# Research and Implementation of Data Intensive Scalable Computing Platform for Scientific Applications

Jianhui Li[1], Yuanchun Zhou[1], Zhiyi Huang[1], Zhen Meng[1], Gang Huang[2], Lijun Fan[2], Zhiduan Chen[3], Hongmei Liu[4], Zhenghua Xie[1], Xiaoguang Lin[1], Qi Liu[1], Yong Liu[1]

[1]Computer Network Information Center, Chinese Academy of Sciences, 4 South 4th Street, Zhongguancun, Beijing, China
email: (lijh@cnic.cn, zyc@cnic.cn, zhiyi@cnic.cn, zhenm99@cnic.cn)
[2]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
[3]Institute of Botany, Chinese Academy of Sciences, Beijing, China
[4]Shenzhen Fairylake Botanical Garden, Chinese Academy of Sciences, Beijing, China

Science is increasingly becoming more and more data-driven. With technological advancements such as advanced sensing technologies that can rapidly capture data at high resolutions and Grid technologies that enable increasingly realistic simulation of complex numerical models, scientific applications have become very data-intensive and involve storing and accessing large amounts of data. Data intensive computing presents a significant challenge for traditional supercomputing architectures that maximize FLOPS since CPU speed has surpassed IO capabilities of HPC systems. Due to these data intensive scientific problems a new challenge is emerging, as many groups in science are facing analyses of data sets in tens of terabytes, eventually extending to a petabyte since disk access and data-rates have not grown with their size. The requirements for the data analysis environment are (i) scalability, including the ability to evolve over a long period, (ii) performance, (iii) ease of use, (iv)some fault tolerance and (v) most important low entry cost. Based on constructing a PB level storage and hundreds of computing nodes environment in Computer Network Information Center of Chinese Academy of Sciences, we present the efficient architecture for a range of data intensive computations operating on petascale data sets. The design goal includes a balanced system in terms of IO performance and memory size, and provides the integrated analysis platform which could do data acquisition, data pretreatment, data analysis and data mining, data visualization, and so on. This paper also takes two typical data intensive applications for example to invalidate the platform. One is Phylogenetic analysis of land plants platform, the other is Atmospheric Scientific Data On-line Analysis Platform.