PLSI 글로벌파일시스템에서 데이터 네트워크 성능 분석

우준*, 장지훈*, 홍태영*
*한국과학기술정보연구원
e-mail : wjnadia@kisti.re.kr, jangoq@kisti.re.kr, tyhong@kisti.re.kr

Performance analysis of data network at the PLSI global file system

Joon Woo*, Ji-Hoon Jang*, Tae-young Hong*
* KISTI

요 약

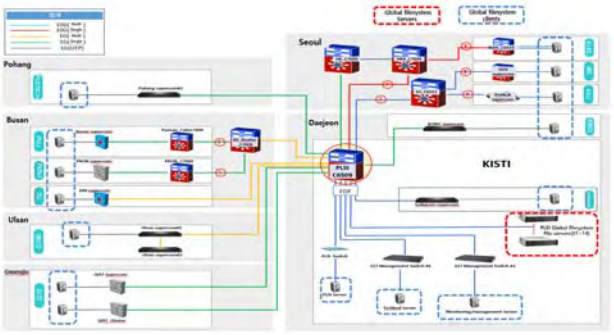
PLSI 통합 슈퍼컴퓨팅 서비스 환경에서는 다수의 사이트에서 클러스터 시스템 간 데이터 공유를 위해 글로벌 파일시스템을 사용하고 있으나, 수백 노드 이상의 클라이언트와 파일 서버 간 통신이 이루어지는 병렬 I/O에서 네트워크 병목 현상 이 발생할 수 있다. 따라서, 본 연구에서는 네트워크 병목 현상이 PLSI 글로벌 파일 시스템의 I/O 성능에 미치는 영향을 분석한다. PLSI 글로벌 파일시스템 테스트 베드에서 실험을 통해 네트워크 스위치의 버퍼 크기가 병목 현상을 유발하며, 네트워크 스위치의 버퍼 용량을 증가하여 I/O 성능을 개선할 수 있음을 보여준다.

1. 서론

국가슈퍼컴퓨팅공동활용체제구축(이하 PLSI) 사업에서는 2009년 이후 국내 주요 슈퍼컴퓨팅센터의 유휴 계산 자원의 상호 연동을 통해 국내 계산과학 분야 연구자에게 통합 슈퍼컴퓨팅 서비스 환경을 제공하고 있다. 또한, 다수의 사이트에서 클러스터 시스템 간 데이터 공유를 위해 글로벌 파일시스템을 사용하고 있다. 지금까지는 글로벌 파일시스템의 I/O 성능 향상을 위해 주로 병렬파일시스템과 디스크 스토리지 중심의 성능 분석 및 최적화에 주력했다. 하지만, 본 연구에서는 PLSI 글로벌 파일 시스템과 유사한 테스트 베드의 구성을 통해 I/O 성능에 영향을 주는 LAN 구간에서 데이터 네트워크의 병목 원인을 분석하고 개선 방안을 도출했다.

2. 글로벌파일시스템

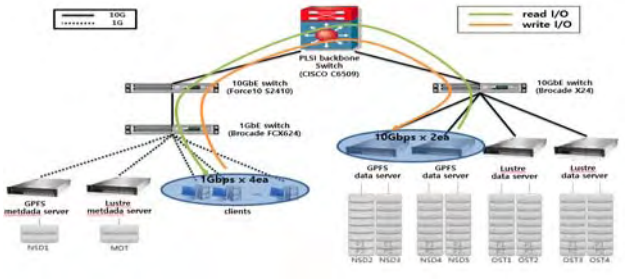
PLSI 파일 공유 서비스[1]는 (그림 1)과 같이 병렬파일시스템인 GPFS와 디스크 스토리지가 설치된 파일 공유 서비스 서버를 구축하고, 타 기관의 클라이언트가 파일 시스템에 직접 접근하는 파일 공유 환경을 구현하고 있다. 하지만, 파일 액세스를 위해 주로 다수의 클라이언트(1GbE)와 파일 공유 서비스 서버(10GbE) 간 여러 트래픽 플로우에 의한 패킷 전송이 동시에 이루어지므로, 데이터 네트워크의 병목이 파일 액세스 성능에 많은 영향을 줄 수 있다.



(그림 1) PLSI 네트워크 구성도

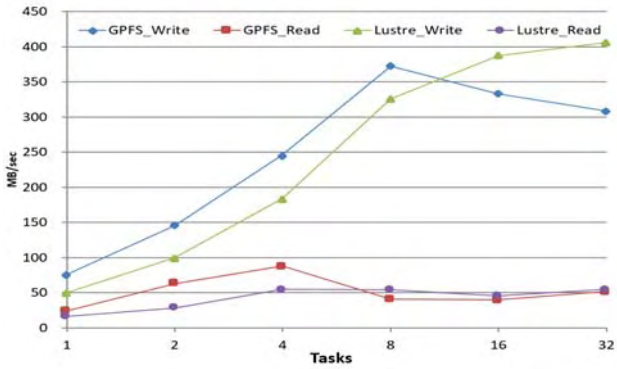
2. 성능 분석

PLSI 네트워크의 성능의 원활한 분석을 위해 (그림 2)와 같이 PLSI 공유 파일 서비스 환경과 유사하게 파일 공유 서비스 서버와 클라이언트 간 통신을 위해 1GbE(클라이언트)와 10GbE(서버) 기반의 데이터 네트워크 구성을 가지는 테스트베드를 구축했다. 파일 액세스 성능 측정을 위해 IOzone[2] 벤치마크 프로그램을 사용 했으며, GPFS[3]와 Lustre[4] 병렬 파일 시스템 각각에 대하여 타스크 수 증가에 따른 성능을 측정했다.



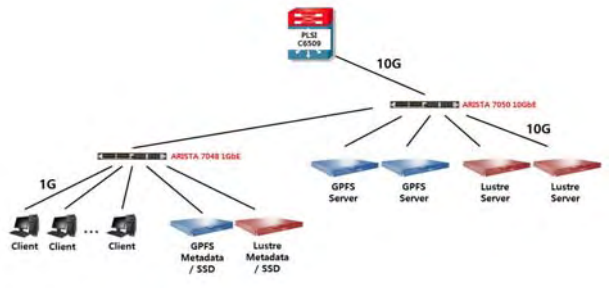
(그림 2) PLSI 글로벌파일시스템 테스트베드 구성도

(그림 3)와 같이 파일 액세스 성능 측정 결과는 병렬파일시스템의 종류와 무관하게 파일 읽기 성능이 파일 쓰기 성능의 약 4분의 1로 현격히 떨어져, 네트워크 구간에서 최소 이론 전송 대역폭인 512MB/sec의 약 5분의 1 에 불과하다.



(그림 3) 데이터 네트워크 재구성 전 파일 액세스 성능

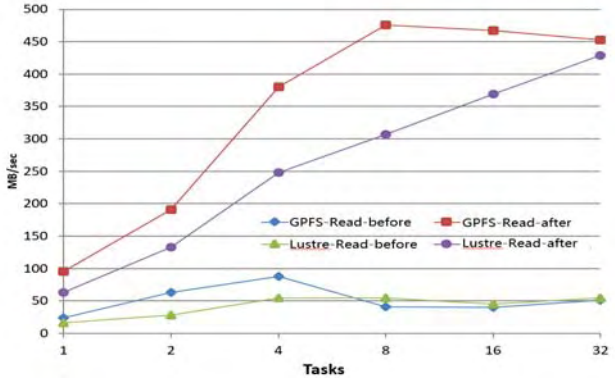
(그림 2)와 같이 전송 구간에서 대역폭이 작아지는 1 GBE 스위치가 병목 지점으로 주목되었다. 이 스위치는 겨우 3MB크기의 버퍼를 가지므로, 네트워크 부하 증가에 따른 스위치 버퍼 부족으로 플로우 별로 서로 다른 대역폭이 할당되어 네트워크 성능이 저하 되는 TCP/IP bandwidth capture 효과[5]를 유발할 수 있다. 따라서, 이로 인한 네트워크 성능 저하가 파일 읽기 성능 저하의 원인으로 보인다.



(그림 4) 데이터 네트워크가 재구성된 PLSI 글로벌파일시스템 테스트베드 구성도

네트워크 병목 현상에 의한 파일 읽기 성능 저하를 개선하기 위해, (그림 4)와 같이 문제의 원인으로 지목된 클라이언트가 연결된 1 GbE 스위치를 768MB의 대용량 버퍼를 가진 스위치로 대체 구성했다.

버퍼가 큰 스위치로 재구성된 테스트베드에서 파일 읽기 성능은, Figure 5 와 같이 재구성 이후 약 450 MB/sec를 상회하여 최소 4배 이상 향상 되었다.



(그림 5) 데이터 네트워크 재구성 후 파일 액세스 성능

4. 결론

본 연구에서는 PLSI 파일 공유 서비스의 성능에 영향을 줄 수 있는 데이터 네트워크 병목 요인을 분석했다. 서비스 환경과 유사한 테스트베드에서 측정된 파일 읽기 성능이 파일 쓰기 성능 대비 현저히 떨어지는 원인이 데이터 네트워크에 대한 부적절한 튜닝에서 비롯됨을 알게 되었다. 데이터 네트워크를 구성하는 스위치의 버퍼 용량을 증가시키면 파일 읽기 성능을 4배 이상 개선할 수 있었고, 이는 데이터 네트워크가 파일 액세스 성능에 미치는 영향이 결코 과소평가될 수 없음을 의미한다.

참고문헌

[1] Hyong-Shik Kim et al., “Development of a Next-Generation Data Sharing System for PLSI”, KISTI project report, 2012.

[2] IOzone Filesystem Benchmark, available at http://www.iozone.org

[3] F. Schmuck, R. Haskin, “GPFS: A Shared-Disk File System for Large Computing Clusters”, Proceedings of the Conference on File and Storage Technologies(FAST’02), pp. 231-244, Jan. 2002.

[4] P. J. Braam, “The Lustre storage architecture”, http://www.lustre.org/documentation.html, Cluster File Systems, Inc., Aug. 2004.

[5] Andreas Bechtolsheim, Lincoln Dale, Hugh Holbrook, Ang Li, “Why big data needs big buffer switches”, white paper, Arista networks, 2016.