http://dx.doi.org/10.14383/SIME.2015.23.2.33

횡단면분석과 추세분석을 이용한 슈퍼컴퓨팅 성능수요 예측*

박만희**

<목 차>

국문초록 : 국가차원의 슈퍼컴퓨팅 성능수요 예측은 슈퍼컴퓨터를 활용하는 계산과학분야의 연구자나 연구개발 인프라를 구축・운영하고 있는 전문기관, 과학기술 인프라구축을 주도할 정부기관에 있어서 매우 중요한 정보이다.

본 연구는 그동안 진행되었던 슈퍼컴퓨터 성능관련 예측활동 분석을 통해 과학기술 역량에 영향을 미치는 요인들을 도출하고 이를 슈퍼컴퓨터 기술진보 추세에 적용한 복합 예측모형을 제안하였다. 횡단면분석에서는 슈퍼컴퓨팅 성능에 영향을 미칠 것으로 판단되는 GDP, GERD, 연구원수, SCI논문수를 고려한 다중회귀분석을 수행하였다. 그리고 횡단면분석 결과에 Top500 자료의 성능(Rmax)값을 이용한 시계열분석을 통해 도출된 기간별 기술진보율을 곱하여 슈퍼컴퓨터의 성능을 예측하였다.

제안된 예측모형을 바탕으로 세계 슈퍼컴퓨터 500위의 시계열자료를 이용하여 한국이 2016년에 보유해야 할 슈퍼컴퓨터 성능규모를 예측하였다. 횡단면분석과 기술진보율을 적용하여 2016년 한국의 슈퍼컴퓨팅 성능수요를 예측해본 결과 현재의 추세를 이용할 경우

* 본 논문은 2013년 KISTI 슈퍼컴퓨팅본부 자문내용의 일부를 수정・보완하여 작성되었음.
** 부산가톨릭대학교 경영학과 교수 (mhpark@cup.ac.kr)

15~30PF 정도, 목표 국가수준의 추세를 이용할 때 20~40PF 정도의 컴퓨팅 역량이 필요할 것으로 예측되었다. 이 결과는 단순 회귀분석을 적용한 결과인 9.6PF와 횡단면분석을 적용한 결과인 2.5PF와 큰 차이를 나타내었다.

주제어 : 슈퍼컴퓨터, 성능수요예측, 고성능컴퓨팅

Supercomputing Performance Demand Forecasting Using Cross-sectional and Time Series Analysis

Manhee Park

Abstract : Supercomputing performance demand forecasting at the national level is an important information to the researchers in fields of the computational science field, the specialized agencies which establish and operate R&D infrastructure, and the government agencies which establish science and technology infrastructure.

This study derived the factors affecting the scientific and technological capability through the analysis of supercomputing performance prediction research, and it proposed a hybrid forecasting model of applying the super-computer technology trends. In the cross-sectional analysis, multiple regression analysis was performed using factors with GDP, GERD, the number of researchers, and the number of SCI papers that could affect the supercomputing performance. In addition, the supercomputing performance was predicted by multiplying in the cross-section analysis with technical progress rate of time period which was calculated by time series analysis using performance(Rmax) of Top500 data.

Korea's performance scale of supercomputing in 2016 was predicted using the proposed forecasting model based on data of the top500 supercomputer and supercomputing performance demand in Korea was predicted using a cross-sectional analysis and technical progress rate. The results of this study showed that the supercomputing performance is expected to require 15~30PF when it uses the current trend, and is expected to require 20~40PF when it uses the trend of the targeting national-level. These two results showed significant differences between the forecasting value(9.6PF) of regression analysis and the forecasting value(2.5PF) of cross-sectional analysis.

Key Words : Supercomputer, Performance demand forecasting, HPC

Ⅰ. 서론

슈퍼컴퓨팅 성능수요 예측은 합리적인 슈퍼컴퓨터 수급계획과 계산과학공학의 연구 수준을 결정하는데 있어 매우 중요한 사안이라고 할 수 있다. 아울러 활용주체, 서비스 주체, 정책주체 등의 목적에 따라 다양한 용도로 미래 슈퍼컴퓨터 성능에 대한 예측이 수행되고 활용되고 있다.

지금까지 슈퍼컴퓨터의 성능수요 예측은 Top500[1]자료에서 제시된 세계 슈퍼컴퓨터 성능측정값(Rmax[2])을 대상으로 성능발전 추세를 추정하고 국가별 동향을 점검하는 등 Top500자료에 의존적인 경향이 있었다. 이는 공신력을 갖은 슈퍼컴퓨터 관련 공식 통계를 집계하기 어려운 탓이기도 하다. 이렇듯 Top500의 자료는 슈퍼컴퓨터 발전추세를 살피는데 있어 매우 중요한 자료로 슈퍼컴퓨터 성능의 시계열적 분석과 한 시점의 현황정보를 제공하는 중요한 자료이다. Rmax도 활용분야에 따라 기여하는 바가 다르기 때문에 일괄적인 대표 변수로 사용하기 어려운 면이 있으나, 한 국가에서 보유한 슈퍼컴퓨터 총 용량이 일정 수준 이상이라고 하면 슈퍼컴퓨터가 일반적으로 사용되는 분야가 대부분 비슷한 비율로 포함한다고 가정하고 이를 대표성이 있다고 판단하여 이용하였다.

본 연구에서는 지금까지 슈퍼컴퓨터 관련 예측 활동을 살펴보고 국가차원에서 슈퍼컴퓨팅 육성 정책수립에 활용할 수 있는 슈퍼컴퓨터성능 예측모델을 개발하여 국가차원에서 향후 구축해야할 슈퍼컴퓨팅 성능규모를 예측하고, 이를 통해 목적에 부합하고 합리적인 정책의사결정이 이루어지도록 지원하고자 하였다.

---

1) 세계 고성능컴퓨터(HPC: High Performance Computer) Top500사이트(www.top500.org)에서 매년 6월과 11월 두 차례 세계에서 가장 빠른 컴퓨터 500위를 발표하고 있다.

2) Dongarra(1986)는 Linpack이라는 기본 방정식을 통해 슈퍼컴퓨터 시스템의 성능을 평가하였는데 Linpack 벤치마크는 연립 1차 방정식의 근을 구하는 프로그램을 통해 부동 소수점연산의 처리 성능을 측정하는 것으로, 측정결과는 FLOPS(Floating Point Operations Per Second)로 표현된다.

Ⅱ. 선행 연구의 검토

1. 슈퍼컴퓨터 성능예측관련 선행연구

일반적으로 예측에 사용되는 분석방법은 정성적인 방법과 정량적인 방법으로 구분되는데 정성적 방법은 전문가들의 의견을 수렴하여 미래 상황을 예측하는 것으로 델파이법이나 시나리오방법 등이 대표적이다. 아울러 정량적 방법은 시계열분석, 인과관계분석, 그리고 성장곡선 모형 등 예측 값을 양적으로 표현할 수 있는 방법이다. 예측기법을 간략히 요약하자면 아래 <표 1>과 같이 정리할 수 있다.

<표 1> 예측기법의 분류

구 분		예측기법
정성적 기법		시장조사법, 전문가 토론, 델파이기법, 시나리오 분석, 역사적 유추 등
정량적 기법	시계열 분석	이동평균법, 지수평활법, ARIMA 분석 등
	인과관계 분석	회귀분석, 계량경제모형, 신경망분석 등
	시스템 모형	투입산출모형(Markov), 시스템다이내믹스 모형 등
	성장곡선 모형	Logistics 곡선, 수정지수 곡선 등

그동안 슈퍼컴퓨터 관련 성능예측은 정성적인 방법과 정량적인 방법이 각각 별도로 적용되어왔다. 특히 Top500의 자료를 활용한 시계열 분석 아니면 국가별 슈퍼컴퓨터 성능에 미치는 요인간의 관계를 규명하는 인과관계 분석 모형이 주로 이용되어 왔다. 최근에 와서 슈퍼컴퓨터 수요자들로부터 향후 필요한 슈퍼컴퓨터의 성능 수요조사를 바탕으로 성능을 예측하는 등 한 국가가 보유해야할 적절한 슈퍼컴퓨팅 성능 규모에 대한 다각적인 연구가 진행되고 있다. 기존에 수행되었던 슈퍼컴퓨팅 성능수요 예측 관련 선행 연구들을 요약하면 <표 2>와 같으며[10], 국내에서는 국가초고속컴퓨팅센터로 지정된 한국과학기술정보연구원을 중심으로 진행되어 왔다.

<표 2> 슈퍼컴퓨터 성능 예측 사례

연도	예측방법	주요변수
1991 [1]	회귀분석	종속변수 : 설치대수 독립변수 : 1인당 GNP, 인구수, 국별수입액, 수출액
2000 [2]	회귀분석	종속변수 : Rmax 독립변수 : GDP, 연구개발투자, 연구원수
2003 [3]	시계열분석	시계열 데이터 : 이용기관수, 로그인수, 작업건수
2005 [4]	회귀분석	종속변수 : $RMAX_i$ = Top500 학술연구용 국가별 성능치 독립변수 : 국내총생산(GDP), 연구개발비, SCI 논문수, 인구 GDP_i = 국가별 국내총생산 RD_i = 국가별 총연구개발비 $Article_i$ = 국가별 SCI 등재 논문수 GDP대비연구개발비i=국가별 연구개발비/국내총생산 POP_i = 국가별 인구수
	시계열분석	$GoRmax_t$ = Top500의 학술연구용 총 성능치 $Year_t$ = 연도 (1993~2004)
2007 [5]	시계열분석	한국 슈퍼컴퓨터 성능
	회귀분석	종속변수 : $RMAX_i$ = 국가별 컴퓨팅성능합 독립변수 : 국내총생산(GDP), 연구개발비, SCI 논문수, GDP와 인구1인당 연구개발지출, SCI논문과 연구원 1인당 연구개발 지출
2009 [7]	시계열분석	슈퍼컴퓨터 1위, 3위, 5위, 10위, 15위, 20위, 30위의 시계열 자료
	전문가 서베이	슈퍼컴퓨터의 중요성, 슈퍼컴퓨터 5호기 수요예측, 슈퍼컴퓨터 5호기의 성능예측 조사
2011 [16]	델파이	6개 연구영역에서 사회적·과학적 중요성, 과제의 규모, 컴퓨팅환경, 슈퍼컴퓨터의 소요연산량, 메모리 사이즈, 입출력 양, 해결방법 등의 조사를 통해 과학연구와 기술개발 과제

국가지원 슈퍼컴퓨터센터의 발전 방안에 관한 연구[2]에서는 Top500에 속하는 슈퍼컴퓨터를 보유하고 있는 19개 국가들을 대상으로 통계적 분포와 슈퍼컴퓨터의 수요결정요인들을 이용하여 국내 슈퍼컴퓨터 수요를 분석하였으며, 인구증가율, 일인당 연구개발비 등이 안정적이라고 가정하고 항상소득이 연간 5% 정도 성장한다면 Rmax 기준으로는 0.951% 증가할 것으로 예측하였다. 슈퍼컴퓨팅 관점에서의 미래 6T 수요분석과 개발활성화 연구[3]에서는 6T발전을 위한 슈퍼컴퓨터의 활용방안으로 KISTI 슈퍼컴퓨팅 센터의 이용기관수, 로그인수, 작업건수 등의 시계열 데이터를 중심으로 수요변화를 예측하였다. KISTI 슈퍼컴퓨터 4호기 도입 타당성분석[4]에서는 거시경제지표 및 학술연구지표를 독립변인으로 하는 회귀분석모델을 통해 국내 학술연구용 슈퍼컴퓨터의 적정 규모

를 예측하였다. 슈퍼컴퓨터 3호기의 경제적 파급효과분석[7]에서는 Top 500에서 제공하는 순위별 성능 자료를 활용하여 슈퍼컴퓨터 순위변화 추이예측을 수행하고, 특정시점에서 예측되는 1위, 3위, 5위, 10위, 15위, 20위, 30위의 성능 예측치를 제시하였다. 일본의 세계최고 컴퓨터 개발프로젝트 기획연구[15]에서는 2015~2020년 사이에 HPC를 사용하여 수행될 가능성이 높은 과학연구와 기술개발 과제를 과학적 중요성, 과제규모, 컴퓨팅 환경, 슈퍼컴퓨터 소요연산량 등을 기준으로 조사하였으며 일본 K컴퓨터 성능목표치 설계에 활용하였다. 슈퍼컴퓨터 기술예측 연구[18]에서는 DEA(Data Envelopment Analysis) 기반 기술예측을 이용하여 엑사수준 컴퓨터의 성능을 측정하고 진보율을 계산하였다. 결과에 따르면 2021년 초와 2022년 후반에 hybrid 시스템이나 다중프로세스(manycore) 시스템 형태로 엑사수준 성능을 달성할 수 있을 것으로 예측하였다. 기존 연구들과 본 연구의 차이점을 요약하자면 슈퍼컴퓨팅 성능에 영향을 미칠 것으로 판단되는 GDP, R&D투자금액, 연구원수, SCI논문수를 고려한 다중회귀분석을 통해 횡단면 분석을 실시하고, 분석결과 값에 Top500 자료의 성능(Rmax)값에 대한 시계열분석을 통해 도출된 기간별 기술진보율을 반영하여 미래 슈퍼컴퓨팅 성능수요를 추정하는 방법을 적용하였다.

2. 슈퍼컴퓨터 성능예측 활동의 특징 및 문제점

슈퍼컴퓨터의 성능예측은 주로 슈퍼컴퓨터시스템 도입과 성과분석 등의 시점에서 이루어졌다. 초기의 예측은 국내 슈퍼컴퓨터에 관한 시계열자료가 충분하지 않아 미국 Cray사와 일본산 슈퍼컴퓨터에 국한하여 시스템 수를 기준으로 예측하였으며, 이후에는 Top500 자료를 바탕으로 시스템 규모와 우리나라 연구개발 규모의 상관성을 바탕으로 당위적인 시스템 규모를 예측하거나 목표 수준을 정하고, Top500자료의 시계열 분석을 통한 성능예측을 수행하고 도입되어야 할 시스템의 규모 산정을 위한 근거로 활용하였다.

이처럼 인과분석을 이용할 때에는 시스템 발전의 동향을 반영하지 못하는 한계점이 있었고, 시계열 분석을 적용할 경우 국내 연구역량과 수준을 고려하지 않은 피상적인 목표설정을 통한 해당 수준의 성능예측에 머무르게 되는 단점을 가지고 있다고 할 수 있다.

Ⅲ. 슈퍼컴퓨팅 성능예측

1. 슈퍼컴퓨팅 성능예측 모델 구축

국가 차원의 슈퍼컴퓨터 성능의 타당한 예측을 위해 앞장에서 정리한 바와 같이 기존 연구의 단점을 보완하고 예측기법의 장점을 살릴 수 있는 방법을 고안하였다. 전술한 것과 같이 Top500자료의 특징은 슈퍼컴퓨팅 성능발전의 추세를 대단히 적절히 표현하고 있는 자료로 추세정보에 강점이 있다. 아울러 국가의 연구개발 역량, 연구인구, 투자금액 등을 종합한 결과가 슈퍼컴퓨터 성능에 미치는 영향의 관계로 해석할 때 현황을 바탕으로 보다 종합적으로 조망할 수 있다.

이러한 특징을 살리고자 슈퍼컴퓨팅 성능에 영향을 미칠 것으로 판단되는 GDP, R&D 투자금액(GERD), 연구원수(REP), SCI논문수를 고려한 다중회귀분석을 통해 횡단면 분석을 실시한 후, 이 값에 Top500 자료의 성능(Rmax)값을 이용한 시계열분석을 통해 도출된 기간별 기술진보율을 곱하여 미래 슈퍼컴퓨터의 성능을 예측하는 방법을 제안한다. 즉, 횡단면 분석을 통한 슈퍼컴퓨터 성능에 영향을 미치는 요인에 대한 영향력을 평가하고 이 결과를 슈퍼컴퓨터 성능발전 추세에 대입하여 목표시점의 성능 값을 예측하고자 하며, 이를 식으로 표현하면 다음과 같다.

$$성능_t = [(요인계수_1 \times 요인_1) + \cdots + (요인계수_n \times 요인_n)] \times 기술진보율$$

t시점의 성능예측 값은 기준시점에서의 요인과 요인 계수, 추정시점에서의 요인변화 예측값, 그리고 기술진보율로 표현된 단위 기간당 성능향상 정도를 이용하여 계산할 수 있다.

2. 슈퍼컴퓨팅 성능예측

2.1 예측 대상과 목표 성능

개발된 예측모형을 적용하기 위해 예측 대상 국가와 대상연도 그리고 목표순위를 우선 설정해야 한다. 예측 대상 국가는 대한민국으로 하며, 실용성을 위해 예측 대상연도

는 KISTI 슈퍼컴퓨터 5호기 도입이 예정된 2016년을 목표로 하여 슈퍼컴퓨터 시스템의 성능을 예측한다. 아울러 대한민국이 정책적으로 추진해야할 성과목표 달성을 위한 슈퍼컴퓨터 성능 보유 목표순위 설정이 필요한데 <표 3>과 같이 우리나라가 단·중기 내에 도달할 수 있는 목표순위로 미국, 중국, 일본, 독일, 영국, 프랑스 등 6대 강국의 슈퍼컴퓨팅 역량은 8~100배의 위치에 있어 이를 단기간 내에 쉽게 넘어설 수 없다. 하지만 이탈리아, 호주, 러시아, 캐나다, 인도 등 경쟁국과의 컴퓨팅 수준은 3배 이내로 근접해 있으므로 7위를 목표수준으로 설정하고 추진해 볼만하다고 판단된다. 또한 한국의 GDP규모와 R&D투자규모를 고려하고 전문가 의견 수렴결과와 정책입안자 협의를 통해 세계 7위의 슈퍼컴퓨팅 역량보유를 목표로 설정하고 예측을 수행하였다.

<표 3> 2013년도 Top500 결과의 국가별 비교

순위	국가	컴퓨터 수	Rmax합계(Gflops)	Rmax비율(%)
1	미국	252	106,833,637.96	47.77
2	중국	66	47,485,017.68	21.23
3	일본	30	20,307,189.00	9.08
4	독일	19	11,351,753.71	5.08
5	프랑스	23	8,938,486.20	4.00
6	영국	29	8,082,236.56	3.61
7	인도	11	2,690,461.00	1.20
8	이탈리아	6	2,422,982.00	1.08
9	호주	5	2,056,248.82	0.92
10	러시아	8	2,012,186.00	0.90
11	캐나다	9	1,772,008.70	0.79
12	스위스	4	1,405,767.00	0.63
13	사우디아라비아	4	1,272,515.00	0.57
14	스웨덴	7	1,161,292.60	0.52
15	스페인	3	1,126,860.00	0.50
16	**한국**	**4**	**1,014,400.00**	**0.45**
17	노르웨이	3	735,400.00	0.33
18	브라질	3	626,000.00	0.28
19	폴란드	3	561,203.00	0.25
20	핀란드	2	378,000.00	0.17
21	이스라엘	2	283,965.00	0.13
22	네덜란드	2	258,034.00	0.12
23	홍콩	1	234,248.00	0.10
24	대만	1	177,100.00	0.08

순위	국가	컴퓨터 수	Rmax합계(Gflops)	Rmax비율(%)
25	덴마크	1	162,098.00	0.07
26	오스트리아	1	152,900.00	0.07
27	벨기에	1	152,347.90	0.07
합 계			223,654,338.13	100

2.2 성능 요인의 도출

국가별 슈퍼컴퓨팅 성능 보유현황은 Top500 등재 데이터의 Rmax값을 기준으로 변동성을 완화하기 위하여 2011년 6월부터 2012년 11월까지 18개월 동안의 데이터 값 평균에 자연로그 값을 취하여 활용하였다. 국가별 슈퍼컴퓨팅 보유 성능에 영향을 주는 요인들로는 국가경제 규모, 연구개발 투자규모, 과학기술의 발전수준이나 고도화 정도 등을 고려할 수 있다. 기존 선행연구들에 따르면 GDP규모와 인구, GDP연구개발투자 비중 또는 1인당 연구개발비 등이 슈퍼컴퓨팅 수요를 설명해줄 수 있는 주요 요인이 될 수 있다고 보았다. 본 연구에서는 경제규모를 대표하는 요소로 GDP, 연구개발의 투자규모를 대표하는 변수로 총연구개발지출(GERD)와 연구원 수, 과학기술의 발전수준이나 고도화 정도를 나타내는 변수로 5년간 SCI 논문수를 슈퍼 컴퓨팅 보유량에 영향을 미치는 주요 변수로 고려하였다(<표 4>). 본 연구에서 고려한 종속변수와 독립변수의 데이터 정규성 및 데이터 치우침 현상 완화를 위해 자연로그를 취한 값을 대상으로 데이터 정규성 검정을 실시하였다(<표 5>). 본 연구에서 고려한 독변수와 종속변수를 정리하면 다음과 같다. 분석대상으로 고려한 국가는 독립변수 데이터의 가용성과 Top500에 지속적으로 등재된 21개 국가[3]를 선정하였다.

<표 4> 다중회귀분석을 위한 독립변수와 종속변수

변수 구분	변수명	설 명	자료출처
종속변수	RmaxAvg	· Rmax(Gflops) · 2011.06~2012.11	Top500
	LNRmaxAvg	· RmaxAvg 자연로그값(billion $)	

3) 분석대상 국가를 영문알파벳 순으로 나열하면 호주, 오스트리아, 벨기에, 캐나다, 중국, 덴마크, 핀란드, 프랑스, 독일, 이스라엘, 이탈리아, 일본, 한국, 폴란드, 러시아, 스페인, 스웨덴, 스위스, 대만, 영국, 미국이다.

독립변수	GDPAvg	· GDP(billion $) · 2007~2011 평균	OECD
	LGDPAvg	· GDPAvg 자연로그값(billion $)	
	GERDAvg	· GERD(million $) · 2007~2011 평균	OECD
	LNGerdAvg	· GERDAvg 자연로그값(million $)	
	REPAvg	· 연구원수(Fulltime R&D Personnel) · 2007~2011 평균	OECD
	LNREPAvg	· REPAvg 자연로그값	
	SCISum	· 2005~2009 SCI논문수(NSI기준)	'09 SCI분석연구/교과부
	LNSCISum	· SCISum 자연로그값	

변수들의 정규성 검정을 위해 기술통계량 분석을 수행한 결과는 <표 5>와 같다.

<표 5> 정규성 분석결과

	Kolmogorov-Smirnova			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
GDPAvg	.267	21	.000	.610	21	.000
LGDPAvg	.169	21	.120	.956	21	**.447**
GERDAvg	.332	21	.000	.554	21	.000
LNGERDAvg	.124	21	.200*	.943	21	**.252**
REPAvg	.277	21	.000	.672	21	.000
LNREPAvg	.175	21	.092	.916	21	**.071**
SCISum	.268	21	.000	.580	21	.000
LNSCISum	.124	21	.200*	.949	21	**.320**
RMaxAvg	.338	21	.000	.458	21	.000
LNRMaxAvg	.145	21	.200*	.931	21	**.142**

a. Lilliefors 유의확률수정

*. 이것은 참인 유의확률의 하한값입니다.

일반적으로 데이터 수가 50개 보다 적은 경우 Shapiro-Wilk 정규성 검정을 이용한다. 정규성 분석결과에 따르면 독립변수와 종속변수 모두 자연로그값을 취한 데이터들의 유의확률이 0.05보다 커서 변수들이 정규분포를 이룬다는 귀무가설을 채택한다. 따라서 본 연구에서는 독립변수와 종속변수 모두 자연로그 값을 취한 변수를 다중회귀분석 대상변수로 선정하였다.

최적의 다중회귀분석 모형을 도출하기 위해 독립변수로 슈퍼컴퓨팅 성능(LNRmaxAvg),

종속변수로 GDP(LGDPAvg), R&D투자(LNGERDAvg), 연구원수(LNREPAvg), SCI논문수(LNSCISum)를 선택하고 다중회귀분석을 실행하면 다음과 같은 결과를 얻을 수 있으며 기술통계량 분석결과는 <표 6>과 같다. 상관계수 분석결과에 따르면 독립(설명)변수들간의 상관계수 값이 커서 다중공선성의 존재할 가능성이 높으므로 독립변수의 선택에 따른 최적모형을 선정할 필요가 있는 것으로 분석되었다.

<표 6> 기술통계량

변 수	평 균	표준편차	N
LNRMaxAvg	13.894729	1.6944130	21
LGDPAvg	7.026176	1.1413620	21
LNGERDAvg	10.126910	1.1698657	21
LNREPAvg	12.228057	1.1717653	21
LNSCISum	11.989567	.9239240	21

<표 7> 모형에 따른 제거변수

진입/제거된변수[b]			
모형	진입된 변수	제거된 변수	방법
1	LNSCISum, LNREPAvg, LNGERDAvg, LGDPAvg[a]	.	입력
2	.	LNREPAvg	후진(기준: 제거할 F의 확률>=.100).
3	.	LNSCISum	후진(기준: 제거할 F의 확률>=.100).

a. 요청된 모든 변수가 입력되었습니다.
b. 종속변수: LNRMaxAvg

본 연구에서는 최적모형 선정을 위해 후진제거법을 이용하여 다중회귀분석을 수행하였다. <표 7>에서 [모형 1]은 모든 변수를 포함시켜 분석이 수행되었고, [모형 2]에서는 연구원수(LGREPAvg) 변수가 제거되었고, [모형 3]에서는 SCI논문수(LNSCISum) 변수가 제거되었다.

잔차의 독립성 여부를 판단하기 위하여 수행된 <표 8>의 분석결과 중 Durbin-Watson 값을 살펴보면 2.364로 0과 4사이에 존재하므로 잔차의 독립성이 보장된다고 할 수 있다. 추정된 회귀모형의 통계적 유의성 분석결과인 <표 9>와 <표 10>에 따르면 3개 모형 모두 통계적으로 유의하고 [모형 1]에서 [모형 3]으로 갈수록 F검정 통계량이 커짐을 알 수 있다. 또한 [모형 1]에서 [모형 3]으로 갈수록 수정된 R^2값이 증가함을 알 수

있으며 이는 [모형 3]이 가장 적합한 모형이라는 것을 의미한다.

<표 8> 모형별 분석결과

모형요약[d]					
모형	R	R 제곱	수정된 R제곱	추정값의 표준오차	Durbin-Watson
1	.964[a]	.930	.912	.5021793	
2	.964[b]	.930	.917	.4874331	
3	**.962[c]**	**.926**	**.918**	**.4859681**	**2.364**

a. 예측값: (상수), LNSCISum, LGREPAvg, LGGERDAvg, LGDPAvg
b. 예측값: (상수), LNSCISum, LGGERDAvg, LGDPAvg
c. 예측값: (상수), LGGERDAvg, LGDPAvg
d. 종속변수: LNRMaxAvg

<표 9> 분산분석 결과

분산분석[b]						
모 형		제곱합	자유도	평균제곱	F	유의확률
3	회귀모형	53.170	2	26.585	112.569	.000[a]
	잔 차	4.251	18	.236		
	합 계	57.421	20			

a. 예측값: (상수), LNGERDAvg, LGDPAvg
b. 종속변수: LNRMaxAvg

<표 10> 모형별 추정계수

계수[a]								
모 형		비표준화계수		표준화계수	t	유의 확률	공선성통계량	
		B	표준오차	베타			공차	VIF
3	(상수)	2.178	1.087		2.004	.060		
	LGDPAvg	.927	.252	.625	3.679	.002	.143	7.007
	LNGERDAvg	.514	.246	.355	2.089	.051	.143	7.007

a. 종속변수: LNRMaxAvg

최적모형 탐색을 통해 도출된 회귀모형은 다음과 같이 주어지며 회귀모형의 설명력은 91.8%이다.

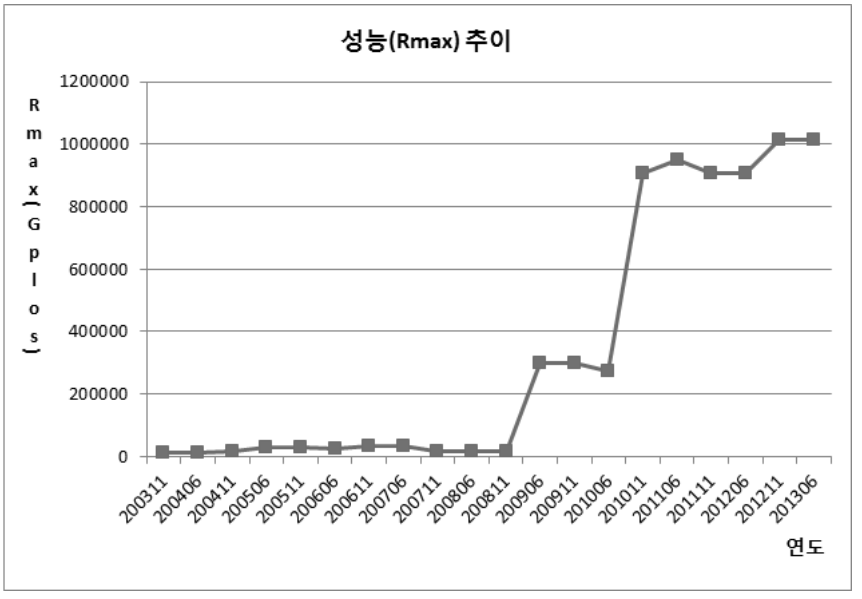
$$LN(Rmax) = 2.178 + 0.927 \times LN(GDP) + 0.514 \times LN(GERD)$$

다른 변수들이 일정하다고 할 때, GDP가 10억 달러 증가하면 슈퍼컴퓨팅의 성능(Rmax)은 0.927 Gflops 증가하고, GERD가 1백만 달러 증가하면 0.514 Gflops 증가한다는 것을 의미한다.

2.3 기술진보율 추정

가. 한국의 슈퍼컴퓨팅 성능 시계열 데이터 분석

Top500 기준 한국의 슈퍼컴퓨팅 성능수요 예측을 위해 2003년 11월부터 2013년 6월까지 10년간의 시계열데이터를 수집하고 성능변화를 살펴보면 10년 동안 전체 변화양상은 지수 함수적으로 증가하는 추세에 있으나, 10년을 몇 개 구간으로 나누어 보면 계단식 증가함수 형태를 보이고 있다. 이는 몇 년 주기로 슈퍼컴퓨터를 신규 도입하고 있는 정부의 슈퍼컴퓨팅 분야 투자정책에 기인하는 것으로 판단된다.



<그림 1> 한국의 슈퍼컴퓨팅 성능 보유추이

슈퍼컴퓨팅 성능(Rmax)값들이 지수함수적으로 증가하는 형태를 취하고 있으므로 성능에 자연로그를 취한 ln(성능)을 종속변수로 하고 성능평가 시점에 해당하는 시간을 독립변수로 하는 회귀모형을 통해 2016년 슈퍼컴퓨팅의 성능을 예측하고자 한다. 회귀분석 모형은 다음과 같이 주어진다.

$$\text{Ln}(\text{성능}) = \alpha + \beta x + \epsilon$$

여기서, Ln(성능)은 성능값에 자연로그를 취한 값이고 α는 회귀상수, β는 회귀계수, x는 시간, ϵ은 잔차를 의미한다. 회귀분석 결과는 다음과 같다

<표 11> 회귀분석 결과

모형요약[b]					
모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차	Durbin-Watson
1	.904[a]	.818	.808	.8009143	.849

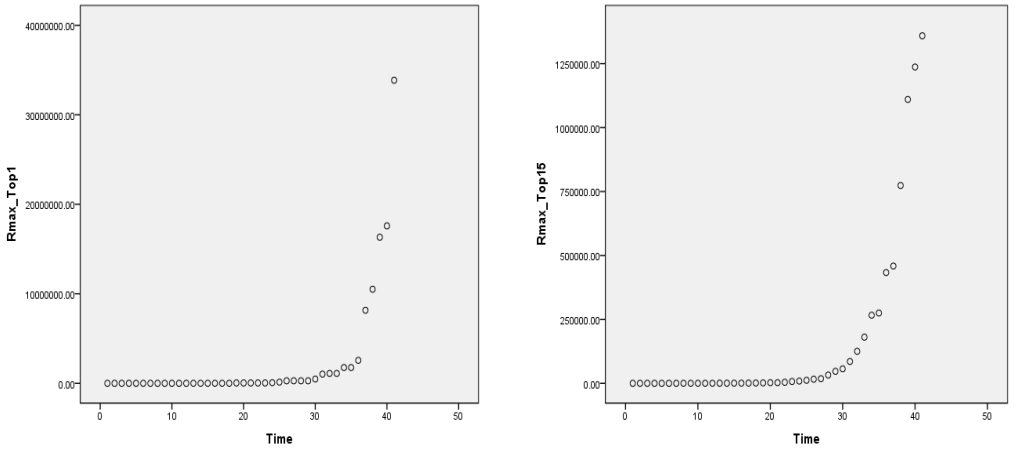
a. 예측값: (상수), Time
b. 종속변수: LnKorea

회귀모형은 통계적으로 유의하고 정규성과 다중공선성 조건을 만족하는 것으로 분석되었으며 도출된 회귀모형은 다음과 같이 주어진다. 모형의 설명력을 나타내는 수정된 R^2는 0.808로 전체 변수의 80.8%를 설명하는 것으로 나타났다. 회귀모형에서 회귀계수는 시간이 1단위 증가할 때 성능은 몇 % 증가하는지를 의미한다. 도출된 모형의 경우 시간이 한 단위 증가할 때 성능은 27.9% 증가한다고 해석할 수 있다. 회귀분석 모형에서 고려한 시간 간격은 Top500 성능 발표주기인 6개월이 시간단위이므로 한국 슈퍼컴퓨터의 경우 6개월 마다 성능이 27.9% 증가한다고 할 수 있다.

$$\text{Ln}(\text{성능}) = 8.543 + 0.279x$$

나. 개별 시스템 기술 진보율

Top500에 등재된 순위별 데이터를 이용하여 1위, 3위, 5위, 7위, 10위, 13위, 15위, 20위, 30위의 성능(Rmax) 시계열 데이터를 이용한다. 시계열 데이터의 변화 형태를 살펴보기 위하여 순위별 데이터의 산점도를 도출하면 <그림 2>와 같다. 산점도 결과에 따르면 순위별 성능값들의 변화는 지수함수 형태를 보이는 것으로 판단된다. 따라서 지수 함수적 성장 패턴을 보이는 성능(Rmax)에 자연로그를 취한 ln(성능)을 종속변수로 하고 1993년 6월부터 2013년 6월까지 시계열을 독립변수로 하는 회귀분석($\ln Rmax_i = \alpha + \beta x + \epsilon$)을 통해 미래 슈퍼컴퓨팅 성능을 예측한다.



<그림 2> Top500 시스템의 1위, 15위 시스템의 성능변화

본 연구에서 순위별 성능예측을 위해 설정한 회귀분석 모형은 아래와 같다.

$\ln Rmax_i = \alpha + \beta x + \epsilon$과 같다.

여기서 $\ln Rmax_i$는 순위가 i인 성능의 자연로그 값을 의미하고, α는 회귀상수, β는 회귀계수, x는 시간, ϵ은 잔차를 의미한다. <표 12>는 회귀모형을 이용하여 Top500에 제시된 성능 1위, 3위, 5위, 7위, 10위, 13위, 15위, 20위, 30위에 대한 회귀분석 결과를 나타낸다. 1위, 3위, 7위, 10위, 13위, 15위를 분석대상으로 고려한 이유는 2013년 기준으로 16위인 한국의 성능순위를 고려할 때 벤치마킹 대상으로 고려해야 할 목표순위의 성능규모를 파악하기 위하여 선정하였고, 20위와 30위는 시의적절한 투자가 이루어지지 못하여 2013년 보다 순위가 하락하는 비관적 경우의 성능규모 파악을 위하여 분석대상으로 고려하였다.

<표 12> 개별 시스템 순위 예측모형별 회귀분석 결과

모형	회귀계수	$p < \alpha$	조정된 R^2	관측수
Rank 1	0.323	0.00	0.991	41
Rank 3	0.319	0.00	0.991	41
Rank 5	0.313	0.00	0.995	41
Rank 7	0.312	0.00	0.996	41
Rank 10	0.307	0.00	0.996	41

모형	회귀계수	$p < \alpha$	조정된 R^2	관측수
Rank 13	0.300	0.00	0.997	41
Rank 15	0.299	0.00	0.997	41
Rank 20	0.297	0.00	0.997	41
Rank 30	0.297	0.00	0.997	41

분석결과에 따르면 9개 분석모형 모두 통계적으로 유의하며 회귀모형의 설명력을 나타내는 조정된 R^2의 값이 0.991~0.997범위에 속하는 것으로 나타나 설명력이 매우 높은 것으로 분석되었다. 회귀계수는 시간이 1단위 증가할 때 성능은 몇 % 증가하는지를 의미하므로 순위가 1위인 Rank 1모형의 경우 시간이 한 단위 증가할 때 성능은 32.3% 증가한다고 해석할 수 있다. 회귀분석 모형에서 고려한 시간 간격은 Top500 성능 발표 주기인 6개월이 시간단위이므로 Top500의 순위 1인 슈퍼컴퓨터의 경우 6개월 마다 성능이 32.3% 증가하며, 15위권의 슈퍼컴퓨터의 경우 29.9%씩 증가한다고 할 수 있다.

다. 국가별 성능 총량 순위분석

위와 같은 방법으로 Top500에 등재된 시스템을 국가별로 취합하여 슈퍼컴퓨팅 성능 보유량 순위 1위, 3위, 5위, 7위, 8위, 9위, 10위, 13위, 15위를 대상으로 시계열 분석을 통해 국가별 성능 진보율을 도출하면 <표 13>과 같다.

분석결과에 따르면 9개 분석모형 모두 통계적으로 유의하며 회귀모형의 설명력을 나타내는 조정된 R^2의 값이 0.993~0.997범위에 속하는 것으로 나타나 설명력이 매우 높은 것으로 분석되었다. Top500의 순위 1위인 국가의 경우 슈퍼컴퓨터 성능이 6개월 마다 성능이 30.6% 증가하며, 순위 7위인 국가는 32.2%씩 증가한다고 할 수 있다.

<표 13> 국가별 성능 총량 예측모형별 회귀분석 결과

모형	회귀계수	$p < \alpha$	조정된 R^2	관측수
Rank 1	0.306	0.00	0.997	41
Rank 3	0.309	0.00	0.995	41
Rank 5	0.321	0.00	0.996	41
Rank 7	0.322	0.00	0.993	41
Rank 8	0.325	0.00	0.996	41
Rank 9	0.326	0.00	0.996	41
Rank 10	0.327	0.00	0.995	41
Rank 13	0.334	0.00	0.995	41
Rank 15	0.336	0.00	0.995	41

3. 한국의 슈퍼컴퓨팅 성능예측

2011년 11월을 기준으로 슈퍼컴퓨터 성능에 영향을 미치는 주요요인은 GDP와 GERD 값이므로 GDP는 IMF의 World Economic Outlook Database(2012.11), GERD 값은 2007년부터 2011년까지 GERD의 연평균증가율인 5%(≒4.91%)를 적용하여 성능수요를 예측해보면 성능수요는 1.24Pflops로 추정되고 실제 값은 0.9Pflops이다.

2016년까지의 주요요인의 증감을 고려하여 예측값을 추정하려면 2016년 GDP와 GERD 값이 필요하다. 2016년 GDP예측치는 IMF의 World Economic Outlook Database(2012.11) GDP 예측치를 바탕으로 1,520.59억달러를 이용하고, GERD 값은 2007년부터 2011년까지 GERD의 연평균증가율인 5%(≒4.91%)를 적용하여 76,029백만달러를 이용한다. 2016년의 경우 기술진보율을 고려하지 않은 2011년 기준의 목표성능은 2.5PF로 추정되었다.

<표 14> 요인수준변화만을 고려한 한국의 슈퍼컴퓨팅 성능예측

기간	2011.11	2012.11	2013.11	2014.11	2015.11	2016.11	2017.11	2018.11
성능(PFlops)	1.24	1.33	1.61	1.85	2.14	2.5	2.93	3.44

이에 국내 슈퍼컴퓨터 성능의 진보율(27.9%/6개월)을 고려했을 때 29.29PF 의 슈퍼컴퓨팅 성능수요가 필요할 것으로 예측 되었다.

<표 15> 요인수준변화와 한국의 기술진보율을 고려한 한국의 슈퍼컴퓨팅 성능예측

기간	2011.11	2012.11	2013.11	2014.11	2015.11	2016.11
성능(PFlops)	1.24	2.18	4.31	8.10	15.32	29.29

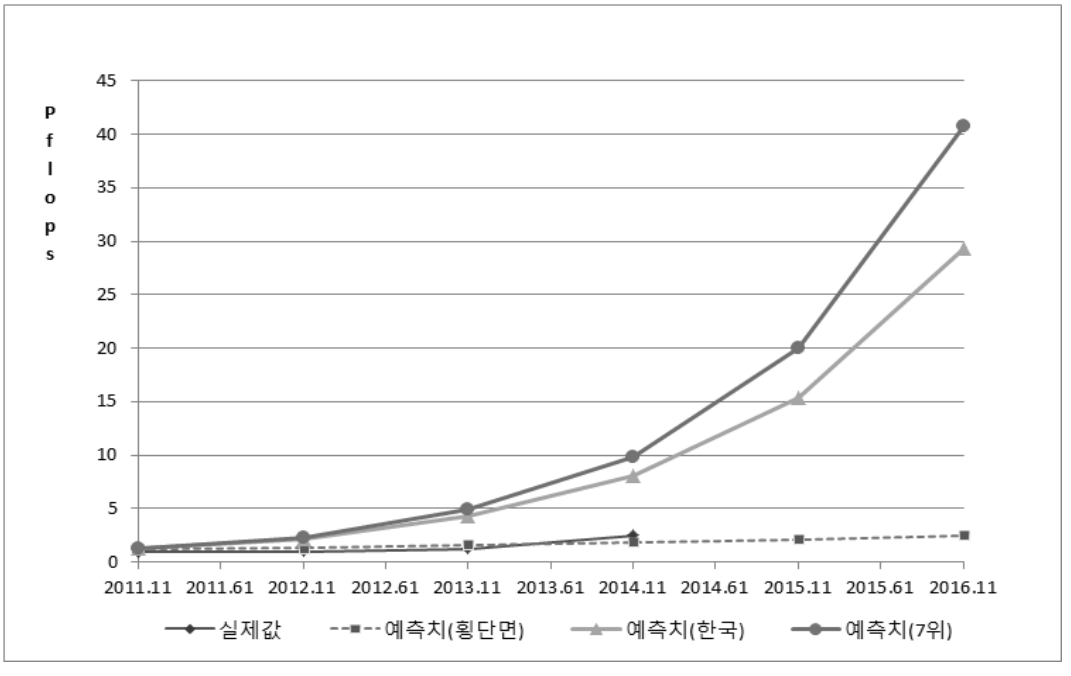
<표 16> 횡단분석과 추세분석을 이용한 한국의 슈퍼컴퓨팅 성능예측

기간			2011.11	2012.11	2013.11	2014.11	2015.11	2016.11	2017.11	2018.11
연도	기준진보율(%)		1.24	1.33	1.61	1.85	2.14	2.5	2.93	3.44
2011	한국	27.9	1.24	2.03	3.32	5.43	8.88	14.53	23.76	38.87
	7위 기준	32.2	1.24	2.17	3.79	6.62	11.57	20.22	35.33	61.75
2012	한국	27.9	-	2.18	3.56	5.82	9.52	15.58	25.49	41.69
	7위 기준	32.2	-	2.32	4.06	7.10	12.41	21.69	37.90	66.24
2013	한국	27.9	-	-	4.31	7.05	11.53	18.86	30.85	50.47
	7위 기준	32.2	-	-	4.92	8.59	15.02	26.25	45.88	80.18

기간			2011.11	2012.11	2013.11	2014.11	2015.11	2016.11	2017.11	2018.11
2014	한국	27.9	-	-	-	8.10	13.25	21.67	35.45	57.99
	7위 기준	32.2	-	-	-	9.88	17.26	30.16	52.72	92.13
2015	한국	27.9	-	-	-	-	15.32	25.07	41.01	67.08
	7위 기준	32.2	-	-	-	-	19.96	34.89	60.98	106.57
2016	한국	27.9	-	-	-	-	-	29.29	47.91	78.37
	7위 기준	32.2	-	-	-	-	-	40.76	71.24	124.50

GDP수준과 GERD 값이 2011년 상황으로 고정이라고 가정하고 한국의 기술진보율을 적용하면 2016년에는 14.53PF급 슈퍼컴퓨팅 성능이 요구되며, GDP와 GERD값의 변동을 감안하고 한국의 기술진보율을 적용하면 29.29PF의 슈퍼컴퓨팅 성능이 요구되는 것으로 예측되었다. 또한 목표순위인 7위의 기술진보율을 적용하고 한국의 GDP수준과 GERD 값이 2011년 상황으로 고정이라고 가정하면 20.22PF이고, 2016까지의 GDP와 GERD 값의 변화를 반영하면 총 40.76PF 규모의 성능이 필요할 것으로 예측되었다.

<표 16>은 연도별 GDP, GERD 변화 수준과 기술진보율을 고려하고 한국과 7위 수준 국가의 기술진보율을 적용하였을 경우 개별 기간별 한국의 슈퍼컴퓨터 성능보유 예측량의 변화를 나타내고, 예측 결과를 그림으로 나타내면 <그림 3>과 같다.



<그림 3> 한국의 슈퍼컴퓨팅 성능수요 예측결과

분석에 이용한 대상 기간과 데이터가 상이하여 기존 연구결과와 본 연구결과를 직접적으로 비교하는 것은 불가능하지만 기존 연구가 본 연구와 동일한 데이터와 분석기간을 적용하였다고 가정하고 기존연구의 분석결과와 본 연구의 분석결과를 요약하면 <표 17>과 같다. 세계적으로 슈퍼컴퓨터 교체주기가 5년인 점을 고려하면 분석시점 기준으로 도입시점까지의 기술진보율을 반영하는 것이 바람직하다고 판단되며 분석결과의 차이는 이에 기인하는 것으로 유추된다.

<표 17> 기존 연구와 본 연구의 결과

구분	기존연구		본 연구	
	횡단면분석	시계열분석	한국	7위 목표
2016년 예측치 (PFlops)	2.5	9.6	15~30	20~40

Ⅳ. 결론

횡단면분석과 기술진보율을 적용하여 2016년 한국의 슈퍼컴퓨팅 성능수요를 예측해 본 결과 현재의 추세를 이용할 경우 15~30PF 정도, 목표 국가수준의 추세를 이용할 때 20~40PF정도의 컴퓨팅 역량이 필요할 것으로 예측되었다. 이 결과는 단순 회귀분석을 적용한 결과인 9.6PF와 횡단면분석을 적용한 결과인 2.5PF와 큰 차이를 나타내었다.

횡단면분석의 경우 국가 경제수준과 R&D투자규모에 영향을 받는 것으로 나타났다. 단일 시스템의 경우 순위가 높을수록 기술 진보율이 높았으며, 국가집계의 경우 상위권 국가보다는 하위권 국가의 기술진보율이 약간 높은 것으로 나타났다. 이처럼 경제 및 연구개발 수준 그리고 기술진보의 특성을 반영한 예측을 통해 슈퍼컴퓨팅 관련 투자정책 의사결정에 필요한 기초정보를 제공할 수 있을 것으로 판단된다.

그 동안 슈퍼컴퓨팅분야의 성능수요에 대한 예측이 단편적으로 이루어져 왔다. 이는 슈퍼컴퓨터 성능에 관한 예측활동이 서두에서 살펴본 바와 같이 목적과 용도에 따라 다양하다는 특성과 예측 데이터의 한계로 어려움이 따르는 것으로 보여진다. 본 연구에서 제시한 방법론을 참고하여 보다 신뢰성 있는 데이터를 생산・확보하고 목적에 맞는 성능예측을 수행한다면 보다 합리적인 예측결과로 목적에 부합한 정책수립과 보다 효과적

인 정책집행에 기여할 수 있을 것으로 판단된다. 연구자들에게 제공되는 서비스의 수요와 공급법칙을 고려해 볼 때, 본 연구에서 도출한 성능수요 예측결과는 정책 의사결정을 위한 하나의 기준자료로 활용 가능하다고 판단되며, 보다 정확한 성능수요 예측을 위해서는 전문가 서베이를 통한 연구분야 및 연구단계별 수요조사를 바탕으로 슈퍼컴퓨팅 적정 성능수요를 도출하고, 두 결과에 대한 비교·평가를 통한 종합적인 검토가 필요할 것으로 판단된다.

참고문헌

(1) 국내문헌

IDC (2012), KISTI 국가슈퍼컴퓨팅센터 정책연구 보고서.

KISTI (2000), 국가지원 슈퍼컴퓨터센터의 발전 방안에 관한 연구.

KISTI (2003), 슈퍼컴퓨팅 관점에서의 미래 6T 수요분석과 개발활성화 연구.

KISTI (2005), KISTI 슈퍼컴퓨터 4호기 도입 타당성 분석.

KISTI (2009), 슈퍼컴퓨터 3호기의 경제적 파급효과분석.

권성훈・홍순기 (2009), "델파이 기술예측의 타당성과 신뢰성 분석에 관한 연구", 「기술혁신연구」 제17권 제3호, pp. 97-117.

김소영 외 (2007), 국가슈퍼컴퓨팅 공동활용체제 구축의 타당성 분석.

시스템공학연구소 (1991), 슈퍼컴퓨터 장기수요예측.

신태영 (2007), "기술혁신과 경제성장: 요소대체율, 기술진보율 및 연구개발투자", 「기술혁신연구」 제15권 제2호, pp. 1-24.

이형진 외 (2012), "슈퍼컴퓨터 예측활동에 관한 연구", 「기술경영경제학회 하계학술대회 논문집」.

(2) 국외문헌

Dongarra, J.J. (1986), Performance of various Computers Using Standard Linear Equations Software in a Fortran Environment.

Feitelson, D.G. (2005), "The supercomputer industry in light of the Top500 data", *Comput Sci Eng*, Vol. 7, pp. 42-47.

PITAC (2007), Computational Science: Ensuring America's Competitiveness.

DoE (2009), Scientific Grand Challenges, 워크숍자료.

National Research Council (2011), The Future of Computing performance.

일본문부과학성 (2011), 향후 HPC 기술의 연구개발 워크숍자료.

Peckham, M. (2013), RIKEN plans exascale supercomputer '30 times faster' than today's fastest in six years, Time.

Lim, D.J., T.R. Anderson and T. Shott (2015), "Technological forecasting of supercomputer development: The March to Exascale computing", *Omega*, Vol. 51, pp. 128-135.

Top500 home page: http://www.top500.org.

HPC Challenge Benchmark home page: http://icl.cs.utk.edu/hpcc.

□ 투고일: 2015. 01. 06 / 수정일: 2015. 03. 24 / 게재확정일: 2015. 03. 30