HPC 환경에서 인터커넥션 네트워크 장애관리 시스템 구축

홍태영*, 윤준원*
*한국과학기술정보연구원 슈퍼컴퓨팅본부
e-mail:tyhong@kisti.re.kr

Fault Management System for Interconnection Network in HPC Environment

TaeYeong Hong*, JunWeon Yoon*
*Supercomputing Center, KISTI

요 약

KISTI 슈퍼컴퓨터 4호기 Tachyon2는 SUN Blade 6275 시스템을 기반으로 구성된 초병렬 컴퓨팅 시스템으로 이론최고성능(Rpeak) 300TFlops를 보이고 있으며 3,200대의 컴퓨팅 노드와 인프라 노드로 구분된다. Tachyon2 시스템은 국내 산학연 연구자들을 위한 공공 목적의 시스템으로 만여 명의 사용자와 200여개의 기관이 사용 중에 있다. 이런 슈퍼컴퓨터와 같은 대형 HPC 환경에서는 대규모의 사용자 작업을 원활하게 수행하기 위해서는 IB의 안정성이 우선적으로 보장되어야 한다.

본 논문에서는 Tachyon2 시스템에서 발생하는 IB 상태를 파악하고 관리하기 위한 자동화 도구를 개발하였다. 이로써 인터커넥션의 상태를 주기적으로 모니터링 할 수 있고, 장애내역 또한 신속하게 파악할 수 있다.

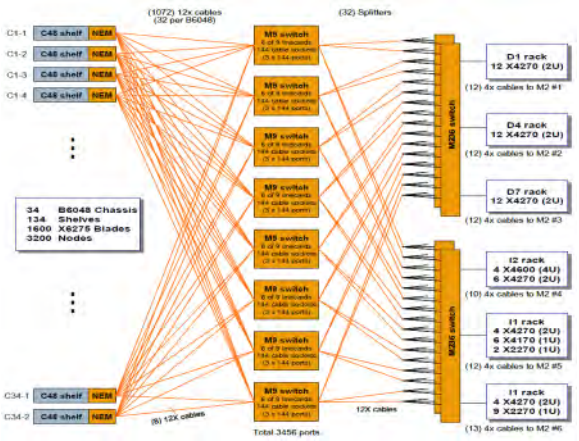
1. 서론

KISTI 슈퍼컴퓨터 4호기 Tachyon2는 SUN Blade 6275 시스템을 기반으로 구성된 초병렬 컴퓨팅 시스템으로 총 3,200개의 계산노드와 각각 117TB 가용 용량의 홈 디렉터리(/home01)와 어플리케이션 디렉터리(/applic) 그리고 가용용량 2.3PB의 스크래치 디렉터리(/scratch2)를 제공하고 있다. 계산노드 간의 메시지 통신 및 파일 I/O를 위해 인터커넥션 네트워크로 인피니밴드(InfiniBand, 이하 IB)를 사용하고 있는데, 슈퍼컴퓨터와 같은 대형 HPC 환경에서 대규모의 사용자 작업을 원활하게 수행하기 위해서는 IB의 안정성이 우선적으로 보장되어야 한다.

본 논문에서는 Tachyon2 시스템에서 발생하는 IB 상태를 파악하고 관리하기 위한 자동화 도구를 개발하였다.

2. 시스템 구성

4호기 Tachyon2 시스템은 3,200개의 계산노드가 고집적도를 가진 34개의 SB 6048 랙에 장착되어 있고 파일시스템은 8대의 SUN X4270 서버와 16대의 J4400 스토리지 연동한 스토리지로 구성되어 있으며, 각각 117TB 가용용량의 홈디렉터리(/home01)와 어플리케이션 디렉터리(/applic)를 제공하고 있다. 2014년도에 추가로 도입된 DDN 스토리지는 Enclosure SS8460(Disk Box)와 10대의 ORACLE X3-2L 서버(MDS 2대, OSS 8대)를 연동하여 구성되어 있고, 가용용량 2.3PB의 스크래치 디렉터리(/scratch2)를 제공하고 있다. 계산 작업은 스크래치 디렉터리의 사용자 작업 공간에서 수행된다. 이는 분산병렬파일시스템인 러스터(Lustre)를 이용하여 구성하였으며, 로그인 노드(Login Node), 데이터 아카이빙 서버(Data Mover), 컴퓨팅 노드, 디버깅 노드에 마운트 되어있다. 노드간 계산 네트워크 및 파일 I/O 통신을 위한 백본 네트워크로 IB를 사용하고 있으며, 4x QDR IB를 사용하여 Fat-tree, Non-blocking IB 네트워크로 구성되어있다. 8대의 Sun Datacenter IB 648 스위치에 모든 계산 노드, 인프라 노드, 파일서버 및 소형 클러스터시스템이 36포트 IB스위치를 통해 채널 당 40Gbps의 대역폭으로 연결되어 있으며 총 11,336개의 링크와 611개의 인피니밴드 칩으로 구성되어 있다.



(그림 1) Tachyon2 IB 구성도

3. 인터커넥션 네트워크 진단 시스템 구축

슈퍼컴퓨터와 같은 HPC 환경에서 대규모의 사용자 작업을 원활하게 수행하기 위해서는 IB의 안정성이 우선적으로 보장되어야 한다. 이를 위해서 인피니밴드 진단 도구에 의해 보고되는 다양한 포트 카운터 값을 참조하여 시스템 상태를 결정하게 된다[2]. Tachyon2 시스템에서 참조하는 인피니밴드 포트 카운터 주요 내용과 임계치는 아래와 같다. 이 임계치를 넘는 경우 해당 장비에 대한 테스트를 통해 수리 또는 교체를 진행하여야 한다.

<표 1> 포트카운터(perfquery) 주요내용 및 임계치

perfquery 타입	설 명	임계치
Symbol Error	전송된 전체 비트 중에 오류 비트의 총 수	120/hr
LinkRecover	포트 트레이닝(점검) 상태에서 정상상태로 복구되는 총 횟수	10/hr
LinkDowned	링크 오류에서 포트 트레이닝(점검) 상태로 복구되는 총 횟수	10/hr
Receive Errors	포트에서 수신된 에러를 포함한 패킷의 총 수	120/hr
XmtDiscards	포트가 다운 또는 혼잡으로 인해 드롭 된 패킷의 총 수	1,000/hr

- BER(Bit Error Rate) 계산
: 시간당 전송률 32Gb(QDR) x 3,600sec =115,200Gb/h
: QDR Threshod(bit): 10e-12 = 1/1,000(Gb)
: 115,200 / 1,000 =115.2 BER (대략 120/hr)

<표 2> perfquery를 통한 포트카운터 수집

perfquery를 통한 포트카운터
[root@ibsm01 27Feb]# perfquery 20 1 # Port counters: Lid 20 port 1 PortSelect:.....1 CounterSelect:.....0x1400 SymbolErrors:.....0 LinkRecovers:.....0 LinkDowned:.....0 RcvErrors:.....0 RcvRemotePhysErrors:.....0 RcvSwRelayErrors:.....0 XmtDiscards:.....0 XmtConstraintErrors:.....0 RcvConstraintErrors:.....0 LinkIntegrityErrors:.....0 XmtData:.....1875106 RcvData:.....3704340 XmtPkts:.....43945 RcvPkts:.....97861
시간별 포트카운터 수집(GetCounters)

// GUID, #Port,Symbol Error, LinkRecover,
// LinkDowned, Receive Errors,XmtDiscards
0x0021283a85b310d2,32,0,0,0,0,0,0,0,0,0,0,0,0,4294967295,4294967295,869873337,595414478,1194960,1347224,2400,2840,5932928123926,6552020979529,0,0
0x0021283a85b310d2,7,0,0,0,0,0,0,0,0,0,0,0,0,147665565,657546391,329280,5433236,0,0,0,0,4998089963238,5462244660488,0,0
0x0021283a85b310d2,26,0,0,0,0,0,0,0,0,0,0,0,0,60552,4294967295,841,185963705,0,404610,0,1198,4604062242695,5600670414659,472025,6384
.....

OFED(OpenFabrics Enterprise Distribution)는 RDMA (Remote Direct Memory Access)를 지원하기 위하여 개발된 오픈소스 프로젝트로 서브넷 내에 구성된 모든 디바이스 포트들의 성능 모니터링 인터페이스를 제공한다[3][4]. <표 1>과 같이 OFED에서 제공하는 포트 카운터 질의(perfquery: query InfiniBand port counters) 정보를 활용하여 아래와 같이 IB 관리 도구를 개발하였다. 이를 통해 Tachyon2 인터커넥션 네트워크 정보를 주기적으로 수집하고 장애 포트를 분별하여 관리함으로써 시스템 관리의 용이성을 가져올 수 있다. 뿐만 아니라 non-blocking 네트워크 특성상 단일 포인트에서의 문제가 전체 시스템으로 파급되어 시스템 레벨에서 부하를 유발하는 증상 및 장애 복구 등에서도 신속하게 대처할 수 있어 시스템의 가용성을 높일 수 있다[5].

① 인피니밴드의 포트카운터(PortCounters) 수집
OFED에서 제공하는 perfquery 명령어를 통해 IB 포트의 성능 및 에러 카운트를 확인할 수 있다.
1시간마다 서브넷에 있는 모든 포트들의 perfquery 정보를 crontab을 이용하여 <표 2> 같이 수집하고 기록한다.
② 인피니밴드의 비활성화 및 에러 포트 검출
<표 2>와 같이 1시간 단위로 수집된 포트카운터 정보를 이용하여 비활성화 및 <표 1>에서 임계치를 초과하는 에러 포트를 자동으로 검출하고 기록한다.
③ 장애 포트 비활성화
IB 포트 장애 히스토리를 검색하여 지속적으로 Symbol Error 값이 임계치를 넘을 경우 해당 포트를 자동으로 비활성화 시킨다, 추후 정비 기간 내에 장비 Swap 테스트 등 다양한 테스트를 통해 문제를 유발한 장비(IB 케이블, IB HCA, 스위치)를 확인하고 교체한다.
④ 비활성화된 링크 출력
위의 포트카운터를 통해 스위치간의 비활성화된 링크의 목록을 출력하는 스크립트로 <표 3>과 같이 해당 날짜와 스위치, 포트 정보를 보여주게 된다.

<표 3> 비활성화 링크 리스트

2016.dis/23MayCmds.warn C31SH3-I4B, 4 <----> S3-1-LC3-I4B, 22 ::: S3-1-LC3-3A_C31SH3-B3 --- 2015.dis/22SepCmds S2-2-LC5-I4C, 34 <----> mid01, 23 ::: S2-2-LC5-I4C_mid01-23 --- 2015.dis/22SepCmds S3-2-LC5-I4A, 33 <----> mid02, 24 ::: S3-2-LC5-I4A-mid02-24 --- 2016.dis/23MayCmds.warn C31SH3-I4B, 4 <----> S3-1-LC3-I4B, 22 ::: S3-1-LC3-3A_C31SH3-B3 --- 2017.dis/08MarCmds C31SH3-I4B, 7 <----> S3-2-LC3-I4B, 22 ::: S3-2-LC3-3A_C31SH3-B6 **replaced(2011-05-20) **replaced(2015-05-13)

4. 향후 계획

Tachyon2 시스템은 IB 스위치 및 계산 노드의 IB HCA 등 연결된 모든 IB 포트를 모니터링해야 하며, 이 포트의 수가 25,000개를 넘고, 포트당 최소 16개 metric 정보를 수집해야 하기에, 모니터링 부하를 고려할 때, 현실적으로 모든 포트를 실시간 환경으로 모니터링 하는 것은 불가능하다. 하지만 경우에 따라서는 현재 수행중인 사용자 작업의 통신 부하 분석 및 통신 속도 저하 원인 분석 등을 위해 실시간으로 일부 포트를 점검할 필요가 있다. 하지만 32개 계산노드를 활용하는 병렬 사용자 작업의 경우라도, 각 노드간의 1:1 연결을 모두 고려하고, 각 연결당 최대 13개의 IB 포트를 지난다는 점을 감안할 때, 수집해야 할 포트의 수는 최대 32x31x13 = 12,896개에 이를 정도로 작지 않아서 실시간으로 모니터링하여 분석하는 것은 불가능에 가깝다. 따라서 전체 시스템에 대한 모니터링 주기를 최소화하는 방안이 검토 및 보완되어야 하며, 이를 위해 최근 몇 년간 개발된 상용 및 오픈소스 기반의 IB 모니터링 도구를 참고할 필요가 있다. 아울러 IB의 Static Routing Table의 기록을 통해, 현재 뿐만 아니라 과거에 수행된 사용자 작업에 대해서도 분석할 수 있는 기반을 마련할 필요가 있다.

4. 결론

Tachyon2 시스템은 다양한 과학 및 공학 분야의 대규모 계산 및 제품개발을 위한 가상실험 지원 등 공공 목적의 슈퍼컴퓨터로 산학연을 대상으로 무중단 서비스를 제공하고 있다. 인터커넥션 네트워크의 안정성은 사용자 작업 수행에 있어 필수적인 요소로 주기적으로 상태를 모니터링하고 장애내역 또한 신속하게 파악할 수 있어야 한다.

본 논문에서는 Tachyon2 시스템의 모든 IB 포트의 상태(성능 및 에러 카운트)를 시간별로 수집, 분석, 처리하는 자동화 도구를 개발하였다. 이로써 인터커넥션의 상태를 주기적으로 모니터링 할 수 있고, 장애내역 또한 신속하게 파악할 수 있다.

참고문헌

[1] Sun Microsystems. Sun Datacenter Infiniband Switch 648 Architecture And Deployment, 2009. http://www.oracle.com/us/sun/.

[2] InfiniBand Trade Association. Infiniband architecture specification, 1.2.1 edition, November 2007. http://www.infinibandta.org/.

[3] Alliance, OpenFabrics. "OpenFabrics enterprise distribution (OFED)." (2010).

[4] Subramanian, Viswanath, Michael R. Krause, and Ramesh VelurEunni. "Remote direct memory access (RDMA) completion." U.S. Patent No. 8,244,825. 14 Aug. 2012.

[5] Dandapanthula, N., et al. "INAM-a scalable infiniband network analysis and monitoring tool." European Conference on Parallel Processing. Springer Berlin Heidelberg, 2011.