ISBN 978-89-5884-997-1 98560

차세대 가시화 시스템의 설계 및 구축

(Design & Construction of Next-generation Visualization System)

구 기 범 (Gee Bum Koo)
voxel@kisti.re.kr
Visualization Team, Supercomputing Center

한 국 과 학 기 술 정 보 연 구 원
Korea Institute of Science & Technology Information

제목 차례

표 차례

그림 차례

1. 서론

Visualization 시스템은 고성능 컴퓨터에서 생산된 대용량 수치 데이터를 가공해서 시각적으로 표현하는 데에 사용된다. 가시화 시스템은 그리드와 같은 차세대 컴퓨팅 환경에서 대규모 데이터를 해석할 수 있도록 도와주는 핵심 구성요소로 자리 잡고 있으며, 해외 주요 연구기관에서는 이미 일정 수준 이상의 가시화 시스템을 구축해서 이를 그리드나 e-Science 관련 프로젝트에서 활용하고 있다. 우리나라는 현재 국가 그리드 기반 구축사업(정보통신부), 국가 e-Science 구축사업(과학기술부) 등 고성능 컴퓨팅 자원 기반의 연구 환경을 구축하는 사업이 진행 중이고 KISTI에서 대규모 슈퍼컴퓨터(이론 성능 250 TFLOPS)의 도입을 추진하고 있다는 점 등을 고려하면, 국내에서도 향후 1~2년 내에 누구나 수 테라바이트 규모의 데이터를 양산할 수 있게 될 것이고, 그에 따라 고성능 가시화 시스템에 대한 수요가 증가할 것임은 어렵지 않게 예상할 수 있다.

일반적으로 가시화 작업은 CPU, GPU, 메모리, 디스크, 네트워크 등 시스템의 거의 모든 구성요소를 동시에 대량으로 사용한다는 특징을 갖고 있다. 뿐만 아니라 많은 가시화 작업이 빈번한 사용자 인터액션을 전제하기 때문에 가시화 시스템은 어지간한 작업을 실시간으로 처리할 수 있는 능력을 요구받는다. 따라서 가시화 시스템은 인터액티브 작업 수행, 가용 자원의 대규모 동시사용, 대용량 데이터의 실시간 처리 등에 최적화돼야 하며, 시스템 설계자도 이러한 사항을 염두에 두고 시스템 설계와 구축을 진행해야 한다.

본 보고서는 2007년 11월 현재 KISTI 슈퍼컴퓨팅센터가 설치하고 있는 '차세대 가시화 시스템'의 설계내용과 최종 사양에 대해 설명한다. 2장에서는 기존의 visualization 시스템이 갖고 있던 문제점을 제시하고, 이를 '차세대 가시화 시스템'에서는 어떻게 해결하는지 설명한다.

2. 차세대 가시화 시스템의 설계

이 장에서는 KISTI가 도입하는 차세대 가시화 시스템의 설계 방침과 세부 설계 내용을 설명한다. 차세대 가시화 시스템의 설계내용에는 다음의 사항들을 중점적으로 반영했다.

○ CAVE/Onyx3400을 운영하면서 드러났던 문제점을 보완한다.

○ 상대적으로 규모가 작은 tiled display를 구축/운영하면서 축적한 노하우(know-how)를 설계에 반영했다. Tiled display는 소규모 시스템이지만 visualization 시스템으로서 갖춰야 할 것은 다 구비하고 있기 때문에 차세대 가시화 시스템의 축소판으로 생각해도 크게 무리가 없다.

○ LLNL, Cal-$(IT)^2$ 등 해외 유수연구기관이 구축한 대형 visualization 시스템을 벤치마크 했다. 특히 단일 시스템의 구성뿐만 아니라 계산 전용 시스템과의 연계, 외부 네트워크의 연결 등 서로 독립적으로 운영되는 컴퓨팅 자원을 연계해서 보다 편리한 고성능 컴퓨팅 환경을 구축하는 방법에 초점을 맞췄다.

○ 주요 컴퓨터 제조업체(IBM, HP 등)에서 판매하는 visualization 솔루션뿐만 아니라 Tungsten graphics, GraphStream 등 소규모 업체이면서도 visualization 시스템 구축에 뛰어난 업체들 관계자들과의 면담을 통해서 시스템 통합에 필요한 기술을 파악했다.

가. CAVE/Onyx3400 시스템의 문제점

2001년 도입 후 2007년까지 운영해왔던 CAVE/Onyx3400 시스템은 출력장치는 가장 높은 수준의 몰입감을 제공한다는 강점을 갖고 있지만 기술적인 면이나 시스템 운영 등에 있어서 다음과 같은 문제점을 보여주었다.

○ **그래픽 성능의 한계** : Onyx3400의 그래픽 카드(SGI에서는 그래픽스 파이프라는 명칭을 사용)인 InfiniteReality3은 설치 당시에는 높은 성능을 갖고 있었다. 하지만 비슷한 시기에 programmable GPU의 개념이 소개된 후 GPU의 성능이 비약적으로 빨라지면서 불과 5년 만에 PC용 그래픽 카드 한 개가 Onyx3400의 전체 그래픽 성능(5 GPU)을 앞지르는 상황이 발생했다. 더 큰

문제는 Onyx3400이 brick 구조를 갖고 있기 때문에 이론적으로는 그래픽을 담당하는 G-brick만 교체하는 것으로 업그레이드가 가능했지만 이후 SGI에서 출시하는 시스템의 brick 자체가 변경되면서 실질적으로는 그래픽 성능의 업그레이드가 불가능했다.

○ **메인 메모리 부족** : Onyx3400은 SMP 시스템이면서 다섯 대의 프로젝터가 연결돼있기 때문에 CAVELib을 사용하면 똑같은 애플리케이션과 데이터를 5개 중복해서 실행시켜야한다는 단점이 있다. 따라서 6GB의 메모리가 장착되어 있다고 해도 이론적으로는 한 애플리케이션이 사용할 수 있는 메모리는 1.2GB로 줄어들고, 실제로는 1GB를 넘기기도 어렵다는 문제가 있었다. COVISE를 이용해서 천문 데이터를 렌더링할 때 Onyx3400에서는 단 한 장의 그림을 만들어내는 것도 버거웠던 것에 비해 PC에서는 전체 데이터를 렌더링하는 데에도 전혀 무리가 없었다.

○ **슈퍼컴퓨터와의 연계 미흡** : 고성능 컴퓨팅 환경의 구성요소로서의 visualization 시스템은 주 계산 시스템과의 효과적인 데이터 공유방법이 반드시 필요하다. 좋은 예로 HP의 SVA(Scalable Visualization Architecture)는 visualization 시스템이 사실상 계산 시스템의 일부로 포함되는 구성을 보여주고 있으며 LLNL의 GAUSS 시스템은 900TB 용량의 스토리지를 공유하고 있다. 하지만 Onyx3400은 단순히 기가비트 이더넷을 통해서 로그인 정도만 가능한 수준의 연계만 구현됐다.

○ **출력장치의 범용성 부족** : CAVE 형태의 장비는 완전한 몰입감을 제공한다는 장점이 있지만 소수의 가상현실 애플리케이션 이외의 애플리케이션에 대해서는 크게 유용하지 않다는 단점도 있다.

○ **외부 사용자의 지원**

나. 차세대 가시화 시스템의 문제점 보완방법

Visualization 시스템에서 실행되는 대부분의 작업은 실시간으로 제어되면서 CPU, 메모리, GPU, 네트워크 등 시스템의 가용자원을 대량으로 동시에 소비한다는 특징을 갖고 있다. 따라서 가시화 시스템은 아래와 같이 대형 인터액티브 작업의 수행에 최적화돼야 한다.

○ **최신 GPU 장착 및 업그레이드** : 시스템 설치 시점 기준으로 가장 높은 성능을 갖는 GPU를 장착한다는 점에서는 Onyx3400과 다른 점이 없으나, 시스템 설치 후 적절한 시점에 GPU의 성능을 업그레이드함으로써 시스템의 운영이 종료되는 시점까지 일정한 수준 이상의 그래픽 성능을 유지하도록 했다.

○ **성능한계 극복** : CAVE를 운영하는 컴퓨터였던 Onyx3400은 도입 당시의 기준으로도 가시화 시스템으로서 가장 높은 성능을 갖는다고 말하기 어려웠다. 하지만 차세대 가시화 시스템은 성능 면에서 미국의 주요 국립연구소에서 운영하고 있는 가시화 시스템과 비교해도 손색이 없을 정도의 성능 수준을 보유하고 있다.

○ **대용량 메모리** : CAVELib을 운영함에도 불구하고, 각 노드의 성능을 현재 구현할 수 있는 최고 수준으로 올렸기 때문에 실제로 각 애플리케이션이 사용할 수 있는 메인메모리용량이 50배 이상 많아졌다.

○ **스토리지 공유를 통한 슈퍼컴퓨터와의 연계 강화** : 슈퍼컴퓨터 4호기의 스토리지를 직접 접근할 수 있도록 해서 계산이 모두 끝난 후 가시화 시스템에서 해당 데이터를 바로 가공할 수 있도록 했다. 계산 시스템의 일부로 포함되는 방안도 강구할 수 있으나 여러 가지 정황상 그렇게 할 수 없었다.

○ **범용적인 출력장치 채택** : CAVE 형태의 장비는 특수한 형태의 애플리케이션 외에는 사용이 거의 불가능했지만 범용성을 갖는 cylindrical 스크린을 채택해서 다양한 형태의 애플리케이션을 사용할 수 있도록 했다.

○ **외부 사용자의 지원 강화** : 비록 입체영상을 볼 수는 없지만 외부 사용자도 가시화 시스템의 컴퓨팅 능력을 사용할 수 있도록 지원하기 위한 장치를 마련했다.

이 외에도 다음과 같은 특징을 갖도록 했다.

○ **실시간 그래픽 처리 능력** : 가시화 시스템은 대용량 데이터를 최소 15 fps 이상의 속도로 렌더링하기 위해 다수의 GPU를 동시에 활용할 수 있어야 한다. 뿐만 아니라 이미 사용되고 있는 범용 가시화 소프트웨어나 일반 OpenGL 애플리케이션도 별다른 수정 없이 모든 GPU의 능력을 동원할 수 있도록 지원해야 한다. 물론 GPU의 수가 증가함에 따라 그래픽 처리능력도 선형적으로 증가하는 확장성(scalability)도 갖춰야 한다.

○ **빠른 응답시간(response time)** : 다수의 GPU를 사용할 때에는 데이터 분산과 이미지 합성 과정이 반드시 필요하게 된다. 이 때 하드웨어 관점에서 가장 중요하게 고려해야 할 사항은 그래픽 메모리에서 메인 메모리로 그림(partial image)을 전송하는 readback 속도와 노드 간 네트워크의 지연시간(latency)이고, 소프트웨어 관점에서는 효율적인 데이터 분산과 이미지 합성 알고리즘, 그리고 노드 간 부하균형(load balancing)을 높은 우선순위에 둬야 한다. 실시간 렌더링에 있어서는 노드 간 네트워크의 지연시간이 성능을 크게 좌우하기 때문에 클러스터를 구축할 때 이를 반영한 노드 간 네트워크를 구축해야 한다.

○ **대용량 데이터의 실시간 가공** : 시뮬레이션에 의해 만들어진 수치 데이터가 그림으로 표현되려면 먼저 수치 데이터를 그래픽스 하드웨어가 처리할 수 있는 형태로 가공하는 작업이 필요하다. 가시화 시스템에서 CPU만으로 데이터를 가공할 때 실시간 처리를 실현하기는 쉽지 않지만, view dependent rendering과 같이 OpenGL 하드웨어가 필요로 하는 데이터를 지속적으로 제공할 필요가 있을 때에는 CPU 작업이라고 해도 거의 실시간 처리에 준하는 성능을 보여줘야 한다.

여기에 더해서 성능과는 직접 관련이 없지만 가시화 시스템에 대해서만 발생하는 요구사항도 존재한다.

○ **데이터의 실시간 외부 전송** : 일부 소프트웨어는 가시화 시스템에 대해 렌더링 이미지를 외부의 호스트에 전송할 수 있는 능력을 요구한다(예: EVL의 SAGE). 일반적으로 클러스터는 사설 IP 어드레스를 이용하지만 데이터의 외부 전송을 위해 특정 노드에 대해 공용 IP 어드레스를 같이 할당하거나 전용 게이트웨이를 설치해야 하는데, 보안을 위해서 전용 게이트웨이를 운영하는 것이 바람직하다.

○ **가상현실** : 사용자가 애플리케이션의 실행을 손쉽게 제어하고, 데이터의 구조 파악을 용이하게 하기 위해서 다양한 가상현실 입/출력 장치를 갖추는 것이 바람직하다. 여기에는 헤드 트래커, 완드 등의 입력장치와 입체영상을 출력할 수 있는 프로젝터 등이 포함된다.

다. 소프트웨어의 리소스 사용 패턴

Visualization 시스템의 구성을 결정하는 데에는 하드웨어 성능이나 구조에 대한 요구사항뿐만 아니라 해당 시스템에서 운용되는 소프트웨어를 원활하게 수행하기 위한 지원방안도 중요한 변수로 작용한다. 따라서 visualization 시스템에서 어떤 소프트웨어가 운용되며, 각 소프트웨어가 어떤 종류의 자원(CPU, GPU, 네트워크 등)을 집중적으로 사용하는지에 대한 정보를 정리할 필요가 있다. 표 [2-1]은 현재까지 CAVE, tiled display 등을 운용하면서 사용했던 소프트웨어들을 실행방법과 주로 사용하는 리소스에 따라서 정리한 것이다.

[표 2-1] 가시화 시스템에서 운영 가능한 소프트웨어의 실행형태

업무	애플리케이션 실행 형태	주요 리소스
일반적인 data visualization	• Interactive MPI job 지원 • Multi-GPU 동시 사용 • 외부와의 고속 네트워크 연결	• 가상현실 입/출력 장치 • GPU • 내부 (고속) 네트워크
고행상도 비디오 스트리밍	• 외부로부터 다수의 고해상도 A/V 스트림을 동시에 전송받아서 스크린에 출력 • HD 카메라의 영상을 외부로 전송	• 가상현실 입/출력 장치 • 외부 네트워크 • 오디오 장비
고해상도 데스크톱	• 프레젠테이션, 동영상 상영 등 일반적인 데스크톱 화면에서 수행하는 작업과 동일	• 프로젝터 / 스크린 • 내부 (고속) 네트워크 • 오디오 장비
일반 수치계산	• 여타의 계산 시스템에서 수행하는, 그래픽 작업이 없는 장시간 계산 • Batch scheduler • 예: render farm, simulation 등	• CPU • 메모리 • 디스크 • 내부 (고속) 네트워크
Steering	• 슈퍼컴퓨터가 생성한 시뮬레이션 데이터를 전송받아서 가공한 후 그 결과를 화면에 출력하되, 모든 작업을 실시간으로 처리 • 슈퍼컴퓨터에서 실행되고 있는 시뮬레이션을 원격제어	• 가상현실 입/출력 장치 • 슈퍼컴퓨터 • CPU • 메모리
GPU 컴퓨팅	• GPU를 이용한 장시간의 계산	• GPU

라. 차세대 가시화 시스템 설계

앞에서 설명한 내용을 종합해서 다음과 같이 시스템을 설계했다.

1) 차세대 가시화 시스템의 특징

여기서는 일반적인 계산 전용 클러스터나 Onyx3400에 비교해서 차세대 가시화 시스템이 갖는 주요 특징을 설명한다.

- 가격대비 성능이 좋은 클러스터 형태로 구축하되 클러스터 환경을 지원하지 않는 소프트웨어를 위해 일정 규모 이상의 SMP 시스템을 노드로 포함시킨다. 이 구성은 CAVELib이나 VR Juggler와 같이 각 노드에 애플리케이션과 데이터를 중복시켜서 실행하는 애플리케이션에게 유리하다.
- 필요에 따라서 interactive 작업과 batch 작업을 실행할 수 있도록 스케줄러 등을 지원한다.
- 일반적인 클러스터는 외부로 대용량의 데이터를 전송하는 것을 고려하지 않지만 visualization 시스템은 필요에 따라서는 실시간으로 고해상도 영상이나 데이터를 외부에 직접 전송할 수 있도록 외부 네트워크와의 연계를 보강했다.
- 필요에 따라서 massively parallel 작업까지 수행할 수 있도록 수 테라플롭스 수준의 계산 능력과 그에 준하는 메인 메모리를 갖도록 했으며, 그래픽 처리능력도 다수의 GPU를 데이터 가시화 작업에 동원할 수 있도록 했다.
- trackd와 CAVELib, VR Juggler로 구동 가능한 트래킹 장비를 지원하고, 입체 영상 출력이 가능한 프로젝터와 스크린, 5.1채널의 오디오 출력 등 가상현실 애플리케이션 실행도 충분히 지원한다.
- HD급 영상 입/출력 : 4k 해상도의 영상 출력

2) 시스템 구성

[표 2-2]는 가시화 시스템의 물리적인 구성요소를 보여준다. 가시화 전용 컴퓨팅 시스템에는 각 노드에 고성능 GPU가 장착된다는 특징이 있는데, 고성능 GPU가 장착된 그래픽 카드는 블레이드나 1U 서버에 장착이 불가능한 크기를 갖고 있기 때문에 공간 효율이 높은 시스템을 구성하기 어렵다는 단점이 있다. 이 외

에도 일반적인 콘솔에 더해서 입체영상을 출력할 수 있는 전용 출력장치가 연결되고, 가상현실 입력 장치와 음향 시스템 등 일반 계산 전용 시스템에서는 보기 어려운 장비들이 추가된다는 특징이 있다.

[표 2-2] 가시화 시스템의 구성요소

종 류	개 요	특이사항
전용 컴퓨팅 시스템	• 다수의 CPU와 고성능 GPU • 대용량 메모리 • 대용량 외부 저장장치 • Interconnection network	클러스터
출력장치	• 프로젝터 • 대형 스크린	입체영상 지원
외부 네트워크	• 슈퍼컴퓨터와 가시화 시스템을 연결하는 초고속 네트워크	
가상현실 입력장치	• 가상현실 애플리케이션의 원활한 실행을 위한 3차원 입력장치	
음향 시스템	• 5.1 채널 이상의 음향 입/출력 시스템	

특히 가상현실 애플리케이션의 원활한 실행을 위해 trackd와 CAVELib, VR Juggler의 지원을 최우선 순위로 두었다.

3) 컴퓨팅 시스템

가시화 시스템 중 컴퓨팅 시스템의 구성은 [그림 2-1]과 같다.

[표 2-3] 가시화 전용 컴퓨팅 시스템의 노드 구분

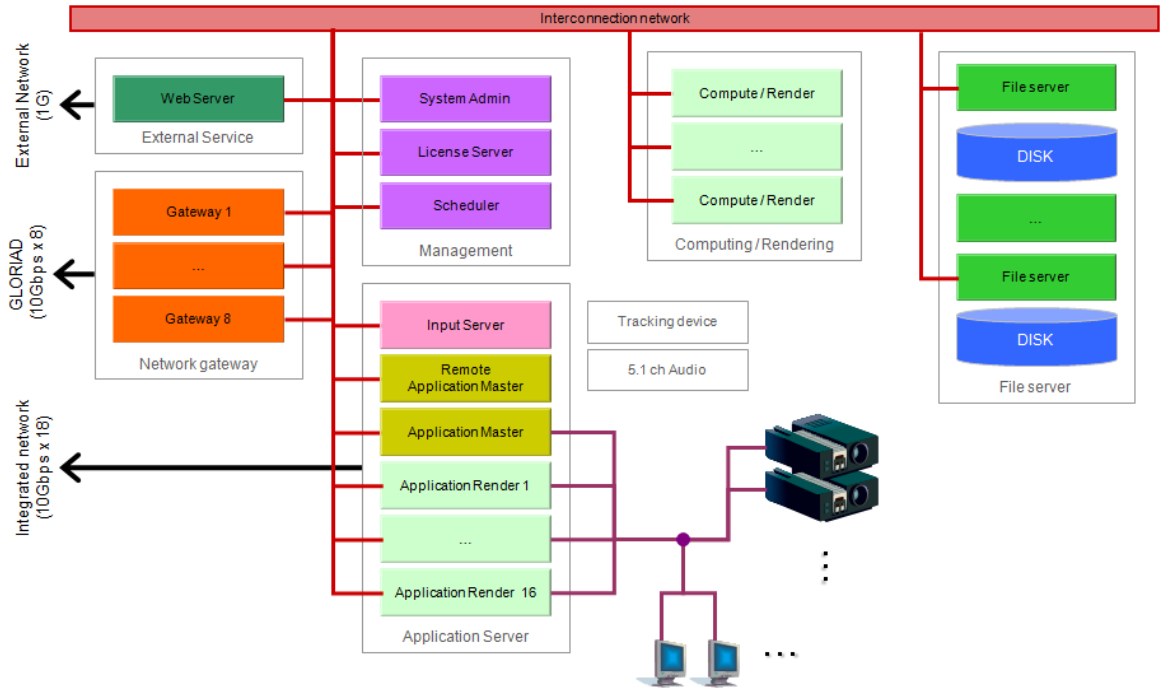
그룹 명칭	노드 명칭	용도
외부 서비스	로그인 노드	• 외부 사용자 접속 • 프로그램 컴파일 및 테스트
	관리 서버	• 시스템 관리 전용 서버
	웹 서버	• 시스템 현황 등 제반 정보 제공 (외부 사용자)
게이트웨이	게이트웨이	• 내부 노드의 외부 네트워크 접속 지원
애플리케이션 서버	라이선스 서버	• 상용 애플리케이션을 위한 라이선스 서버
	스케쥴러	• Batch job 스케쥴링 전용 서버
	입력장치 서버	• 헤드 트래커, 완드 등의 입력장치 전용 서버
	로컬 애플리케이션 마스터	• User interactive 애플리케이션 실행의 시작점
	원격 애플리케이션 마스터	• User interactive 애플리케이션 실행의 시작점
	컴퓨팅/렌더링 노드	• 병렬 계산 / 병렬 렌더링을 수행하는 노드의 집합
파일 서버	파일 서버	• 내부 노드에 파일 서비스 제공

[표 2-3]은 가시화 전용 컴퓨팅 시스템을 구성하는 노드의 역할에 대해 설명하고 있다. 대부분의 노드는 일반 계산 전용 클러스터와 유사하지만 애플리케이션 마스터, 입력장치 서버는 가시화 전용 컴퓨팅 시스템에서만 볼 수 있는 노드다.

◯ 애플리케이션 마스터 : 모든 인터액티브 작업의 시작점이 되는 노드다. 이 노드는 클러스터 환경을 지원하지 못하는 가시화 소프트웨어(예: OpenDX)로도 일정 규모 이상의 데이터에 대한 가시화 작업을 수행할 수 있도록 하기위해 설치한다. 따라서 일반 노드보다 4배 이상 많은 CPU 코어와 메모리를 갖춘 SMP 시스템으로 구성하는 것이 바람직하다.

◯ 입력장치 서버 : 헤드 트래커, 완드 등 3차원 가상현실 입력장치가 연결되는 노드다. 시스템의 구성에 따라서 애플리케이션 마스터로 포함될 수도 있고, 독립 노드로 존재하면서 입력장치 신호를 애플리케이션 마스터에 네트워크로 전송해 주는 식으로 구성할 수도 있다.

◯ 게이트웨이 : 컴퓨팅/렌더링 노드에서도 중간 결과를 외부에 전송할 수 있도록 하기 위해 외부 네트워크로 나가는 관문을 운영할 필요가 있다.



[그림 2-1] 가시화 컴퓨팅 시스템의 노드 구성

3. 차세대 가시화 시스템 세부사양

이 장에서는 KISTI 슈퍼컴퓨팅센터가 도입하는 차세대 가시화 시스템의 구성과 세부사양에 대해 설명한다. 이 장에서 설명하는 내용은 2007년 11월까지 확정된 사양이며, 최종 설치가 완료되는 시점에는 약간의 변화가 있을 수도 있다. 변경된 세부사양은 다음 보고서에 반영할 예정이다.

가. 컴퓨팅 시스템

1) 노드 구성

차세대 가시화 시스템의 노드는 [표 3-1]과 같이 구분했다. 최초 구상 단계에는 별도의 로그인 노드와 콤포지터가 필요할 것으로 예상했으나, 설계과정에서 로그인 노드는 원격 애플리케이션 마스터로 대체하고 콤포지터는 소프트웨어 솔루션으로 대체하기로 했다. 파일 서버도 슈퍼컴퓨터 4호기의 스토리지를 액세스할 수 있도록 해주는 외부 파일 서버와 내부 파일서버를 분리했으나, 충분한 성능(대역폭)을 구현하는 것이 거의 불가능한 것으로 판단해서 외부 파일 서버를 없애고 애플리케이션 서버들만 직접 통합 네트워크를 통해 4호기의 스토리지에 접근할 수 있도록 했다.

[표 3-1] 가시화 컴퓨팅 시스템의 노드 구분

그룹 명칭	노드 명칭	용도
외부 서비스	웹 서버	○ 시스템 현황 등 제반 정보 제공
게이트웨이	게이트웨이	○ 내부 노드의 외부 네트워크 접속 지원
관리	관리 서버	○ 시스템 관리 서버
	라이선스 서버	○ 상용 애플리케이션 라이선스 서버
	스케줄러	○ 배치 작업 스케줄링 서버
애플리케이션	입력 장치 서버	○ 완드 등 입력 장치 운영 서버
	애플리케이션 마스터	○ 인터액티브 작업 수행
	애플리케이션 렌더링	
	원격 애플리케이션 마스터	○ 원격 사용자의 인터액티브 작업 수행
컴퓨팅/렌더링	컴퓨팅/렌더링 노드	○ 계산 / 렌더링을 수행하는 노드
파일 서버	파일 서버	○ 내부 노드에 스토리지 서비스 제공

2) 노드 사양

2007년 하반기에 인텔과 AMD가 각각 새로운 칩셋과 CPU를 발표할 예정이었기 때문에 노드 사양을 결정하는 것이 어려운 문제였다. 특히 5년 이상의 운영기간을 생각하면 이미 출시돼있는 제품을 사용하는 것은 성능과 업그레이드의 한계가 분명히 존재하기 때문에 바람직하지 않은 접근방법이다. 하지만 출시 예정제품을 선택하는 것도 정확한 출시일자(=시스템 설치일자)를 예측하기 어렵고, 상대적으로 도입가격이 높으며, 하드웨어의 안정성 등이 충분히 검증되지 않은 제품을 선택하는 것이기 때문에 원활한 시스템 운영에 지장을 줄 수 있다는 단점이 있다.

차세대 가시화 시스템은 애플리케이션 서버, 컴퓨팅/렌더링 노드 등 성능과 직접 관련이 있는 노드는 출시 예정제품을 채용해서 높은 성능을 유지하고, 웹 서버 등 높은 성능을 필요로 하지 않는 제품에 대해서는 기존 제품을 사용해서 도입가격을 낮추는 방법을 채택했다.

○ **외부 서비스 노드 그룹** : 외부 서비스 노드에는 웹 서버만 존재한다.

[표 3-2] 웹 서비스 사양

노드 명칭	CPU	RAM	GPU
웹 서버	AMD Opteron 2218	8 GB	메인보드 내장

○ **게이트웨이 그룹** : 게이트웨이는 CPU 성능은 조금 중요할 수 있겠으나 메인메모리는 크게 중요하지 않다. 그 대신 인피니밴드와 10G 이더넷을 동시에 지원해야 하므로 이를 위해 PCI express 슬롯이 두 개 이상 존재해야 한다.

[표 3-3] 게이트웨이 노드 사양

노드 명칭	CPU	RAM	GPU
게이트웨이	AMD Opteron 2218	8 GB	메인보드 내장

○ **관리 서버 그룹** : 관리 서버는 모두 높은 성능을 필요로 하지는 않는다.

[표 3-4] 관리서버 사양

노드 명칭	CPU	RAM	GPU
관리 서버	AMD Opteron 2218	4 GB	메인보드 내장
라이선스 서버	AMD Opteron 2218	4 GB	메인보드 내장
스케줄러	AMD Opteron 2218	4 GB	메인보드 내장

○ **애플리케이션 서버 그룹** : 애플리케이션 마스터는 사용자가 키보드와 마우스를 이용해서 직접 애플리케이션을 실행하기 위해 존재한다. 애플리케이션 렌더링 노드는 모두 프로젝터와 직접 연결돼서 영상을 출력하는 역할을 담당한다. CAVELib의 특성상 애플리케이션 렌더링 노드와 애플리케이션 마스터 노드는 모두 동일한 사양을 갖고 있는 것이 바람직하며, 가능한 많은 메모리와 CPU를 확보하는 것이 유리하다. 그리고 애플리케이션 마스터와 애플리케이션 렌더링 노드는 영상 출력에 있어서 각 노드 사이의 동기화(synchronization)가 이루어져야 하기 때문에 별도의 추가 장비를 장착한다.

[표 3-5] 애플리케이션 서버 사양

노드 명칭	CPU	RAM	GPU
입력 장치 서버	AMD Opteron		메인보드 내장
애플리케이션 마스터	Intel Xeon 5450 3.0 GHz	64 GB/node	NVIDIA QuadroFX5600 G-Sync 채택
애플리케이션 렌더링			
원격 애플리케이션 마스터			

○ **컴퓨팅/렌더링 노드 그룹** : 컴퓨팅/렌더링 노드는 필요에 따라서 그래픽 처리, 일반 수치계산, GPU를 이용한 계산 등 다양하게 사용될 수 있기 때문에 비교적 높은 성능을 유지하고 있어야 한다.

[표 3-6] 컴퓨팅/렌더링 노드 사양

노드 명칭	CPU	RAM	GPU
컴퓨팅/렌더링 노드	Intel Xeon 5450 3.0 GHz	32 GB/node	NVIDIA QuadroFX5600

○ **파일 서버 그룹**

[표 3-7] 파일서버 사양

노드 명칭	CPU	RAM	GPU
파일 서버	AMD Opteron 2220	16 GB	메인보드 내장

3) 스토리지 & 파일시스템

스토리지는 물리적으로 RAID가 장착된 다수의 파일 서버로 구성한다. RAID 스토리지가 직접 연결되는 파일서버는 총 20대이며, 각 파일서버에는 3대의 RAID 스토리지(Lustre에서의 OST)가 연결된다. 각 RAID 스토리지에는 750GB 용량의 SATA 하드디스크 10개가 장착된다. 따라서 전체 물리용량은 450TB에 이른다. 이외에도 파일시스템을 구성한 후 메타데이터를 제공하기 위해 두 대의 MDS 서버를 운용한다.

클러스터를 위한 파일시스템은 IBM의 GPFS, Lustre, PanFS 등 다양한 솔루션이 존재하지만 실제로 디스크 I/O가 잦은 환경에서 운용할 때 충분한 안정성과 성능을 제공하는 솔루션은 의외로 찾아보기 어려운 것이 사실이다. 차세대 가시화 시스템에서는 우선 Lustre를 채용했고, 전체 throughput은 12GB/sec 수준이 될 것으로 예상한다.

나. 네트워크

1) 노드 간 네트워크 (interconnection network)

노드 간 네트워크로는 Infiniband 4X DDR을 사용한다. 대역폭은 단 방향 20Gbps를 제공한다.

[표 3-8] 인피니밴드 사양

구 분	내 용
방 식	InfiniBand
총 스위치 수량[대]	1
포트 종류 및 개수[개]	4X DDR / 156개
모델명	QLogic InfiniIO 9280
확장슬롯	24개
IB Leaf Card	12 Port InfiniBand 4X DDR (20Gbps)
Chassis Bandwidth	11.52Tbps Full Duplex
Switching	Cut-through
Switching Latency	140ns - 420ns
Power	N+3 Redundant
Power consumption	Up to 2100W
Spine Board	Full CBB for each Port
Subnet Manager	Embedded Subnet Manager

2) 외부 네트워크

외부 네트워크는 슈퍼컴퓨터 4호기를 도입하면서 구축되는 '통합 네트워크'와 GLORIAD의 두 가지가 존재하고, 모두 10Gbps 이더넷을 사용한다. 차세대 가시화 시스템은 통합 네트워크를 통해서 슈퍼컴퓨터 4호기의 스토리지에 접근할 수 있도록 구축한다. GLORIAD는 해외 연구기관의 시스템과 연계해서 실험할 때 주로 사용하는데, 가시화 시스템의 모든 노드는 게이트웨이 노드를 통해서 GLORIAD를 접속할 수 있도록 했고, 전체 대역폭은 80Gbps에 이른다.

3) 관리 네트워크

관리 네트워크는 일반 기가비트 이더넷을 사용한다.

[표 3-9] 관리 네트워크 스위치 사양

모델명	Netgear GS748TS
포트제공	48개. 최대 288개 확장가능
Switch Backbone Bandwidth	96 Gbps
Stack Bandwidth	20 Gbps
Latency	20 Microsec 이하
크기	1U
Routing	140ns - 420ns 이하
QOS	Layer 3 (DSCP) Quality of Service (QoS)
MTBF	84,000 시간

4) 콘솔 네트워크

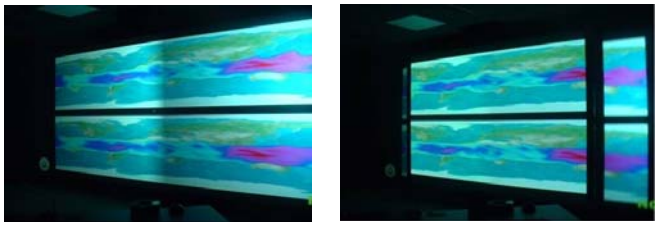
콘솔 네트워크는 다음과 같이 구성된다.

다. 출력장치

출력장치는 크게 프로젝터와 스크린으로 구성되는데 프로젝터는 지금까지 출시된 프로젝터 중 가장 높은 해상도(4096 x 2160)를 지원하는 SONY 프로젝터를 사용하고, 스크린은 실린더 형태로 구축한다. 가시화 룸의 크기에 따른 스크린의 크기, 프로젝터 영상의 aspect ratio 등을 고려해보면 2채널이 최적의 구성이었으며, 여기에 입체영상을 위해 총 4대의 프로젝터를 설치한다. (SONY 프로젝터는 active stereo를 지원하지 않는다)

입체영상 구현은 active, passive, INFITEC stereo 등을 모두 고려했는데, active stereo는 SONY 프로젝터가 아예 지원하지 않았기 때문에 고려 대상에서 제외시켰다.

Passive stereo와 INFITEC stereo는 사용하는 프로젝터의 수에는 차이가 없고, 편광필터



[그림 3-1] 2채널 edge-blended screen

와 INFITEC 필터 중 어떤 것을 사용하느냐에 따라 달라진다. 일반적으로 빛 손실은 passive stereo가 유리하지만 hot spot, edge blending을 적용할 때의 image quality 등을 고려해서 INFITEC stereo를 적용하기로 했다.

1) 프로젝터

출력장치는 SONY의 4k 프로젝터로 구성한다. 2007년 6월 기준으로 4k 급 해상도(4,096 x 2,160)를 지원하는 프로젝터는 SONY만 출시한 상태였고, Barco, Christie 등 주요 프로젝터 제조업체는 극장용 솔루션을 위해 2k 급 해상도의 프로젝터나 active stereo를 지원하는 HD급 해상도 프로젝터를 내놓고 있었다. JVC는 당시 4k 출력을 지원하는 DI-ILA 칩을 독자적으로 생산해서 이를 적용한 프로젝터를 시연제품 수준으로만 갖고 있었다.

[표 3-10] SONY 4k 프로젝터 사양

항목	내용	비고
모델명	SRX-S110	
제조업체	SONY	
해상도	4,096 x 2,160	
밝기	10,000 ANSI Lumen	
Contrast	1,800 : 1	
무게	110 kg	
호스트와의 연결	DVI	4개의 DVI 포트 동시사용

2) 스크린

스크린의 형태는 cylindrical 스크린으로 정했다. 가시화 룸 설치 예정공간이 일반 사무실로 쓰기에는 전혀 문제가 없지만 front projection을 구현하기 위한 공간으로는 높이가 충분하지 않기 때문에 rear projection 방식을 선택했다. 프로젝터를 2채널로 구성하기 때문에 edge-blending을 필요로 하는데, 두 채널의 blending 영역은 18%로 정했다.

[표 3-11] 스크린 사양

		스크린과 프로젝터	비고
모델명		AeroPlex	
제조업체		Stewart	
재질		Rigid real acrylic screen	
Gain		1.0	그림 참고
규격(m)	반지름	12.152 / 12.253 (내/외)	스크린 두께로 인한 차이
	높이	2.314	
	Arclength	7.961	

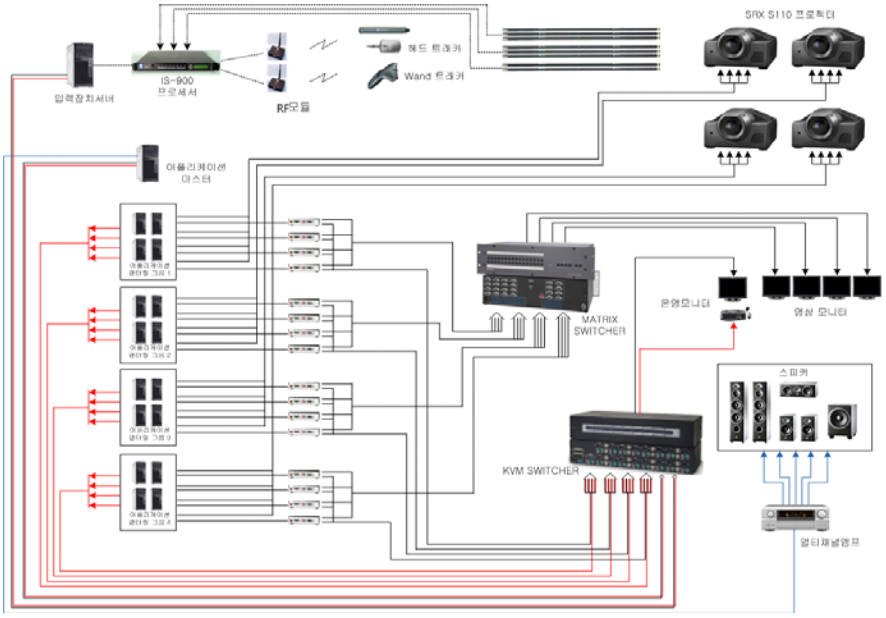
3) Edge-blending

출력장치를 LCD가 아닌 프로젝터를 사용했고, 2채널로 구성했기 때문에 edge blending 솔루션이 반드시 필요한데, 짝수 채널로 출력장치를 구성할 경우 스크린의 정면에 blended 영역이 존재하기 때문에 가능한 좋은 image quality를 보장하는 솔루션을 사용해야 한다. 그리고 edge blending 솔루션 중 4k 해상도를 완전히 지원하는 제품을 선택해야 한다는 문제도 있다. 가시화 시스템에서는 3D Perception사의 프로세서와 Mechdyne사의 OptiBlend® 기술을 조합해서 edge blending을 구현할 계획이다.

디지털 프로젝터는 검정색의 표현이 어렵고, 특히 다채널의 blended edge의 존재는 검정색의 표현을 더 어렵게 한다. 이 를 해결하기 위해 OptiBlend는 특수한 광학 마스크로 blend 된 영역에 도달하는 빛을 조절해서 더 부드러운 blend 영역을 구현한다. 여기에 디지털 edge blending 솔루션을 조합해서 검정색의 표현 수준을 높임과 동시에 blend 영역의 image quality를 더욱 향상시킨다.

4) Visualization solution의 구성

Visualization 솔루션은 [그림 3-2]와 같이 구성된다.



[그림 3-2] Visualization solution 구성도

라. 트래킹 장비

트래킹 장비는 원래 CAVE에서 사용하던 IS-900을 그대로 사용하기로 했지만 도입한지 5년이 넘은 장비에 대한 무상 유지보수가 불가능하다는 제조사의 입장에 따라 신형 IS-900을 새로 도입해야 했다. 최신의 트래킹 장비 중 광학식 장비의 경우 주위 조명의 영향을 받을 수 있고, 모션 캡쳐 장비는 '슈퍼컴퓨팅센터'의 역할을 고려해볼 때 필요성이 낮다는 점 등을 고려해서 가장 익숙한 IS-900을 그대로 채용했다.

[표 3-12] IS-900 SimTracker 무선 트래킹 시스템

장비	비고
MicroTrax head tracker	
MicroTrax wand tracker	
6-feet revision 3 SoniStrips	8개를 천정에 설치
MicroTrax recharging MicroTrax device cradle	무선 장비를 위한 충전기 세트

[표 3-13] Tracking device 상세 사양

Degree of freedom		6 with MicroTrax™
Angular range		Full 360° ~ all axis
Resolution	Standard	0.75 mm, 0.05°
	Wireless	1.5 mm, 0.10°
Static accuracy	Standard	2.0 ~ 3.0 mm 0.25° RMS in pitch & roll, 0.50° RMS in yaw
	Wireless	3.0 ~ 5.0 mm 0.50° RMS in pitch & roll, 1.00° RMS in yaw
Update rate	Standard	180 Hz
	Wireless	120 Hz
Latency		4 ms typical
Maximum tracked devices		4 for SimTracker

4. 가시화 룸

가시화 룸에 대해 가장 중요하게 생각하는 조건은 사용자가 장시간 머무르면서 작업할 수 있는 환경을 조성하는 것이다. 처음에는 KISTI 본원의 본관 2층의 빈 공간을 가시화 룸으로 이용하는 방안을 고려했으나, 컴퓨터와 프로젝터 사이의 거리가 지나치게 멀어지고, 좋은 환경을 구성하기 어렵다는 문제가 있어서 (구) 슈퍼컴퓨팅사업실과 (구)슈퍼컴퓨팅센터장실을 이용하는 것으로 최종 결정을 내렸다.

설치 예정 장소는 평범한 사무실이었기 때문에 쾌적한 환경을 만들 수 있다는 장점이 있지만 천정 높이가 충분하지 않아서 rear projection 형태로 스크린과 프로젝터를 설치해야 했다.

Rear projection은 프로젝터에 의해 소비되는 공간을 최소로 줄이기 위해서 거울을 사용할 수 있는데, hot spot 등 image quality가 낮아지는 부작용이 발생할 수 있기 때문에 거울은 사용하지 않기로 했다. 그 대신 프로젝터에 장착할 수 있는 렌즈 중 가장 초점거리가 짧은 렌즈를 이용해서 불필요한 공간을 최소로 줄이고자 했다.

[그림 4-1] 가시화 룸 조감도(좌상, 우) 및 공사 중인 가시화 룸(좌하)

가시화 룸에는 신규 시스템뿐만 아니라 현재 운영하고 있는 tiled display도 이전해서 한 공간 내에서 필요한 모든 형태의 visualization 자원을 사용할 수 있도록 지원할 예정이다.

5. 기반시설

차세대 가시화 시스템의 기반시설은 슈퍼컴퓨터 4호기의 기반시설 중 일부를 활용한다. 전체적으로 보면 그래픽 카드 때문에 단위 노드의 전력소요가 계산 전용 노드보다 조금 많은 편이고, 한 대의 캐비닛이 수용하는 노드의 수가 적어서 하중은 큰 문제가 되지 않는 대신 그만큼 면적을 많이 차지하는 문제가 있다.

가. 전기

컴퓨팅 시스템의 전기 소요는 다음과 같다. 절대적인 노드의 수가 많지 않기 때문에 큰 문제가 되지는 않는다.

- 정격 전력용량 : 149.6 kW
- 전류 : 1395A(단위) / 698A(운전)

그리고 출력장치를 구성하는 프로젝터가 대당 30A의 전기를 필요로 해서 프로젝터만 120A의 전기를 필요로 한다.

나. 설치 공간

가시화 시스템은 모든 노드에 full-size 그래픽 카드가 장착되기 때문에 블레이드나 1U 서버로는 노드를 구성하기 어렵다. 따라서 대규모 계산 전용 시스템에 비해 집적도가 상대적으로 떨어지고, 같은 성능을 구현하기 위해 보다 더 많은 공간을 차지하므로 사전에 설치 공간을 충분히 확보해야 한다.

차세대 가시화 시스템은 20개의 캐비닛으로 구성돼있으며, 각 캐비닛은 표준 크기를 갖고 있으므로 전체 설치면적은 $12m^2$가 된다. 다른 클러스터도 마찬가지지만 캐비닛의 배치에 따라서 고속 네트워크 케이블의 가격이 많이 차이가 날 수 있다.

다. 하중

앞에서 설명했듯이 가시화 시스템은 시스템 집적도가 낮지만 그만큼 하중에 대한 부담이 적은 편이다.

[표 5-1] 캐비닛 별 하중

캐비닛 번호	노드 구성	노드 수 (대)	하중(kg)
1 ~ 11	컴퓨팅/렌더링 노드	8	338
12	네트워크 스위치		201
13 ~ 16	파일서버(OSD)	5	606
17 ~ 18	애플리케이션 서버	8	338
19	애플리케이션 서버	3	126
20	파일서버, 기타서버	파일서버 2, 기타서버 15	284

6. 기타사항

과거 CAVE/Onyx3400의 문제는 GPU(그래픽스 파이프)의 사후 업그레이드가 이뤄지지 않았기 때문에 설치 후 4년이 경과한 시점에서는 사실상의 데이터 visualization 작업을 진행할 수 없을 정도로 낮은 성능을 갖고 있다. 차세대 가시화 시스템을 도입할 때에는 설치 후 일정 기간이 경과한 시점에 전체 GPU의 업그레이드를 진행하기로 했다.

7. 결론 및 향후계획

본 보고서는 2007년 KISTI 슈퍼컴퓨팅센터의 차세대 가시화 도입 프로젝트를 진행한 중간결과 중 시스템의 세부 사양을 중심으로 다루었다. 차세대 가시화 시스템의 설치는 2008년 1월 20일에 완료할 예정이며, 모든 설치가 마무리되면 시스템의 최종 사양, 성능 등을 중심으로 세부 분석을 수행할 계획이다.