Gupta's BigOnc Request Summary
Date submitted: Apr 18, 2013

Print

Request Summary and Awards

File Download

It may take a few minutes to zip or tar your document for download.

Click here to download all documents in ***ZIP*** format. (44.5 KBs)

Click here to download all documents in ***TAR*** format. (112.6 KBs)

Top of page

Title / FOS

Title	BigOnc
Request Number	MCB130121
Request Type	New
Primary Field Of Science	413 - Genetics and Nucleic Acids

Keywords

Oncology

Genomics

BigOnc

Top of page

PI Information

First Name	Amarnath
Middle Name	
Last Name	Gupta
Organization	San Diego Supercomputer Center
Position	Center Non-Researcher Staff
Degree	PhD
Degree Field	Computer Science
Department	Next Generation Tools for Biology

Address1	9500 Gilman Drive
Address2	Mail Code: CASCE
City	La Jolla
State	CA
Zip Code	92093
Country	United States
Email	gupta@sdsc.edu
Phone	858 822-0994
Fax	858 822-0906
URL(s)	http://bigonc-stsi.sdsc.edu

Top of page

Co-PI(s) Information

First Name	Mark
Middle Name	Alan
Last Name	Miller
Organization	University of California-San Diego
Position	Center Researcher Staff
Degree	PhD
Degree Field	Molecular Biology
Department	SDSC - Next Generation Tools for Biology
Address1	9500 Gilman Dr
Address2	Mail Code: 5050
City	La Jolla
State	CA
Zip Code	92093
Country	US
Email	mmiller@sdsc.edu
Phone	858-822-0866
Fax	
URL(s)	

DN(s)	CN=Mark Miller 0,O=National Computational Science Alliance,C=US
	/C=US/O=National Computational Science Alliance/CN=Mark Miller 0
	CN=Mark Miller 0,O=National Center for Supercomputing Applications,C=US
	/C=US/O=National Center for Supercomputing Applications/CN=Mark Miller 0
	CN=Mark Miller,O=TACC Classic CA,O=UT-AUSTIN,DC=TACC,DC=UTEXAS,DC=EDU
	/DC=EDU/DC=UTEXAS/DC=TACC/O=UT-AUSTIN/O=TACC Classic CA/CN=Mark Miller
	CN=Mark Miller,O=TACC MICS CA,O=UT-AUSTIN,DC=TACC,DC=UTEXAS,DC=EDU
	/DC=EDU/DC=UTEXAS/DC=TACC/O=UT-AUSTIN/O=TACC MICS CA/CN=Mark Miller
	CN=Mark Miller 0,OU=People,O=National Center for Supercomputing Applications,C=US

/C=US/O=National Center for Supercomputing Applications/OU=People/CN=Mark Miller 0
CN=Mark Miller 0,O=Pittsburgh Supercomputing Center,C=US
/C=US/O=Pittsburgh Supercomputing Center/CN=Mark Miller 0
CN=Mark Miller 0,O=National Institute for Computational Sciences,DC=TENNESSEE,DC=EDU
/DC=EDU/DC=TENNESSEE/DC=NICS/O=National Institute for Computational Sciences/CN=Mark Miller 0

Top of page

Supporting Grant(s) Information

PI Name	Mark Miller
Funding Agency	National Institutes of Health (any institute)
Funding Agency Division	CTSA
Program Officer Name	Unknown
Program Officer Email	Unkown
Funding Title	Scripps Translational Science Institute
Award Number	2008-1226
Awarded Amount	1245000
Percentage of award supporting this request	100
Start Date	05/01/2008
Expiration Date	04/30/2013
Field Of Science	Genetics and Nucleic Acids
Comment	We have been informed that funding will be extended although a grant extension has not been signed yet.

Top of page

Resources Requested

Please estimate what percentage of the work you expect to do in this allocation will be the following types (the 3 numbers should sum to 100):

- Production (actually doing research): 10
- Exploration/porting (preparing to do research): 90

Please estimate what percentage of the jobs you expect to run in this allocation will be the following types (the 3 numbers should sum to 100):

- Submitted through command line/script: 50
- Submitted through a metascheduler (to run on one of a set of resources, without user control over which of the set is chosen): 50

Please estimate what percentage of the science runs you expect to perform in this allocation will be the following types (the 4 numbers should sum to 100):

- Tightly-coupled (multiple jobs that will run simultaneously): 20
- Dependent (multiple jobs such as in a workflow): 80

Resource Name	XSEDE Extended Collaborative Support
Resource Requested Amount	12
Resource Awarded Amount	*12*

What do you want to accomplish with the help of expert staff? Have you already done any work on this aspect of your software?

- I have had some exposure to Hadoop and Cassandra on a cluster of nodes at the center. I hope to gain important experience with the use of MapReduce functionality within Hadoop and any types for object designs in Cassandra would be very beneficial.

How would the success of this collaboration benefit your project?

- This project would benefit by this collaboration because the standard hardware tecnologies as well as traditional software applications, operating systems and file systems are clearly inadequate i the face of this volume of data.

Which member(s) of your team would collaborate with ECSS staff?

- Mostly myself (Bill West). O would anticipate some involvement by Amarnath Gupta as well

Have you had significant interaction on previous projects related to your current proposal or discussed your extended support needs with any XSEDE staff? If so, please indicate with whom

- No. However, Amarnath Gupta had a high level discussion with Mike Norman concerning the efficacy of building out such a framework on a platform such as Gordon.

Have you received TeraGrid/XSEDE advanced support in the past?. If so, please indicate the time period, and how the support you received then relates to the support you request now.

- No.

Resource Name	SDSC Appro with Intel Flash I/O Nodes (Gordon ION)
Resource Requested Amount	1
Resource Awarded Amount	*1*

Top of page

Abstract

The BigOnc project is a collaboration between SDSC and the Scripps Translational Science Institute (STSI). The concept of the BigOnc project is to create a Big Data framework that stores massive amounts of data describing the genotype, treatment regimes, and clinical outcomes for large numbers of cancer patients. The goal of BigOnc project is to allow clinicians to submit queries that consist of their patient tumor genomes, and receive information that will help them choose appropriate therapies for their specific patient/tumor. The project requires the ability to store, access, and query big data rapidly, since each uncompressed human genome approaches 1 TB. The unique architecture of Gordon will be extremely useful in evaluating the viability of the BigOnc framework, because the combination of high memory and fast data access via SSDs will allow us to analyze the large amounts of data required in an environment designed for exactly this kind of use case. The startup allocation requested here will allow us to begin benchmarking experiments so we can understand how to organize and query these large volumes of data on the Gordon architecture. Initially experiments will focus on evaluating genetic variations of patients in much smaller files (10 to 20 GB each). These

preliminary experiments will help us benchmark the performance of Gordon for as a platform for this type of large data querying, and provide proof of concept for the BigOnc project. The results will guide us in scaling up to the full production version of BigOnc resource that we envision.

Top of page

Print