# 슈퍼컴퓨터 단일 컴퓨팅 노드 성능 측정을 위한 벤치마크 기법

권민우*, 윤준원*, 홍태영*
*한국과학기술정보연구원 슈퍼컴퓨팅센터
e-mail:mwkwon81@kisti.re.kr

# Benchmarking techniques to evaluate single computing node of HPC

Min-Woo Kwon*, JunWeon Yoon*, TaeYoung Hong*
*Dept. of Supercomputing Center, KISTI

요 약

한국과학기술정보연구원에서 운영 중인 슈퍼컴퓨터 4호기인 Tachyon 2차 시스템은 이론최고성능 300TFlops인 SUN Blade 6275 시스템을 기반으로 구성되어있다. 로그인 노드 4대와 컴퓨팅 노드 3200대로 구성되어 있으며 컴퓨팅 노드 중에 24대는 디버깅 노드로 사용되고 있다. 3200대의 컴퓨팅 노드가 동일한 하드웨어로 구성이 되어 있으므로 Tachyon 2차 시스템의 전체 계산 성능을 결정하는 가장 중요한 요소가 단일 컴퓨팅 노드의 성능이 되겠다. 본 논문에서는 다양한 벤치마크 기법을 통해 단일 노드의 성능을 측정하여 분석하였다.

## 1. 서론

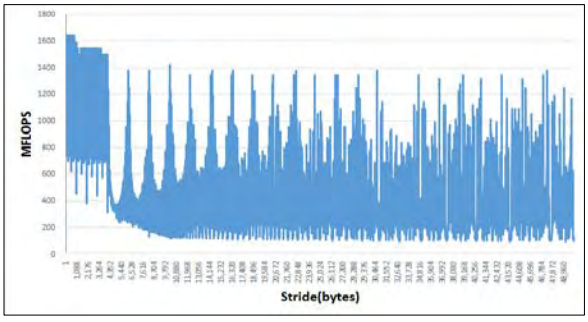
한국과학기술정보연구원의 슈퍼컴퓨터 4호기인 Tachyon 2차 시스템은 컴퓨팅노드 3200대로 구성되어 있으며 모든 컴퓨팅 노드가 동일한 하드웨어로 구성되어 있다[1]. 슈퍼컴퓨터의 성능을 측정할 때 HPL(High Performance Linpack)이 가장 많이 쓰이지만[2], 본 논문에서는 좀 더 다양한 벤치마크 기법을 통해 단일 컴퓨팅 노드의 성능을 다양하게 측정하여 분석하였다. 본 논문에서는 총 4개의 벤치마크 툴을 이용해 단일 컴퓨팅 노드의 성능을 측정하였다.

## 2. Stream 벤치마크

Stream은 프로세서와 메모리 간의 대역폭을 측정하는 벤치마크 도구이다. 일반적으로 서버의 CPU는 메모리 시스템에 비해서 속도가 빠르다. 이로 인해 메모리 대역폭에 따라 서버의 성능이 제한되는 문제가 발생할 수 있다. Stream 벤치마크는 시스템의 캐시 메모리 용량보다 훨씬 큰 데이터 량을 가지고 시스템을 테스트함으로 매우 큰 벡터타입의 데이터를 사용하는 어플리케이션에 대한 시스템의 성능을 예측해볼 수 있다[3]. 본 논문에서는 컴퓨팅 노드가 동일한 하드웨어로 구성되어 있으므로 한 개의 컴퓨팅노드를 택하여 CPU와 DDR 메인메모리 간의 성능 대역폭을 측정하였고 MPI나 OpenMP를 통해 단일 노드의 최대 대역폭을 측정하였다. 실험결과, 8 core를 사용했을 때, 최대 35,690MB/s의 전송속도를 가지는 것으로 측정되었다.

## 3. Stride 벤치마크

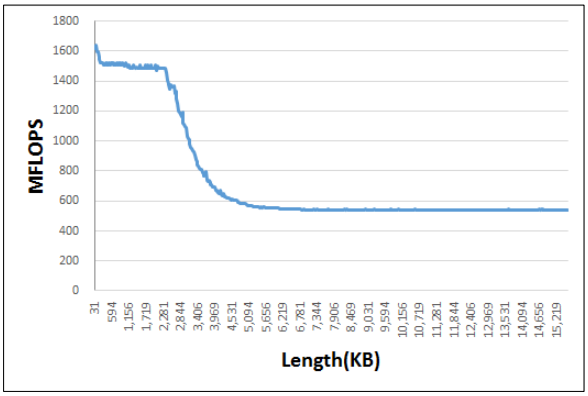
Stride는 메모리나 캐시의 성능 측정 및 스트레스 테스트를 수행하는 벤치마크 도구이다. 벤치마크는 스칼라와 벡터 연산의 루프 형식으로 구현되어 있고 메모리 접근 패턴과 사용하는 벡터의 길이에 따른 전송된 MFLOP rate를 측정하는 방식으로 테스트를 진행한다. 따라서 다양한 메모리에 대한 접근 패턴을 이용해 다양한 메모리 사용조건을 가지는 어플리케이션에 대한 스트레스 테스트가 가능하다[4]. 본 논문에서는 Stride 테스트 중에 대표적인 STRID3, CACHE에 대한 성능 테스트를 한 개의 컴퓨팅노드에서 진행하였다.




(그림 1) Stride(strid3) 벤치마크 결과

그림1는 Stride 간격을 1Byte 에서 262KByte까지 변화시키며 동일한 연산량에 대한 계산속도(MFLOPS)를 측정한 결과이다. 결과에서 확인할 수 있는 것처럼 동일한 연산량에 대한 Stride 간격의 차이에 따라 최대 1,638

MFLOPS에서 최소 67 MFLOPS까지 처리량의 차이를 보이는 것을 확인할 수 있었다.




(그림 2) Stride(cache) 벤치마크 결과

그림3은 동일한 데이터량에 대해 한 루프에서 처리되는 데이터 량을 64Byte에서 16MByte까지 변화시키며 계산속도(MFLOPS)를 측정한 결과이다. 결과에서 확인할 수 있는 것처럼 한 루프에서 처리되는 데이터량이 Tachyon 2차 시스템의 L3 캐쉬 사이즈인 8Mbyte를 넘어갈 때 더이상 캐시의 기능이 발휘되지 못하고 최저 계산성능인 540 MFLOPS로 유지되는 것을 확인할 수 있었다.

## 4. PSNAP 벤치마크

PSNAP은 OS의 간섭과 노이즈를 측정하여 OS Jitter를 측정하는 벤치마크 도구이다. PSNAP 벤치마크 테스트는 주어진 시간 동안 스핀 루프를 돌려서 이론적으로 걸리는 시간과 실제 측정되는 시간의 오차 비율을 통해 OS의 간섭과 노이즈를 측정한다. 스핀 루프 1회 처리 시간은 이론적으로 1ms이고 이를 반복 수행했을 때 이론 시간과 실제 시간의 오차를 구할 수 있다. PSNAP 벤치마크 테스트는 CPU마다 수행되어야 한다[5]. 본 논문에서는 한 개의 컴퓨팅노드를 택하여 각각의 CPU에 대한 스핀 루프의 실제 시간을 측정하여 이론 시간(1ms*반복수행회수)과 비교함으로 Tachyon 2차 시스템의 컴퓨팅 노드에 대한 성능 테스트를 진행하였다. 한 개의 컴퓨팅 노드의 단일 CPU에서 스핀루프를 100,000번 수행했을 때, 104,374 ms가 걸렸고, 이를 통해 4.374%의 OS Jitter가 발생함을 확인할 수 있었다.

## 5. CLOMP 벤치마크

CLOMP는 OpenMP의 오버헤드와 그 밖의 성능을 측정하는 도구이다. CLOMP 벤치마크 테스트는 Part와 Zone으로 구성된 메쉬타입의 데이터 구조를 이용하여 다양한 인자값의 변화를 통해 시스템의 구성에 따라 OpenMP의 성능에 어떠한 영향을 주는지를 다양하게 측정할 수 있다. 본 논문에서는 CLOMP 매뉴얼에서 제공하는 6가지의 기본적인 테스트를 진행하여 컴퓨팅 노드에 대한 OpenMP 성능 테스트를 진행하였다. 6가지의 기본적인 테스트는 Target input 테스트, Target/NUMA friendly 테스트, Cache friendly 테스트, Cache/OpenMP friendly 테스트, Mem-bound input 테스트, Mem-bound/NUMA friendly 테스트가 있다[4]. 본 논문에서는 OpenMP 성능의 가장 실질적인 평가가 될 수 있는 Manual Case(계산결과를 보장하는 실질적인 테스트)의 Speedup(직렬 프로그램 동작 시간/병렬 프로그램 동작 시간)값으로 성능을 측정하였다. Speedup 값이 전체 Core 개수에 근접할수록 좋은 성능이라 평가할 수 있다[4].

<표 1> CLOMP 벤치마크 결과

| Test Cases | Speedup |
|---|---|
| Target input | 5.01 |
| Target/NUMA friendly | 4.37 |
| Cache friendly | 4.57 |
| Cache/OpenMP friendly | 7.22 |
| Mem-bound input | 5.33 |
| Mem-bound/NUMA friendly | 7.74 |

표1은 CLOMP의 기본적인 6가지 테스트에 대한 각각의 Manual Case의 Speedup 값을 보여준다. 테스트 결과 Tachyon 2차 시스템의 컴퓨팅 노드에서는 Mem-bound/NUMA friendly한 테스트에서 core 개수인 8개에 가장 근접한 성능을 보임을 알 수 있었다.

## 5. 실험 결과 분석 및 결론

슈퍼컴 4호기인 Tachyon 2차 시스템은 CPU와 메모리간에 최대 35,690MB/s의 전송속도를 가지며 Stride 간격의 차이에 따라 1,638~67 MFLOPS의 처리량의 차이를 보였다. OS Jitter는 4.374%가 발생하며 Mem-bound/NUMA friendly인 케이스에서 가장 좋은 OpenMP 성능을 보여줬다. 본 논문의 실험결과를 향후 슈퍼컴퓨터 5호기의 도입에 기초 데이터로 활용하고 좀 더 다양한 시스템과의 비교, 분석을 수행할 예정이다.

## 참고문헌

[1] National Institute of Supercomputing and Networking, KISTI, http://www.nisn.re.kr

[2] 홍태영, 홍정우, 김성호 “리눅스 클러스터 시스템 계산노드용 단일서버 벤치마크” 한국컴퓨터종합학술대회, 32(1), 52-54, 2005

[3] McCalpin, John D. “Sustainable memory bandwidth in current high performance computers” Silicon Graphics Inc, 1995).

[4] ASC Sequoia Benchmark Codes, Lawrence Livermore National Laboratory (LLNL), https://asc.llnl.gov/sequoia/benchmarks/#clomp

[5] PAL System Noise Activity Program, Los Alamos National Laboratory, http://www.c3.lanl.gov/pal/software/psnap