슈퍼컴퓨터 작업 로그 기반 실패 작업 특성 연구[1]

박주원
한국과학기술정보연구원
e-mail: juwon.park@kisti.re.kr

Analysis of failed job based on scheduler job logs

Ju-Won Park
Korea Institute of Science and Technology Information

요 약

최근 기초 과학 분야뿐만 아니라 빅데이터 분석, 인공 지능과 같은 컴퓨터 과학 분야에서도 대용량의 컴퓨팅 자원을 많이 활용함에 따라 슈퍼컴퓨터와 같은 고성능 컴퓨팅 자원에 대한 요구가 더욱 증가하고 있다. 이러한 대규모 컴퓨팅 자원을 안정적으로 운영하기 위해서는 실패 작업의 특성에 대한 상세한 분석이 필수적이다. 본 논문에서는 한국과학기술정보연구원에서 운영하고 있는 슈퍼컴퓨터(Tachyon)에서 1 년 동안 수집된 작업 데이터를 기반으로 고성컴퓨팅 시스템을 활용하는 작업의 특성을 파악하기 위해 다음 3 가지의 분석 결과를 제시한다. 첫째는 실패한 작업의 비율, 평균 사용한 processor 수, 전체 작업 시간 중 실패 작업이 차지한 비율과 같이 간단한 통계적 분석 결과를 제시한다. 둘째는 실패한 작업의 inter-arrival time 분포 모형을 제시한다. 마지막으로 시간에 따른 실패 작업 확률을 분석하기 위해 inter-arrival time 값을 이용하여 hazard rate 결과를 제시한다.

1. 서론

전통적으로 고에너지 물리, 해양, 천문등 기초 과학 분야의 대규모 워크로드 실행하기 위해 대용량 컴퓨팅 자원이 많이 활용되고 있다. 최근 들어 이와 같은 기초과학 뿐만아니라 빅데이터 분석, 인공 지능과 같은 컴퓨터 과학 분야에서도 대용량의 컴퓨팅 자원을 많이 활용함에 따라 슈퍼컴퓨터와 같은 고성능 컴퓨팅 자원에 대한 요구가 더욱 증가하고 있다. 이러한 요구로 인하여 지난 10 년간 고성능 컴퓨팅의 성능은 기하급수적으로 증가하였으며 2023 년에는 Exa-scale 의 컴퓨터가 등장할 것으로 예상된다[1]. Exa-scale 컴퓨팅 시스템은 수십만 노드 규모로 커질 것으로 예상되며 시스템을 안정적으로 운영하기 위해서는 체크포인팅, failure-aware 작업 스케줄링, 시스템 시뮬레이션과 같은 운영 관리 기술이 점점 중요해지고 있다. 안정적이고 robust 한 시스템을 유지 및 관리하기 위해서는 실패한 작업의 특성에 대한 상세한 분석이 필수적이다.

본 논문에서는 병렬 프로그램 작업의 특성을 파악하고 실패 작업 분석을 위해 다음 3 가지 방법을 사용한다. 첫째는 실패한 작업의 비율, 평균 사용한 processor 수, 전체 작업 시간 중 실패 작업이 차지한 비율과 같이 간단한 통계적 분석 결과를 제시한다. 둘째는 실패한 작업의 inter-arrival time 분포 모형을 제시한다. 이를 위해 Normal, Exponential, Gamma, Log-normal, Weibull, Pareto 총 6 개의 이론적 분포도와 Kolmogorov-Smirnov(K-S) 검정[2] 결과를 통해 가장 적합한 분포 모형을 찾는다. 마지막으로 시간에 따른 실패 작업 확률을 분석하기 위해 inter-arrival time 값을 이용하여 hazard rate 결과를 제시한다.

본 논문의 구성은 다음과 같다. 먼저 제 2 절에서는 분석 데이터 및 관련 연구에 대해 살펴본다. 제 3 절에서는 슈퍼컴퓨터 작업 로그 기반 실패 작업의 특성을 제시하고 제 4 절에서 결론을 맺는다.

2. Background

2.1 분석 데이터

한국과학기술정보연구원에서는 국내 기초 과학 연구 지원을 위해 슈퍼컴퓨팅 서비스를 제공하고 있으며 1988 년 슈퍼컴퓨터 Cray-25 를 시작으로 현재 4 호기를 통해 고성능 컴퓨팅 자원을 제공하고 있다. 타키온으로 이름 붙여진 4 호기는 SUN 의 Blade 6275 시스템을 기반으로 구성되었으며 Rpeak 는 300TFlops 의 연산 성능을 가지고 있다. 총 3,200 노드, 25,600 CPU 코어로 구성되어 있으며 작업 관리를 위한 스케줄러는 Sun Grid Engine[3]을 사용하고 있다. 본 논문에서 사용한 작업 로그 데이터는 표 1 에서 제시한 특성(feature) 값을 측정한 것으로 2015 년 1 월에서 2015 년도 12 월까지 1 년동안 총 300,995 작업, 6,259 억 CPU 시간에 이른다.

[1] 본 논문은 [13] 연구 내용을 기반으로 작업 로그 데이터 (2015 년 1 월 ~ 2015 년 12 월)를 활용하여 분석된 내용임.

<표 1> 분석 데이터의 작업 특성.

FEATURE	Description
J_DATE	Job submit date
USER ID	User ID
JOB_NUM	Number of job
JOB_NAME	Name of job
QUEUE	Name of the queue
Q_DATE	Job submit date to queue
S_DATE	Job start date
W_TIME	Wait time
E_DATE	Job end time
R_TIME	Job run time
CPUS	Number of used CPU cores
MEM	The integral memory usage
CPU_TIME	The cpu time usage
JOB_STATUS	The status of Job (D or E)
EXIT_CODE	Exit status of the job
FAILED_CODE	Failed code of job
OMP_NUM_THREADS	Number of threads to use

2.2 관련 연구

[4-6]와 같이 몇몇의 연구에서 실패 작업 특성 및 성능 분석에 대한 연구가 이루어졌으나 몇가지 한계점이 있다. 첫째는 실패 작업과 분석되는 특징에 대한 연관 관계를 제시하지 못했다. [7-8]등 많은 논문에서 실패 작업을 실행한 노드의 시간적/공간적 분석 결과를 제시하고 있으나 정작 실패한 작업이 실행 노드 및 inter-arrival time 이라는 특성과 어떠한 연관성을 가지고 있는지 제시하지 못했다. 둘째, 실패 작업의 특징 분석을 제한된 필드 정보를 통해 이루어졌다. 사실 보안 문제로 인하여 슈퍼컴퓨터와 같은 대용량 컴퓨터를 운영하는 기관에서는 작업 로그 데이터를 공개하지 않고 있다. [9]의 노력으로 인하여 많은 슈퍼컴퓨터의 작업 로그들이 표준화된 형식 (Standard workload format)으로 공개되어 스케줄러 성능 분석, 작업 시간 예측등 많은 연구가 이루어지고 있으나 데이터를 가공하는 단계에서 많은 정보가 유실되어 실패 작업의 상세한 특성 분석에는 한계가 있다.

3. 슈퍼컴퓨터 로그 기반 실패 작업의 특성

본 절에서는 먼저 고성능 시스템의 작업의 특성을 파악하기 위해 다음 3 가지 방법을 사용한다. 첫째는 실패한 작업의 비율, 평균 사용한 processor 수, 전체 작업 시간 중 실패 작업이 차지한 비율과 같이 간단한 통계적 분석 결과를 제시한다. 통계적 분석 결과는 [9]에서 제공하는 parallel workload log 중 유사한 아키텍쳐를 가지는 LLNL-ATLAS, LLNL-Thunder, CTC-SP2, SDSC Blue 시스템의 작업 로그 분석 결과와 함께 제공한다. 둘째는 실패한 작업의 inter-arrival time 분포 모형을 제시한다. 이를 위해 Normal, Poisson, Gamma, Log-normal, Weibull, Pareto 총 6 개의 이론적 분포도와 Kolmogorov-Smirnov(K-S) 검정 결과를 통해 가장 적합한 분포 모형을 찾는다. 마지막으로 시간에 따른 실패 작업 확률을 분석하기 위해 inter-arrival time 값을 이용하여 hazard rate 결과를 제시한다.

<표 2> 실행 작업의 통계량.

DATA SET	SUCCESS JOB			FAILED JOB		
	#job ratio (%)	avg. # of processor	Runtime (%)	#job ratio (%)	avg. # of processor	Runtime (%)
Tachyon	91.4	71.7	70.1	8.6	162	29.9
LLNL-ATLAS	67.8	373	43.8	32.2	456	56.2
LLNL-Thunder	92.2	32.9	92.3	7.8	91.9	7.6
CTC-SP2	78.4	9.8	68.9	21.6	15.3	31.1
SDSC BLUE	97.5	17.9	91.1	2.5	34.6	8.9

3.1 실행 작업의 통계 분석

표 2 는 4 개의 표준화된 형식으로 수집된 슈퍼컴퓨터 운영 센터의 작업과 한국과학기술정보연구원에서 운영 중인 Tachyon 작업의 전반적인 통계량이다. Number of job 과 Runtime 필드는 전체 시간 중에서 성공 또는 실패한 작업이 차지한 비율(%)을 보여주고 Average number of processor 는 각 작업이 사용한 프로세서의 수에 대한 평균 값이다. 전체 작업 중 실패 작업의 비율은 2.5 % (SDSC BLUE) ~ 32.2% (LLNL-ATLAS) 로 매우 다양하며 실패 작업의 실행 시간도 7.6% (LLNL-Thunder) ~ 56.2% (LLNL-ATLAS) 로 매우 다양하다. 그러나 본 테이블에서 다음 2 가지의 공통점을 찾을 수 있다. 첫째는 'observation 1: 실패 작업의 사용 processor 의 수가 크다' 는 점이다. 이는 사용되는 프로세서의 수가 많을 수록 작업이 실패하는 경우가 많다는 것을 의미한다. 둘째는 'Observation 2: LLNL-Thunder 를 제외한 나머지 데이터 셋들에서 실패한 작업의 비율 대비 runtime 비율이 크다' 는 점이다. 특히 tachyon 의 경우 실패한 작업은 전체 작업의 8.6%에 지나지 않지만 runtime 은 전체 실행 시간의 27.7%를 차지한다. 이는 실패 작업의 실행 시간이 성공한 작업의 실행 시간 보다 평균적으로 크다는 것을 의미하는 것으로 [10]에서 가정한 사실과 배치되고 [6]에서 분석한 결과와 일치한다. 실제로 실패한 작업의 평균 CPU time 은 49,148 초로 성공한 작업의 평균 CPU time 인 10,874 초 보다 4.5 배 높았다.

3.2 Inter-arrival time 분포 모형

Tachyon 에서 수집된 특성 중에서 작업이 제출된 시점인 Q_DATE 값을 이용하여 inter-arrival time 을 계산하고 Normal, Exponential, Gamma, Log-normal, Weibull, Pareto 총 6 개의 이론적 분포 모형과의 적합도를 비교한다. 이를 위해 본 논문에서는 두 분포도의 CDFs 사이의 최대 거리 값 (D-value)를 통해 적합도를 판단하는 Kolmogorov-Smirnov (K-S) 검정 방법을 사용한다. K-S 검정을 통해 얻어진 D-value 는 값이 작을 수록 적합도가 우수하다고 판단할 수 있다. 표 3 은 6 개의 이론적 분포도와 실패 작업의 inter-arrival time 의 KS 검정 결과 값을 보여준다. 이를 통해 Obervation 3: 타키온의 실패 작업 inter-arrival time 의 분포는 log normal 분포와 가장 적합하다'는 것을 확인할 수 있다.

<표 3> K-S 검정 결과.

Fit of distribution	Max. distance
Normal	0.34955
Exponential	0.36415
Gamma	0.2859
Log-normal	0.057639
Weibull	0.060154
Pareto	0.10427

3.3 Hazard rate

어떤 이벤트가 빈번히 발생할 경우 최근 발생한 이벤트로부터 시간이 경과함에 따라 이벤트의 발생 확률을 분석하기 위해 많은 논문에서 hazard rate 를 많이 활용한다[10-11].

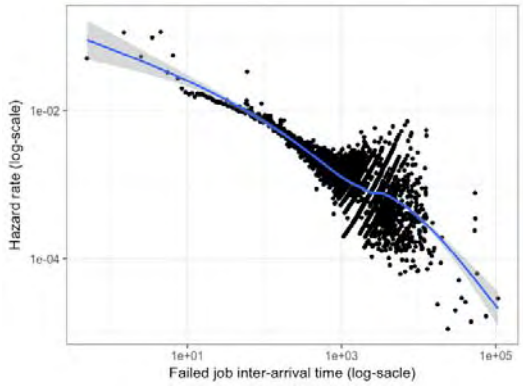

그림 1 실패 작업의 Hazard rate.

그림 1 은 실패 작업의 inter-arrival time 을 이용한 hazard rate 를 보여준다. 그림에서 확인할 수 있듯이 'observation 4: 최근 발생한 이벤트로부터 시간이 증가함에 따라 failed job 발생 확률은 감소한다' 는 것을 확인할 수 있다. 이러한 현상이 발생하는 이유 중 하나는 batch 형태로 실행되는 HPC 작업의 특성으로 기인한다. HPC 작업은 대부분 interative 가 없는 batch 형태로 스케줄러에 제출되어 실행되기 때문에 짧은 시간 안에 많은 작업이 제출된다. Tachyon 의 경우 전체 작업의 49% 이상이 10 초 이내의 inter-arrival time 값을 갖는 것으로 나타났다. 또한 HPC 작업은 동일한 로직에 파라미터만 변경하여 반복 실행함으로써 최적값을 찾는 parameter sweep 형태가 빈번하다. 이러한 작업 형태에서는 실패 작업이 각각 독립적으로 발생하기 보다는 앞 작업과 연관성을 가지는 특징이 있다.

4. 결론

본 논문에서는 2015 년 1 월 ~ 2015 년 12 월까지 1 년동안 Tachyon 을 통해 실행된 작업 300,995 개의 작업 데이터를 기반으로 실패한 작업의 비율, 평균 사용한 processor 수, 전체 작업 시간 중 실패 작업이 차지한 비율과 같이 간단한 통계적 분석 결과와 함께 K-S 검정 방법을 통해 inter-arrival time 의 분포 모형을 제시하였다. 또한 시간이 경과함에 따라 실패할 확률을 분석하기 위해 Hazard rate 결과를 제시하였다. 본 논문을 통해 다음 4 가지의 사실을 확인할 수 있었다.

1. 실패 작업의 사용 processor 의 수가 크다
2. LLNL-Thunder 를 제외한 나머지 데이터 셋들에서 실패한 작업의 비율 대비 runtime 비율이 크다
3. 타키온의 실패 작업 inter-arrival time 의 분포는 log normal 분포와 가장 적합하다'
4. 최근 발생한 이벤트로부터 시간이 증가함에 따라 failed job 발생 확률은 감소한다

참고문헌

[1] U.S Department of Energy Office of Science and National Nuclear Security Administration, "Preliminary conceptual design for an exascale computing initiative," U.S. Department of Energy Office of Science and National Nuclear Security, Tech. Rep., Nov. 2014.

[2] A. Justel, D. Pen˜a, and R. Zamar, "A multivariate kolmogorov-smirnov test of goodness of fit," Statistics & Probability Letters, vol. 35, no. 3, pp. 251–259, 1997.

[3] G. Borges, M. David, J. Gomes, C. Fernandez, J. Lopez Cacheiro, P. Rey Mayo, A. Simon Garcia, D. Kant, and K. Sephton, "Sun Grid Engine, a new scheduler for EGEE middleware," in IBERGRID–Iberian Grid Infrastructure Conference, 2007.

[4] H. Li, D. Groep, L. Wolters, and J. Templon, "Job failure anal- ysis and its implications in a large-scale production grid," in IEEE International Conference on E-Science and Grid Computing (E- Science'06), 2006, pp. 27–27.

[5] Z. Zheng, L. Yu, W. Tang, Z. Lan, R. Gupta, N. Desai, S. Coghlan, and D. Buettner, "Co-analysis of RAS log and job log on Blue Gene/P," in IEEE International Parallel & Distributed Processing Symposium (IPDPS'11), 2011, pp. 840–851.

[6] Y. Yuan, Y. Wu, Q. Wang, G. Yang, and W. Zheng, "Job failures in high performance computing systems: A large-scale empirical study," Computers & Mathematics with Applications, vol. 63, no. 2, pp. 365–377, 2012.

[7] S. Gupta, D. Tiwari, C. Jantzi, J. Rogers, and D. Maxwell, "Under- standing and exploiting spatial properties of system failures on extreme-scale HPC systems," in IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'15), 2015, pp. 37–44.

[8] S. Di, R. Gupta, M. Snir, E. Pershey, and F. Cappello, "Logaider: A tool for mining potential correlations of HPC log events," in IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'17), 2017, pp. 442–451.

[9] FeitelsonDG, TsafrirD, KrakovD, "Experience with the Parallel Workloads Archive," Technical report, 2012.

[10] W. Cirne and F. Berman, "A comprehensive model of the supercomputer workload," in Proceedings of IEEE International Workshop on Workload Characterization, 2001, pp. 140–148.

[11] Cox, David Roxbee, "Analysis of survival data," Routledge, 2018.

[12] Collett, David, "Modelling survival data in medical research," Chapman and Hall/CRC, 2015.

[13] Park, Ju-Won, and Eunhye Kim. "Exploiting the behavior of the failed job in high performance computing system." 2018 18th International Conference on Computational Science and Applications (ICCSA). IEEE, 2018.