국가 슈퍼컴퓨팅 공동활용체제 구축을 위한 글로벌공유파일시스템 성능 분석

Performance Analysis of Global Shared Filesystem for the PLSI

우준, 박석중*, 이상동, 김형식*
한국과학기술정보연구원, 충남대학교*

Woo Joon, Park SeokJung*, Lee SangDong,
Kim HyongShik*
KISTI, Chungnam Univ.*

요약

국내 슈퍼컴퓨팅자원을 상호 연동하여 활용성을 극대화하기 위해 추진되고 있는 국가슈퍼컴퓨팅공동활용체제 구축사업(PLSI)에서는 워크플로우에 따라 여러 기관의 자원을 사용하여 응용 시뮬레이션 및 가시화를 수행하기 위해 병렬파일시스템을 통해서 글로벌한 액세스를 가능하게 하는 데이터 공유 인프라의 구축이 요구되었다. 이에 따라, 본 연구에서는 KISTI 및 부산대 슈퍼컴퓨팅센터 간에 병렬파일시스템인 GPFS를 기반으로 리모트 파일시스템의 데이터를 상호 공유할 수 있는 글로벌공유파일시스템 테스트베드를 구축하고, 양 기관의 연동망으로 1Gbps급 WAN에서 네트워크 및 파일시스템의 성능을 분석하였다.

Abstract

The purpose of the PLSI(Patnership & Leadership for the national Supercomputing Infarastructure) is to maximize a utilization of public supercomputing resources by linking with each other. When someone performs a simulation and visualization of an application using it's resources on each sites, it needs to construct the infrastructure, so that afford to access the data globally. So, in this research, I implemented the global shared filesystem mutually to share remote filesystem's data between KISTI and Pusan National University's supercomputing center based on GPFS of parallel file system, and analyzed a performance of network and filesystem on 1Gbps WAN

I. 서 론

1. 국가슈퍼컴퓨팅공동활용체제구축사업

국가슈퍼컴퓨팅공동활용체제구축사업(PLSI)은 국내 슈퍼컴퓨팅자원을 상호 연동하여 국가 과학기술 개발에 활용함으로써 공공 슈퍼컴퓨팅자원의 활용을 극대화 하기 위해 그림 1과 같은 자원연동 계획에 따라 KISTI를 중심으로 추진되고 있는 프로젝트이다[1]. 2007년 현재 서울대, 부산대, 기상청과 MOU를 체결하고 각 기관의 가용 자원을 확보하여 시범적으로 태풍 시뮬레이션과 같은 대규모 응용과제를 수행하기 위해 인프라를 구축하고 있으며, 이러한 인프라 중 효율적인 데이터 공유를 위해 글로벌공유파일시스템의 구축을 추진하고 있다.

▶▶ 그림 1. PLSI의 자원연동 계획

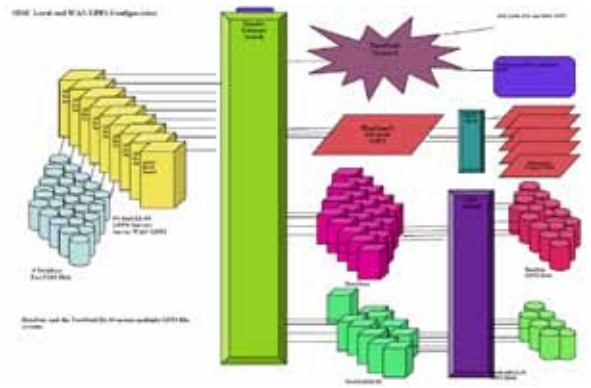
2. 글로벌공유파일시스템

LAN에 연동된 KISTI 내부 이기종 슈퍼컴퓨터 간 또는 WAN에 연동된 KISTI와 타 기관의 슈퍼컴퓨터 간 데이터의 공유를 통한 효율적이고 편리한 슈퍼컴퓨팅 자원 연동을 위해서는 기존 GridFTP, SRB(Storage Resource Broker), NFS(Network File System) 등 보다는 성능, 안정성, 이기종 간 호환성 등이 보장될 수 있는 병렬파일시스템 기반의 글로벌 공유파일시스템의 구축 이 필요하다. 특히, 미국의 TeraGrid, NERSC(National Energy Research Scientific Computing Center), NCAR(National Center for Atmospheric Research), 유럽의 DEISA(Distributed European Infrastructure for Supercomputing) 등과 같은 해외의 그리드 및 멀티 클러스터 서비스 환경에서는 이러한 글로벌 공유파일시스템을 각 서비스 모델의 핵심 구성 요소로 활용하고 있으며, 아래는 이러한 해외 구축 사례에 대하여 소개하고 있다.

2.1 TeraGrid GPFS-WAN

TeraGrid는 미국 과학재단의 지원으로 미국 내 8개 대학

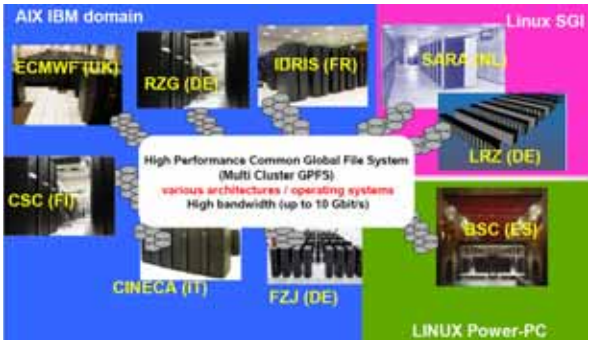
슈퍼컴퓨팅센터를 비롯 11개 슈퍼컴퓨팅 센터를 연결한 초대형 그리드로 100TFlops 이상의 자원이 연동되어 있으며, 병렬파일시스템인 GPFS를 WAN(30Gbps)을 통해서 3개 기관(SDSC, NCSA, ANL)의 시스템에 마운트하고 있으며, 해당 기관의 모든 노드에서 접근 가능하다[2].



▶▶ 그림 2. TeraGrid GPFS-WAN 구성도

2.2 DEISA GPFS-WAN

DEISA(Distributed European Infrastructure for Supercomputing)는 유럽 지역의 11개 슈퍼컴센터가 주축이 되어 구축된 그리드로 145TF의 자원이 연동되어 있다. TeraGrid와 마찬가지로 GPFS를 기반으로 WAN(1-10 Gbps)을 통해서 글로벌한 데이터 공유 환경을 제공하고 있다[3]. 현재 다양한 이기종 클라이언트에서 GPFS에 직접 액세스 할 수 있도록 포팅하는 작업을 진행 중이다.

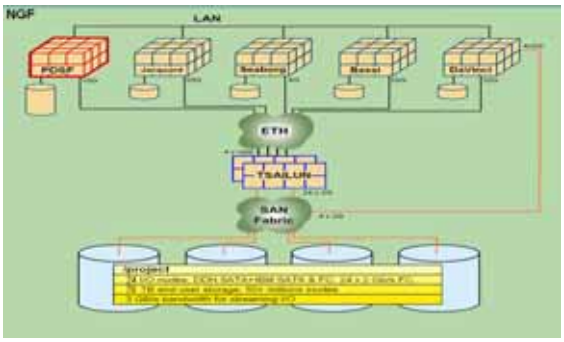


▶▶ 그림 3. DEISA GPFS-WAN 구성도

2.3 NGF(NERSC Global Filesystem)

NERSC(National Energy Research Scientific Computing Center)는 미국 에너지성 산하 슈퍼컴퓨팅센터로서 수년 간의 GUPFS(Global Unified Parallel File System) 프로젝트를 통하여 NGF(NERSC Global Filesystem)를 구축을 위한 기술 조사 및 시험을 완료하고, 최근 서비스 가능한 NGF를 구축하였다[4].

NGF는 GPFS를 기반으로 로컬 이기종(AIX/Linux) 자원 간 파일시스템을 공유하고 있으며, 지속적으로 I/O 성능 및 디스크 용량을 확장하는 등 업그레이드를 추진할 예정이다.



▶▶ 그림 4. NGF(NERSC Global Filesystem) 구성도

Ⅱ. 글로벌공유파일시스템 테스트베드 구성

본 연구에서는 본격적인 글로벌공유파일시스템 시범 서비스 환경 구축 전에 병렬파일시스템이 WAN 구간에서 어떤 성능 특성을 보이는 지 파악하기 위해 KISTI-부산대 간 1Gbps급 전용망에서 네트워크 및 파일 I/O 성능을 분석하였다. 이러한 성능 분석을 수행하기 위해 IBM의 병렬파일시스템인 GPFS를 기반으로 그림 5와 같은 테스트베드를 구축하였다. 이 테스트베드는 표 1과 같이 3개의 독립적인 GPFS 클러스터로 구성되어 KISTI에 X4500 및 KGFS 클러스터와 부산대에 PGFS 클러스터가 존재한다. X4500 클러스터가 자체 디스크를 기반으로 파일시스템 서비스를 제공하는 GPFS 서버의 역할을 수행하였고 KGFS, PGFS 2개의 GPFS 클러스터가 X4500의 리모트 파일시스템을 마운트하여 GPFS 리모트 클라이언트로 동작하였다.

[표 2] 글로벌공유파일시스템 테스트베드 사양

GPFS 클러스터명	사양
X4500	SUN X4500 1대 디스크 SATA 500GB*48개
KGFS	IBM x335 5대
PGFS	Dell PE1950 5대

▶▶ 그림 5. 글로벌공유파일시스템 테스트베드 구성도

Ⅲ. 글로벌공유파일시스템 성능 분석

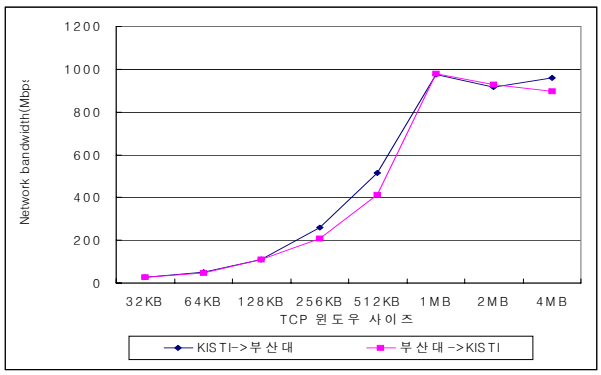
KISTI-부산대 간 WAN 구간을 통한 글로벌공유파일시스템 구축은 1Gbps WAN에 비해 상대적으로 높은 대역폭을 가지는 LAN에서 파일 I/O 성능이 WAN 성능을 상회하므로 WAN 구간의 대역폭이 I/O 성능 증가의 장애 요인이 될 것으로 예상되었다. 이에 따라, 본 장에서는 WAN 구간의 네트워크 성능 및 글로벌공유파일시스템의 파일 I/O 성능을 측정하였으며, WAN 구간의 네트워크 성능에 영향을 미치는 네트워크 파라미터들을 분석하였다.

1. 네트워크 성능

네트워크의 성능은 Full Bandwidth와 Response Time에 크게 영향을 받게 되는데 특히 Bandwidth가 높고 Response Time이 긴 네트워크(Long Fat Network : LFN)에서는 일반적인 TCP 설정으로는 Full Bandwidth 만큼의 성능을 낼 수 없다.

이러한 LFN에서 Achievable Bandwidth를 높이기 위해서는 TCP의 윈도우 크기를 적절하게 변경시켜줘야 한다. 송신지와 수신지 사이의 물리적 매체를 논리적으로 하나의 파이프로 생각 했을 때, 파이프에 담길 수 있는 최대의 크기를 Bandwidth Delay Product(BDP)라 부르며, BDP 크기는 Full Bandwidth * RTT(Round Trip Time)로 계산 할 수 있다[5]. 적절한 TCP 윈도우 크기는 BDP * 2 로 언급되고 있으며, KISTI-부산대 WAN 구간의 Full Bandwidth는 1Gbps, RTT는 4.8ms로 적절한 TCP 윈도우 크기는 1.2MBytes가 될 것으로 예측되었다.

그림 6의 KISTI와 부산대의 네트워크 성능 측정 결과에서 보는 것처럼 윈도우 크기를 변경하여 1MB에서 achievable bandwidth가 네트워크의 full bandwidth에 근접하는 것을 볼 수 있다.



▶▶ 그림 6. TCP 윈도우 크기에 따른 네트워크 성능

이와 더불어 아래와 같은 TCP 커널 변수를 조정하는 것이 네트워크 성능에 많은 영향을 끼침을 알 수 있었다.

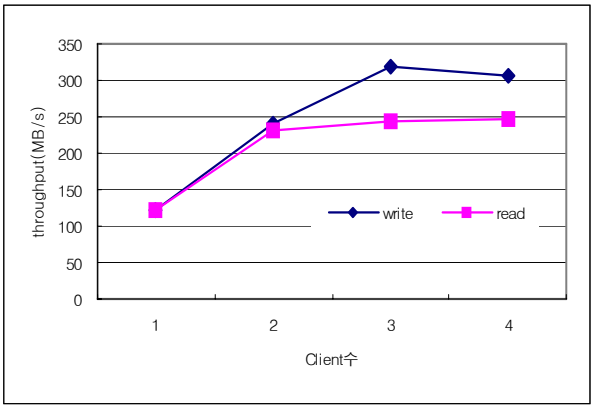
- rmem_max : 최대 receive 윈도우 사이즈
- wmem_max : 최대 send 윈도우 사이즈
- tcp_rmem : TCP receive 버퍼에 할당된 메모리
- tcp_wmem : TCP send 버퍼에 할당된 메모리
- tcp_window_scaling : 윈도우 창 배율
- tcp_sack : Selective Acknowledgement
- lp_no_pmtu_disc : 가장 큰 MTU설정 옵션

2. 파일 I/O 성능

파일 I/O 성능 측정은 KISTI X4500(GPFS 서버)의 공유파일시스템(/mnt/midgpfs)을 KGFS(GPFS 클라이언트)와 부산대 PGFS(리모트 GPFS 클라이언트)에 마운트 하여 수행하였다.

파일 I/O의 성능은 워크로드가 되는 데이터의 크기, 블록 크기, 액세스 패턴 등 여러 가지 요소에 의해 영향을 받게 된다. 여기서는 과학응용 애플리케이션의 일반적인 워크로드를 가정하여 파일크기 4GByte, 블록크기 1MByte, 순차적인 액세스 패턴에서 네트워크 대역폭을 충분히 활용하기 위해 클라이언트의 수를 변수로 하고 테스트를 진행하였다.

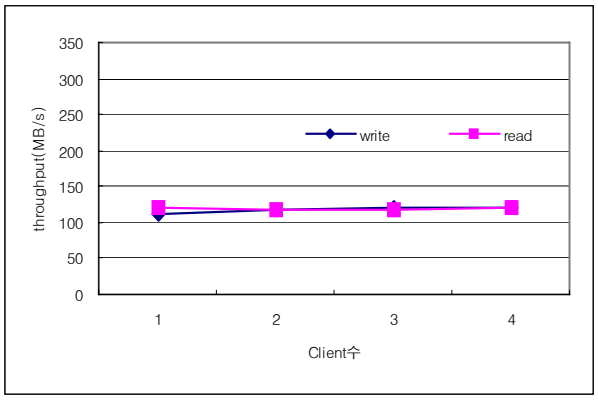
첫 번째로 KISTI X4500(GPFS 서버)와 KGFS(GPFS 클라이언트) 간 I/O 성능을 측정하였다. 그림 7의 그래프에서 볼 수 있듯이 KGFS 클라이언트 수가 증가함에 따라 쓰기의 경우는 클라이언트 3대에서 읽기의 경우는 2대에서 I/O throughput 성능이 각각 약 310MB/sec와 250MB/sec로 수렴하고 있다. 이 경우 X4500과 KGFS 간 네트워크의 이론 대역폭이 4Gbps(500MB/s)이기 때문에 X4500의 I/O 성능이 병목구간이 되고 있다.



▶▶ 그림 7. X4500(GPFS 서버)과 KGFS(GPFS 클라이언트) 간 I/O 성능

두 번째로 KISTI X4500(GPFS 서버)과 부산대 PGFS(리

모트 GPFS 클라이언트) 간 I/O 성능을 측정하였으며, 그림 8의 그래프에서와 같이 WAN 구간의 네트워크 성능인 1Gbps(125MB/sec)에 근접하고 있다. 따라서, WAN 대역폭을 늘리지 않는 한 더 이상의 I/O 성능 개선은 어렵다.



▶▶ 그림 8. X4500(GPFS 서버)과 PGFS(GPFS 클라이언트) 간 I/O 성능

Ⅳ. 결 론

본 연구에서는 국가슈퍼컴퓨팅공동활용체제구축사업에서 여러 서비스 모델의 핵심 구성 요소로 효율적인 데이터 액세스 환경을 제공할 것으로 예상되는 글로벌공유파일시스템의 구축을 위해 테스트베드를 구성하여 네트워크 및 파일 I/O 성능을 분석하였다.

결론적으로 KISTI-부산대 간 WAN을 통한 글로벌공유파일시스템의 파일 I/O 성능에 가장 큰 영향을 미치는 것은 WAN 구간 대역폭이지만, 네트워크 대역폭을 최대한 활용하기 위해서는 TCP 파라미터와 최적의 값을 도출해야 함을 알게 되었다. 이에 따라, 네트워크 및 파일 I/O 성능 분석을 통해 이러한 파라미터와 값을 찾을 수 있었고, WAN 구간의 네트워크 대역폭에 근접하는 I/O 성능을 얻을 수 있었다.

향후, 본 연구를 바탕으로 2007년 말까지 글로벌공유파일시스템 시범 서비스 환경을 구축할 예정이다. 하지만, 실제 서비스 환경에서는 I/O 성능 이외에도 글로벌공유파일시스템의 안정성과 보안 문제, 서비스 정책 등도 고려해야 될 것으로 예상된다.

▌참 고 문 헌▐

[1] 이상동, “국가슈퍼컴퓨팅공동활용체제구축사업”, 사업소개 발표자료, pp30, 2007
[2] Phil Andrews, Patricia Kovach, Christopher Jordan, “Massive High-Performance Global File Systems for Grid computing”, SC05, 2005.
[3] Thomas Bonisch, “The DEISA Infrastructure”, HLRS, 2006.
[4] NERSC, “The NERSC Global File System”, NERSC, 2006.
[5] 권기환, 한대희, 조기현, 오영도, 서준석, 손동철, 이지수, “데이터 그리드를 위한 네트웍 퍼포먼스 측정”, 한국정보과학회 춘계학술발표회, pp1.2-10, 2004.