 파라미터, 스케쥴링 정책, 스케쥴링 모드 및 장치 부하이다. 장치 형태는 2단계 수준 분류법을 분류한다. 첫째 수준은 장치가 CPU, 버스, 네트워크, 디스크 중 어느 것인가를 지정하는 것이고, 둘째는, 서브 형태 즉, CPU, 운영체제, 버스, 네트워크, 혹은 디스크의 형태 등을 지정하는 것이다. 장치 형태는 자원 할당의 목적으로 이용된다. 장치 고유의 파라미터는 특정한 장치의 서브 형태에 대한 숫자 정보를 제공한다. 예를 들면 클럭 비율, 대역폭, 명령어당 싸이클 수 등이다. 장치 부하는 자원에서 수행되는 LUoW의 집합을 나타내며, 장치 스케쥴링은 자원이 수행되는 스케쥴링 정책을 의미한다. 스케쥴링 정책은 자원에서 수행되는 LUoW를 위한 우선순위 할당을 지정하며, 장치의 실행 행위에 영향을 미친다. 장치 스케쥴링 모델은 특정 장치에 스케쥴된 부하를 나타내는 장치 고유 및 스케쥴링 정책 고유 모델이다.

5. 결 론

시스템에 의해 지원되어야 하는 QoS 요구 사항을 가지는 분산 응용이 점차로 증가하고 있다. 이러한 QoS 요구 사항들은 동적인 운용 조건하에서 분산된 이질적 자원들(CPU, 네트워크, 저장 장치 등)의 효과적인 QoS 지원을 수행하는 자원 관리자에 의해서만 만족될 수 있다. 자원 관리자는 각 응용의 부하 특성과 QoS 요구 사항을 파악할 수 있을 때 이러한 기능을 수행할 수 있다.

본 논문에서는 QoS 기반의 자원 관리 모델을 제시하고, 응용 모델에서 발견되어야 하는 속성의 형태를 연구하였다. 응용 모델은 응용 흐름의 명세, 계층적 방법에 의한 서비스 요구 사항, 서비스 각각의 자원 이용 특성, 응용이 허용할 수 있는 QoS의 범위, 시스템이 제공하는 QoS의 함수로서 응용 사용자가 받게 되는 이득을 파악할 능력이 있어야 한다.

본 논문에서는 또한 LRoS로 약칭되는 추상화 방식을 제시하였다. LRoS는, 시스템 자원관리 체계에 의해 각각의 자원이 가장 잘 이용될 수 있도록 서비스의 계층적 집합을 동적으로 결합하는 것으로 특정한 응용을 모델화하는 데에 이용된다. 또한 QoS 요구 사항은 QoS 요구 사항과 선호도를 파악하는 추상화 방식인 이득 함수를 이용하여 구체화됨을 지적하고, LRoS와 이득 함수 추상화의 유용성을 확인하기 위하여 멀티미디어 응용 예에 대한 모델을 제시하였다.

끝으로 본 논문의 제안 방식의 특징을 정리해 보면, 우선 최근의 일부 연구를 제외한 많은 자원 관리 연구들이 개별적인 자원 관리[1, 2, 9]에 초점을 두고 있는 데 비해 본 연구는 종단간의 자원 관리 뿐 아니라 응용과 하부구조를 포함하여 총체적인 관리를 위한 종합적인 모델이라고 할 수 있다. 즉, 다양한 시스템 모델과 TPM이나 LRoS, LuoW 등 여러 가지 추상화 개념을 소개하여 CPU, 통신망, 버스, 소프트웨어 등 다양한 이질적인 환경을 쉽게 수용하도록 하고 있다. 또한 다른 연구들[7, 15, 16]이 QoS 요구 사항이나 시스템 목표만을 고려한 데 비해 본 연구는 응용의 QoS 요구 사항과 시스템 목표의 최적화를 고려한다는 차이가 있다고 하겠다. 특히 자원 관리에서의 이득 함수는 멀티미디어 시스템을 인간의 지각활동과 연계시켜 자원 관리의 효율화와 사용자 및 응용 우선적인 실질적인 자원 관리를 가능하게 한다고 하겠다.

참 고 문 헌

[1] D. Anderson, R. Herrtwich, and C. Schaefer, "SRP:A Resource Reservation Protocol for Guaranteed Performance Communication in the Internet," *University of California at Berkeley Technical Report TR-90-006*, Feb. 1990.

[2] J. Beck and D. Siewiorek, "Automated Task Allocation and Processor Specification Strategies

for Multicomputer Systems," *CMU-EDRC 18-50-94 Technical Report*, 1994.

[3] J. Coolahan and N. Roussopoulos, "Timing Requirements for Time-Driven Systems Using Augmented Petri Nets," *IEEE Transactions on Software Engineering*, pp. 603-616, September 1983.

[4] S. Chatterjee and J. Strosnider, "A Generalized Admissions Control Strategy for Heterogeneous, Distributed Multimedia Systems," *Proceedings of ACM Multimedia*, 1995.

[5] D. Ferrari, "Client Requirements for Real-Time Communication Services," *IEEE Communications*, vol. 28, no. 11, Nov. 1990.

[6] D. Gilles and J. Liu, "Scheduling Tasks with AND/OR Precedence Constraints," *University of Illinois, Urbana-Champaign Technical Report UIUCDCS-R-90-1627*, March 20, 1991.

[7] S. Golestani, "Congestion-Free Transmission of Real-Time Traffic in Packet Networks," *Proceedings of INFOCOM*, San Francisco, June 1990.

[8] M. Harbour, M. Klein, and J. Lehoczky, "Fixed Priority Scheduling of Periodic Tasks with Varying Execution Priority," *Proceedings of the 1991 Real-Time Systems Symposium*, December 1991.

[9] D. Katcher, H. Arakawa, and J. Strosnider, "Engineering and Analysis of Fixed Priority Schedulers," *IEEE Transactions on Software Engineering*, 19(9), September 1993.

[10] E. Knightly, "H-BIND: A New Approach to Providing Statistical Performance Guarantees to VBR Traffic," *Proceedings of IEEE Infocom*, 1996.

[11] E. Knightly and H. Zhang, "Traffic Characterization and Switch Utilization using a Deterministic Bounding Interval Dependent Traffic Model," *Proceedings of IEEE INFOCOM*, 1995.

[12] R. Perlman, "Interconnections: Bridges and Routers," Addison-Wesley, 1992.

[13] C. Paris and D. Ferrari, "A Dynamic Connection Management Scheme for Guarantee Performance Services in Packet switching Integrated Services Networks," *University of California at Berkeley Technical Report TR-93-005*, 1993.

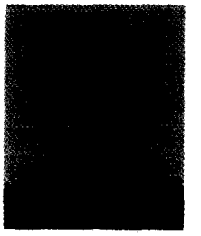
[14] B. Sabata, S. Chatterjee, M. Davis, J. Sydir, and T. Lawrence, "Taxonmy for QoS Specifications," unpublished, October 1996.

[15] R. Widyono, "The Design and Evaluation of Routing Algorithms for Real-Time Channels," *University of California at Berkeley Technical Report TR-94-024*, June 1994.

[16] H. Zhang and E. Knightly, "Providing end-to-end Statistical Performance Guarantees with Interval Dependent Stochastic Models," *Proceedings of ACM Sigmetrics*, May 1994.

궁 상 환

1977년 2월 숭실대학교 전산학과 졸업
1977년~1983년 육군 제2군수지원사령부 전산장교
1983년 고려대학교 경영대학원 전산정보 석사
1991년~현재 충북대학교 대학원 박사과정
1982년~현재 한국전자통신연구소 책임연구원(컴퓨터연구단 멀티미디어연구부 시각언어연구실)
관심분야: multicast transport, distributed system architecture등



윤 석 환

1982년 2월 아주대학교 산업공학과(공학사)
1984년 2월 건국대학교 산업공학과(공학석사)
1996년 8월 아주대학교 산업공학과(공학박사)
1992년 8월 품질관리 기술사 자격취득
1995년 1월~현재 한국정보처리학회지 편집위원
1986년 1월~현재 한국전자통신연구소 책임연구원(컴퓨터연구단 멀티미디어연구부)
관심분야: 그룹웨어, S/W 공학, 생산정보시스템, 개발방법론

황 승 구

1979년 서울대학교 전기공학과 졸업(학사)
1981년 서울대학교 전기공학과 졸업(석사)
1986년 University of Florida 전기공학과(박사)
1982년 7월~현재 한국전자통신연구원 멀티미디어 연구부 부장
1994년~1995년 미국 SRI International(International Fellow)
관심분야: 멀티미디어 시스템, HCI, 로보틱스