THE PATH TO PETASCALE COMPUTING IN GEODYNAMICS

A report by the Science Steering Committee
Computational Infrastructure for Geodynamics (CIG)

December 5, 2006

Members:

Peter Olson, Chair
Johns Hopkins University

Brad Aagaard
United States Geological Survey
Menlo Park, California

Wolfgang Bangerth
Texas A&M University

Omar Ghattas
The University of Texas at Austin

Louise Kellogg
University of California, Davis

Laurent Montesi
Woods Hole Oceanographic Institution

Jeroen Tromp
California Institute of Technology

Shijie Zhong
University of Colorado at Boulder

Introduction

The National Science Foundation is examining the feasibility of a Petascale computing collaboration within the Geosciences Directorate at NSF. This initiative offers the U.S. geodynamics community an opportunity to actively participate in planning and implementing future directions in large-scale scientific computing. As a community we welcome this opportunity, and we look forward to working closely with other disciplines represented by the GEO Directorate in making it a genuine success.

The Computational Infrastructure for Geodynamics is in a unique position in regard to gauging the computer resource needs of our own geodynamics community. In the past year, we have sponsored several workshops on a variety of topics in computational geodynamics. At these workshops the hardware needs of each discipline in our community were discussed in some detail. In addition, we have just completed a broad-based survey of our member institutions, in which the present computing hardware resources and future computing aspirations of over 100 CIG member researchers have been tabulated and analyzed. This report is a summary of our findings. The entire results of the survey are attached as Appendix A.

Computing in Geodynamics – a spectrum of approaches and needs

Research in computationally intensive Geodynamics spans a broad spectrum of styles and needs for cycles. In many areas, cutting-edge research attacks how best to parameterize complicated multi-scale physical and chemical processes. At present, most advances in geophysics are done at a considerably more modest level of computing, with most researchers using a mid-sized cluster. The reason for this is inherent in the nature of geodynamics research. Progress in most areas of solid-earth geophysics still centers on deciphering the complex interactions between many geophysical and geochemical processes. Experience has shown that the most effective computational approach is to explore these phenomena at different spatial and temporal scales, using the widest possible range of parameter values. This style of research necessarily entails many realizations of the same model, to determine parameter sensitivity, test model assumptions, and compare with observations. It will likely remain at the heart of computational geodynamics for the foreseeable future.

In a few areas, the underlying equations are sufficiently well understood and the algorithms sufficiently stable that cutting edge research involves access to the largest computing resources to make possible calculations at sufficient spatial and temporal resolution. The question we address here is how to develop a program that allows simultaneous progress on both types of applications.

The Potential for Petascale Computing in Geodynamics

There is a consensus opinion in the geodynamics community that petascale computing will play an important role in future Earth Science research in general, and for geodynamics in particular. Petascale machines open the possibility of exploring

geodynamic phenomena over a much more extensive range of spatial and temporal scales than can be done at present. For the most part, approaching realistic conditions and parameters in geodynamic computations is linked to broadening the range of spatial and temporal scales that a model simulates, as well as incorporating physical and chemical processes over each of these scales. Already several groups within the CIG and elsewhere in the solid-earth sciences are engaged in preparations for petascale computing applications. Examples of these efforts include: (1) An effort by the CIG working group in mantle convection for CitcomS to be a component of the UC NPAR[1] bid; (2) the efforts of Jeroen Tromp to have his code simulating seismic wave propagation, SPECFEM, be part of several bids for the NSF Track 1 petascale competition; (3) the efforts of SCEC to have TeraShake as a key component of the NPAR bid; and (4) the CMU/UT-Austin seismic wave propagation forward/inverse code.

We recognize that petascale computing has the POTENTIAL to transform geodynamics modeling. However, we also recognize that achieving this lofty goal is a far greater challenge than stating it. For our community, success at petascale computing entails a broader commitment than simply joining a hardware consortium. In spite of the fact that some areas in geodynamics research are prime candidates for first-stage application of petascale computing (see above) and that nearly all of the disciplines represented in CIG could potentially benefit by computing at this scale, to date only a limited number of applications in our field are now ready, or even suitable, for this level of computing. In the next section we summarize what we learned from our members about how to prepare the geodynamics community for the petascale computing challenge, and we outline the additional resources that will be required.

Our survey findings are consistent with the recommendations of the 2005 Cohen report. The Cohen survey findings of the report showed that 78% of research undertaken at the time by earth scientists was on clusters of 64-256 processors, but that there was considerable dissatisfaction with the resources available to researchers at the time (Appendix 1, Cohen Report). The Cohen report's conclusions called for three tiers of computing power, a large facility for leading edge earth science research; ten regional clusters to serve mid-range computing users; and increased funding for small to medium clusters at the research group/department level, which would serve developing codes, as well as education/training of future computational geophysicists[2].

The situation has not changed significantly in the past two years. Only a very few researchers within the CIG community have an urgent need for petascale computing at the present time. The vast majority of researchers in our community needs access to well-supported, readily available computers with about 100 processors, which will likely grow to an order of 1000 processors over the next 5 years. Increased access to, and additional

[1] The National Petascale Applications Resource (NPAR)

[2] Cohen, R.E., ed., 2005. From overview section of *High-Performance Computing Requirements for the Computational Solid Earth Sciences.* 94pp, http://www.geo-prose.com/computational_SES.html

funding for, these smaller machines are absolute prerequisites for successful future use of the petascale computer by our community. The access formula to computing on this scale should be comparable to the machine accessibility within the researchers' home institutions.

Even more daunting are the modeling, algorithmic, and software developments and the community training needs that will be required to make use of a petascale computer. There will be a need to support activities (like CIG, and probably other activities) that help provide researchers with the proper tools to move up from one scale of computing to the next. This is especially important at the expected bottlenecks where performance scaling up becomes ever more of a problem.

Summary of the Survey

From Oct. 26 to Nov. 7, 2006, CIG conducted an online survey of its members. During this time, we received 114 responses to the multiple-choice questions with more than a third of respondents offering additional written comments (see Appendix A).

Overall, the community voiced a fair degree of wariness about the concept of a petascale resource dedicated to very specialized geosciences research, as an addition to the facilities that NSF has committed to provide in all research areas through its Track 1/2 programs that are accessible to all geodynamicists.

Rather, an overwhelming majority of the respondents wanted assurance that computing resources available to them would be increased substantially, with that increase more squarely centered on computing platforms with several teraflops of performance. The respondents also believe that there must be an appropriately balanced investment in system size, ranging from smaller clusters to larger systems, and also an appropriate balance between hardware and software funding. These resources are seen as crucial to develop the models, algorithms, and applications software that will make petascale machines useable to our community.

Themes that were repeated in the survey included:

* Hardware must be complemented by software. Usability of the computational facility will be much improved if a suite of well-tested and optimized software is available on it. We need more people who know how to develop and build efficient parallel software that would be suitable for petascale computing.

* Well-supported, efficient, widely available access to capacity computing will result in the greatest scientific advancement in geodynamics.

* The community needs substantially more access to tools for constructing large, complex models.

* Many currently available parallel geodynamics codes will not scale well to the petascale.

* Researchers believe that national computing centers that cater to capacity computing are less than successful due to long queues.

* There is not yet enough support at computing centers for porting and maintaining codes.

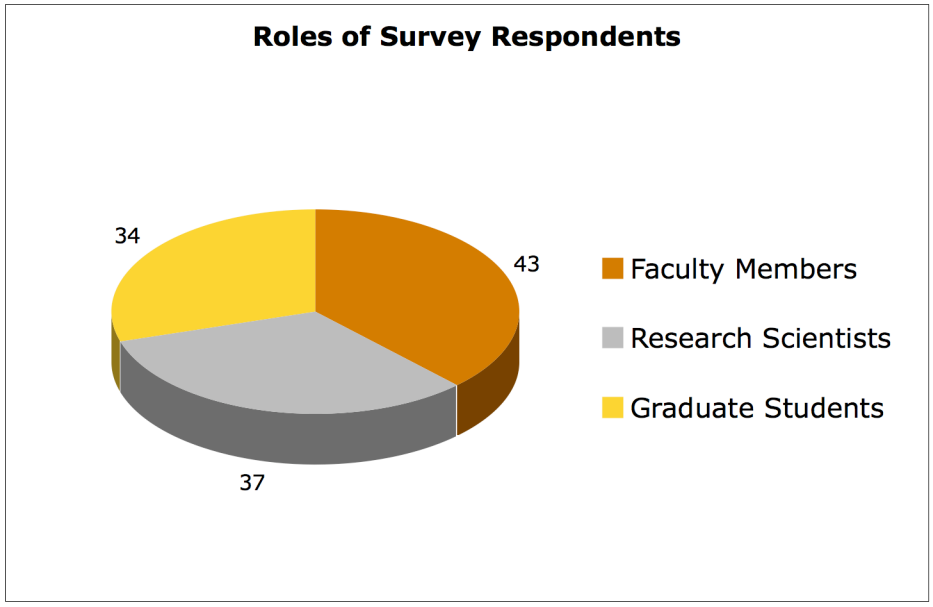
* Experiences within the community suggest that department and university operated computing remains an essential element in the hardware and software hierarchy.

* Because much research focuses on model development, concept testing, and trying different approaches for input and solvers, smaller machines with very short queues are optimal.

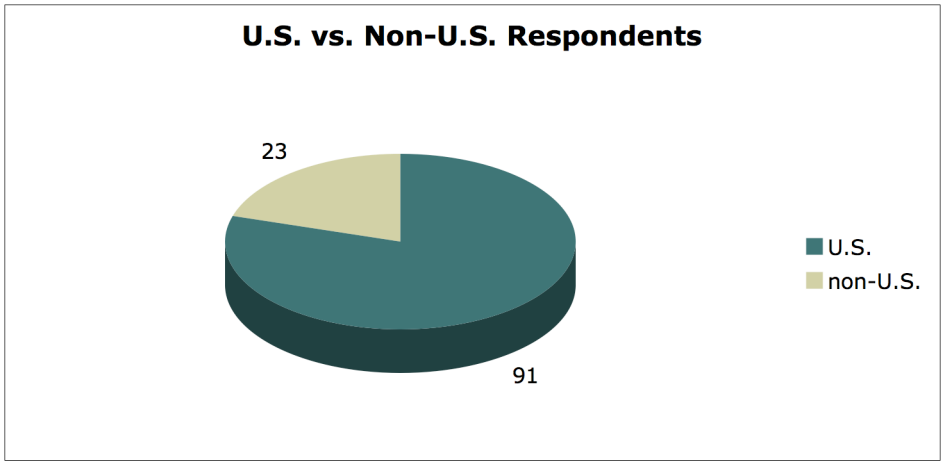
Appendix A:
CIG Computing Needs Survey Report

The Computational Infrastructure for Geodynamics (CIG) Science Steering Committee recently asked the CIG community at large to participate in a survey on the computational needs of the community so as to provide feedback to the National Science Foundation on plans to develop a national geosciences facility that will have sustained petaflops performance. As of November 6, 2006, CIG received 114 responses.

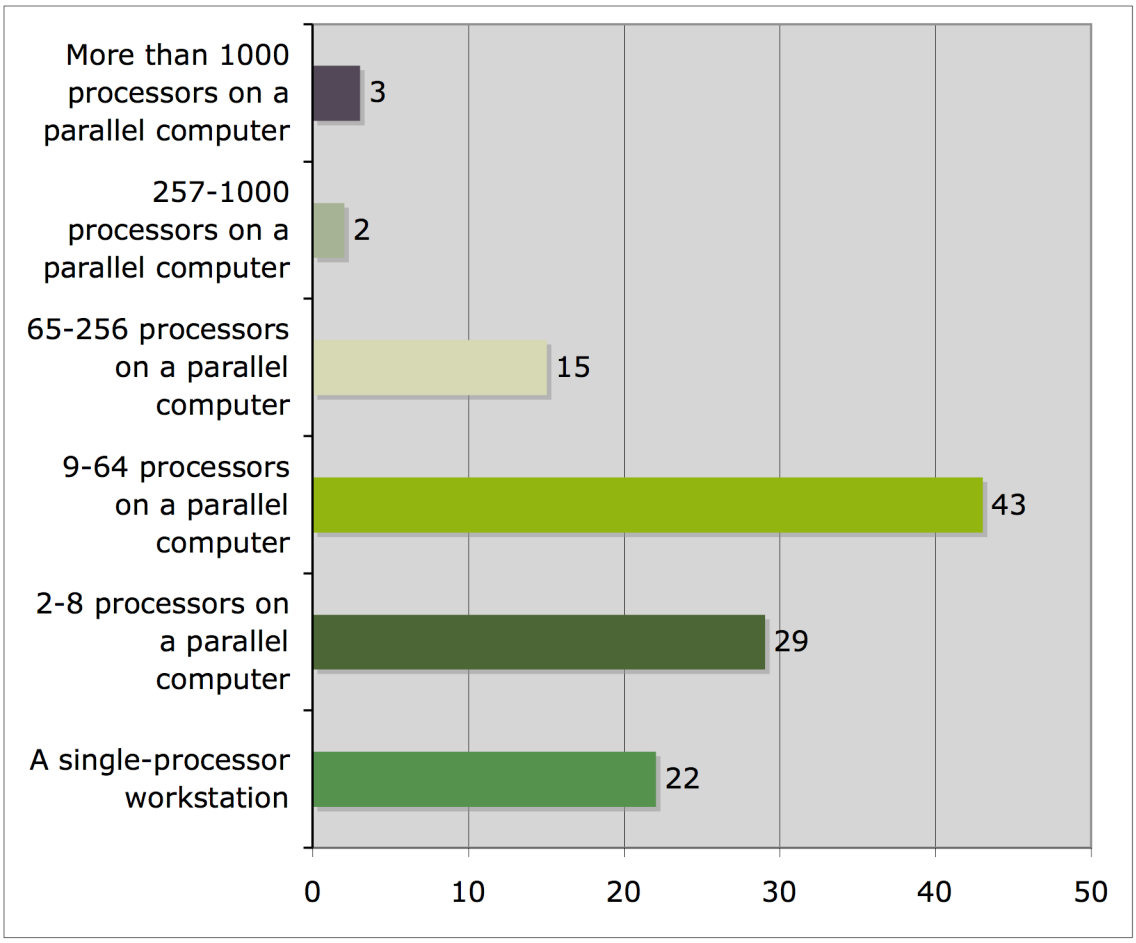
Here are the results of the survey, some further analysis of some answers received, and additional comments submitted by respondents.



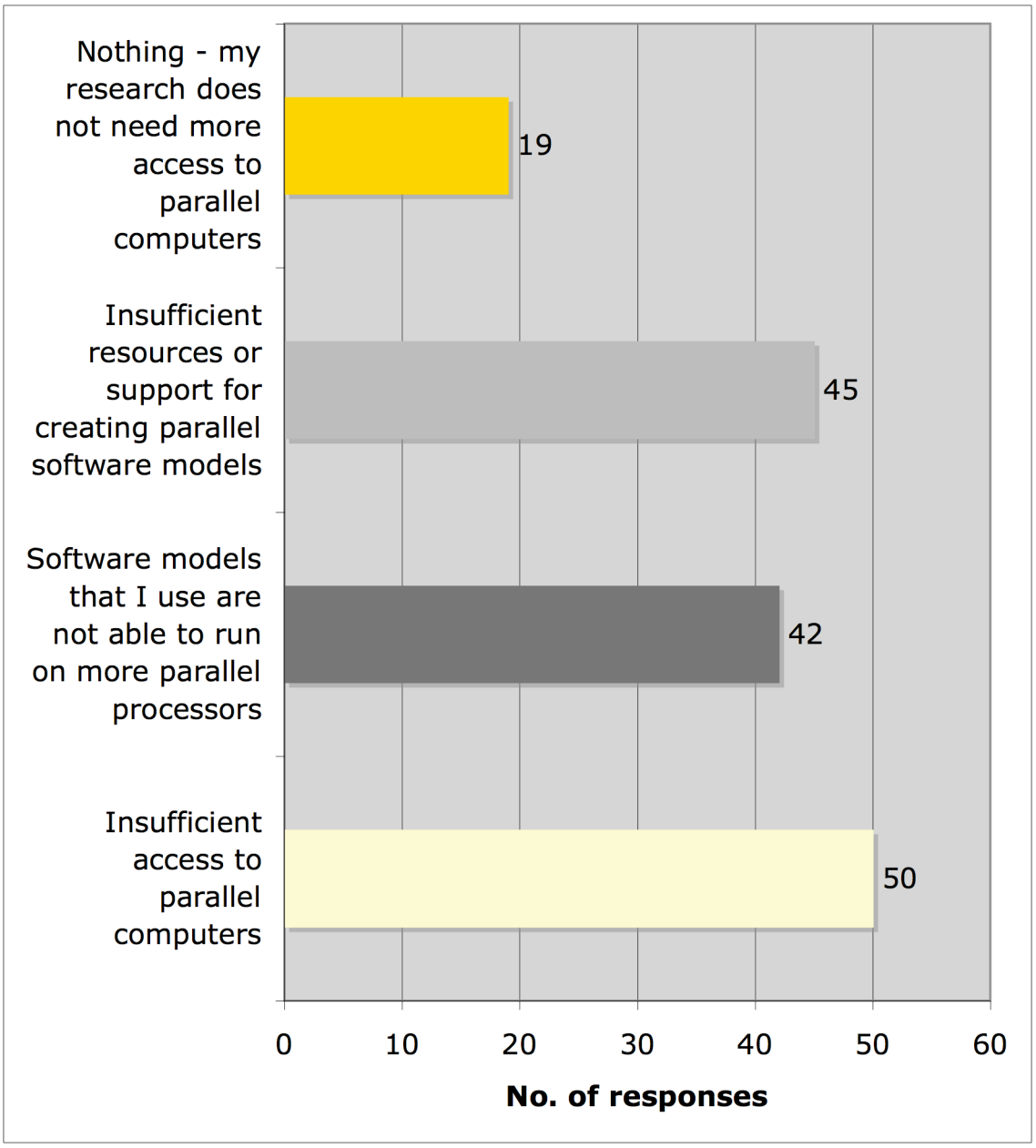
Note: No undergraduate students were among the survey respondents.



1. What is the largest number of processors that you regularly use for your current research?

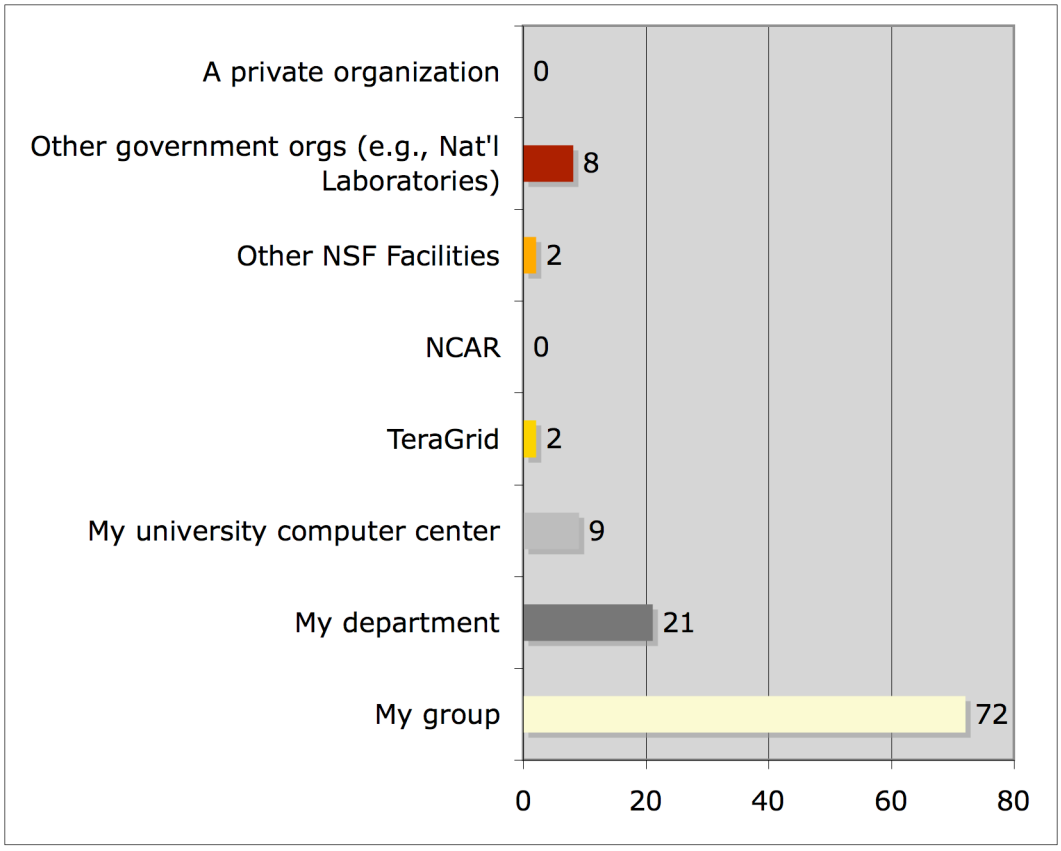


2. What limits your use of parallel processors?

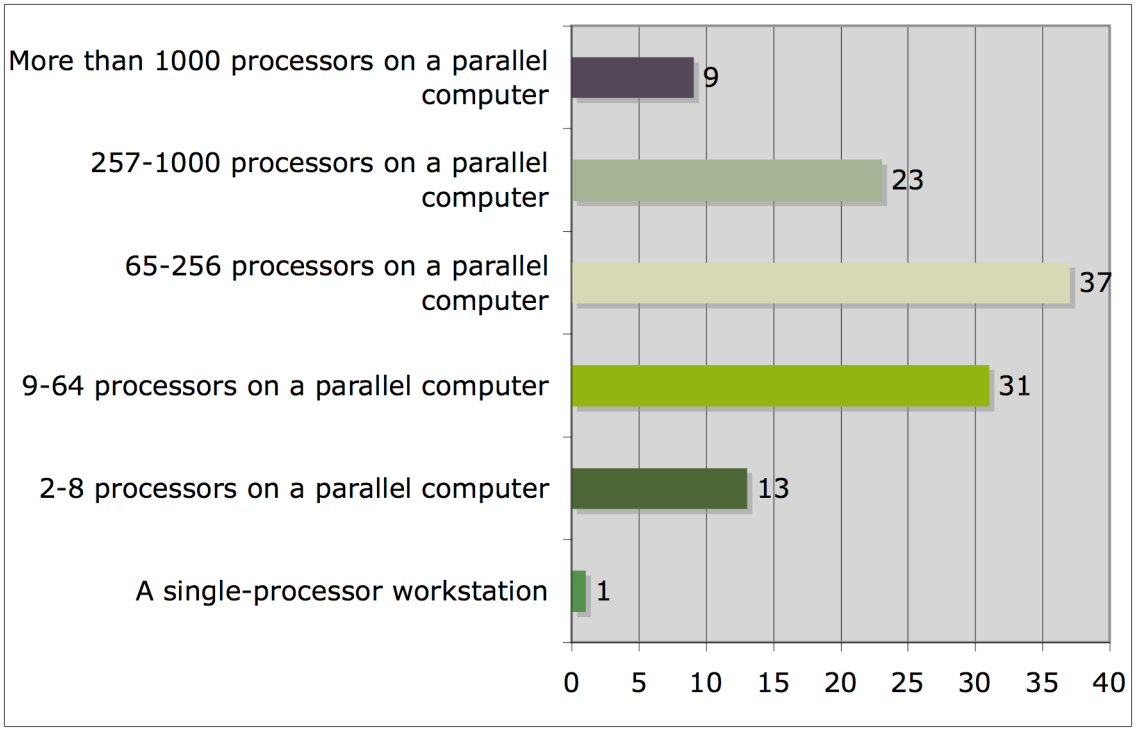


Note: Respondents could choose more than one limiting factor.

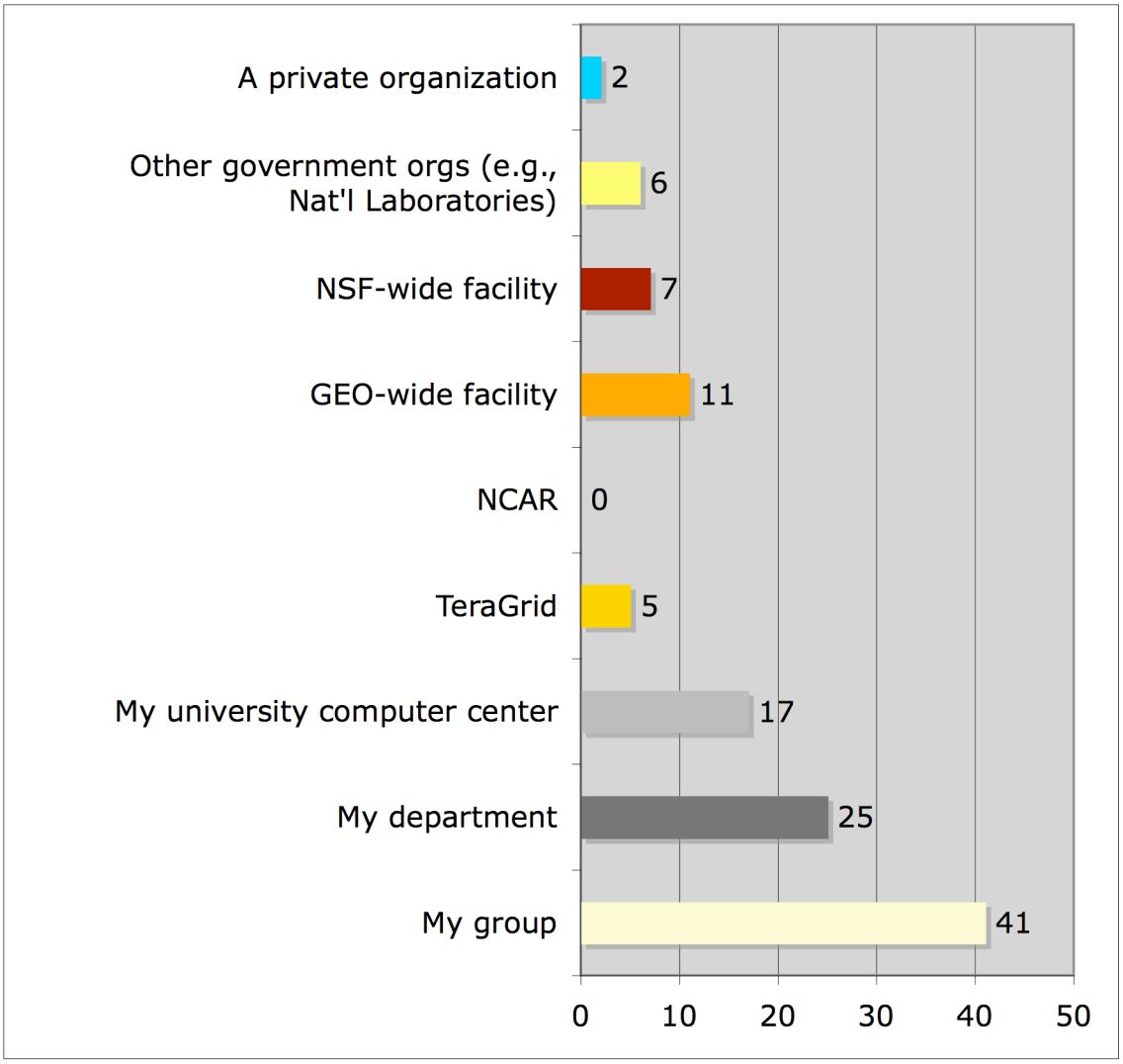
3. For the computing facilities in question 1, who owns/operates them?



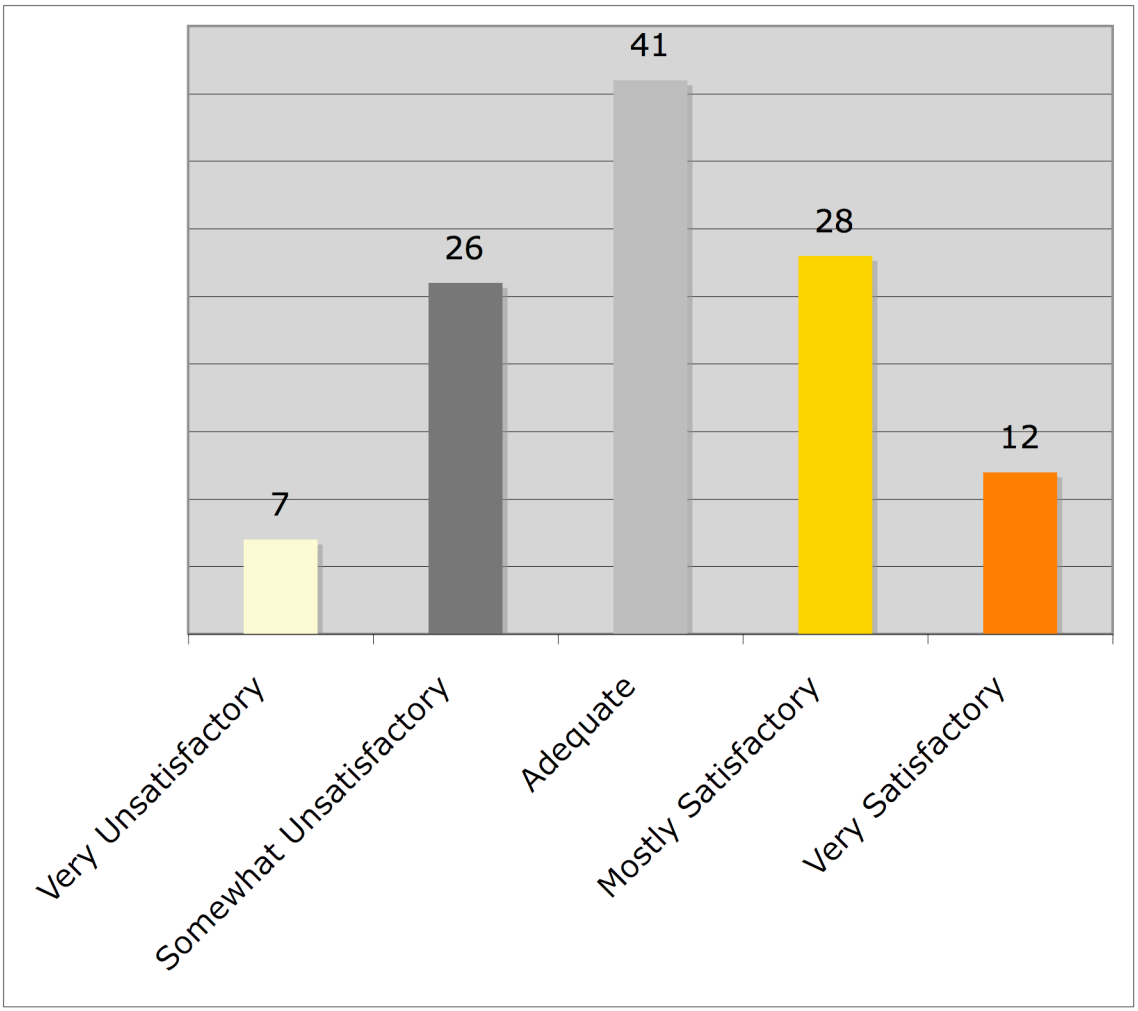
4. For the next 5 years, please estimate/project the type of computer systems you think that you will need for your research:



5. For the computing facilities in question 4, who should operate them?



6. How would you characterize how well your current resources satisfy your scientific needs?

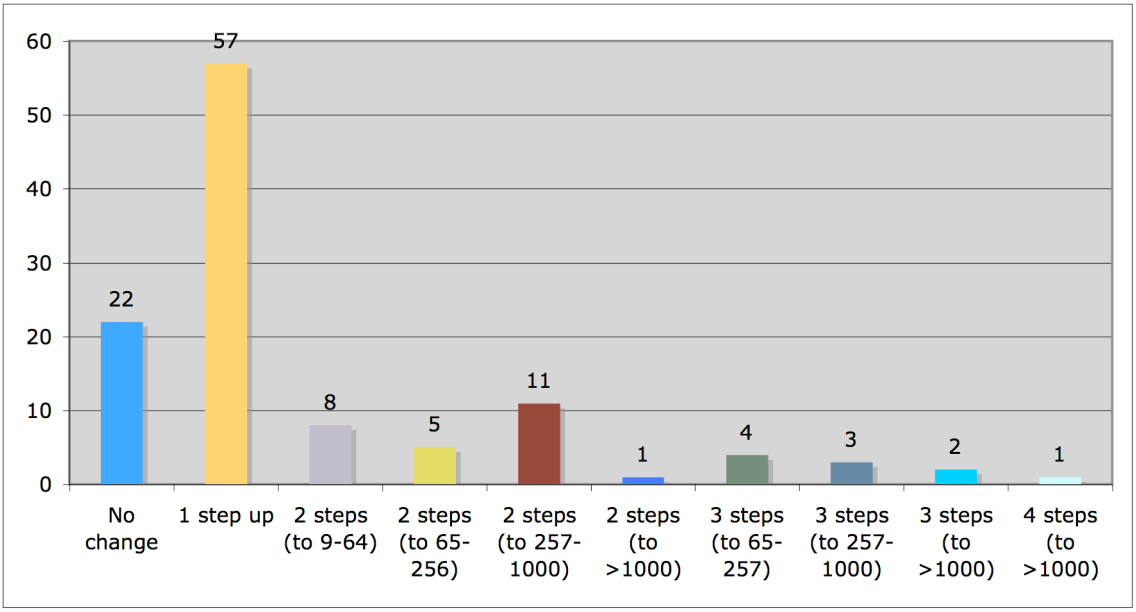


Response Comparisons

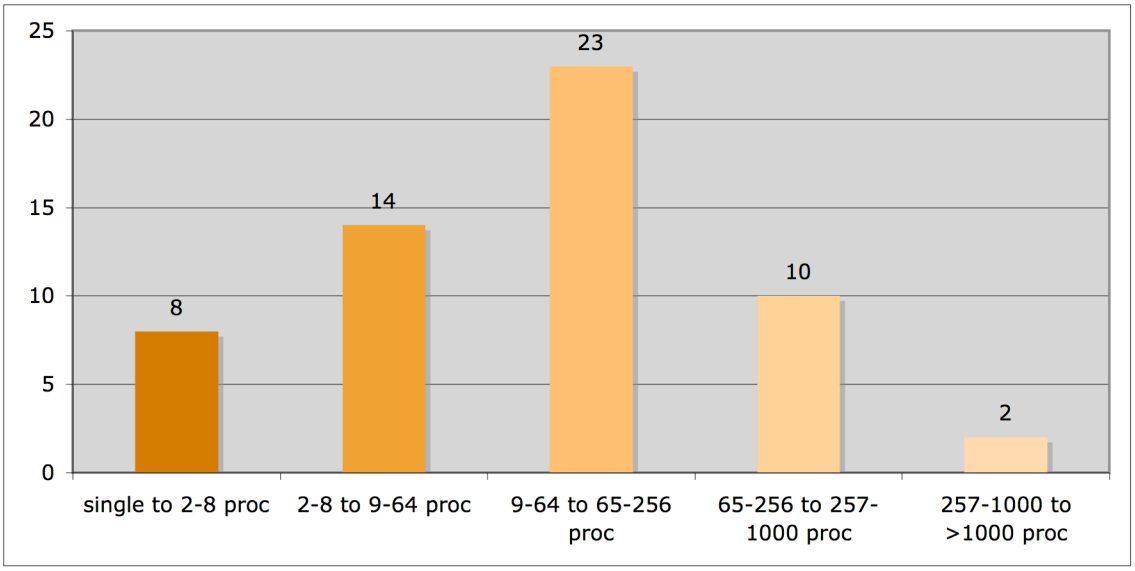
Questions 4 and 5: A comparison between the number of processors respondents desired in the next 5 years and their opinion as to who should own/operate the facility offering that computing power:

	single proc	2-8 proc	9-64 proc	65-256 proc	257-1000 proc	>1000 proc
My Group	1	8	16	11	5	
My Department		1	7	7	6	2
My University Computer Center		2	3	10	4	
TeraGrid					3	2
Other Gov't Orgs		1	1	2	1	1
NSF-wide		1	2	3	1	
GEO-wide			2	4	2	3
Private Org					1	1

Questions 1 and 4: A comparison of the number of levels of computing power that respondents expected to step up to in 5 years versus their current levels. (**Note:** Three that answered "No change" were already at >1000 level.)



Of the 57 who anticipated only a one-step increase in computing power needs, the breakdown of which particular increase would be needed in 5 years is as follows:


25
20
15
10
5
0
8
14
23
10
2
single to 2-8 proc
2-8 to 9-64 proc
9-64 to 65-256 proc
65-256 to 257-1000 proc
257-1000 to >1000 proc

Additional Comments Received

While I see the usefulness of large computational hardware, I feel I cannot judge its usefulness until I make the jump to a limited number of processors. This jump is particularly difficult in our group at WHOI because of its limited size (currently no student or postdoc) and no one here has experience in parallel computing. Support on how to use this hardware is important.

Exciting idea!

A bigger hammer is not the most efficient way to solve our problems. In my opinion, the mantle convection community is more limited by the extreme inefficiency of its software than lack of hardware capabilities. We could, in principle, have software that runs at least an order of magnitude faster than current codes allow. Thus throwing money at machinery is a very crude way to solve our current problems. Rather, money needs to continue to flow toward projects requiring expert personnel to construct more efficient software, such as the Computational Infrastructure for Geodynamics (CIG).

I am mostly looking to increase my local capabilities, by growing our local computer clusters. We have trouble attracting funds to keep even a rather modest computer facility working and relatively up to date.

Peta-Scale computing at remote super-computer sites is over-kill for us, and is also not particularly efficient, as it would necessitate moving huge amounts of data around -- which our university's network is not well suited for. We do a lot of computationally intensive data processing, followed closely by inversion and use of our own facility is most efficient. These needs are driven largely by the fact that I'm a seismologist rather than a geodynamicist.

Some larger scale computing capabilities for forward simulation are needed, too. In this case, use of a remote super-computing facility is more feasible, but it's more at the Tera-Scale than the Peta-Scale.

I fear that a Peta-Scale facility, built and financed to serve a tiny fraction of the geoscience community, will gut small grants for local computing -- which, in effect, will make our science even harder, or perhaps even impossible, to complete. This must also be understood in the context of university retrenchment in matching funds -- which increases reliance on federal dollars to support computing. It's a delusion to believe that the answer will come in the form of a giant government facility or that once the facility is built greater funds will shake free for local facilities. Rather, what's needed is greater investment in distributed computing at the group or departmental level.

The current resources (assuming reasonably regular upgrades) are sufficient for my *current* research, which involves 2D and 3D regional models. I have resisted moving to 3D

global models thus far because of the lack of capacity; this is something that would most likely change if computing on a much larger scale would become available to my group.

I would certainly prefer someone else to take care of operating it; but anything on a national level may become too administratively cumbersome (like NPACI) to be attractive for anything but the most desperate cases.

For many of the problems I have worked on, desktop workstations have become sufficiently fast that there is little incentive to spend the effort to parallelize the codes. More recent problems involving FEM codes will clearly require parallel processing. A local 16-node cluster already available to me may meet my needs for the immediate future. However, the problem of parallelizing code remains an energy barrier to optimal use of this cluster. (I commend CIG's efforts to offer parallelized standard codes as tools for researchers!) At some point, as this project matures, easy access to remote HPC facilities will most likely be needed.

Currently have 16 dual processors with primary access only for my group. Maintenance and access shared with a larger group of roughly 100 processors with Physics and Math department in my university. I can think of wave propagation problems that would require on the order of 1000 processors for high freqs and long range to finish a problem in 1 day or less. I do not want to have to write additional proposals to have access to a facility. I thought the NPACI model did not work well in that it required separate proposals reviewed by non geophysical specialists.

The idea of a community facility with tightly controlled access and time allocations should be contrasted with the idea of distributed clusters having some common hardware and software standards. It might be better to spend money via NSF-IF toward individual investigators to build and enhance their own clusters, which can occasionally be linked easily for collaborative work enhanced by some common standards.

The pervasive presence of multicore processors over the coming years will be important for considerations of future distributed infrastructure. They will strengthen local vs. centralised facilities. A petascale facility would allow us to arrive at the next level of earth models (for example, seismic or tectonic).

There is value in having NSF build a large GEO computing facility. But I am afraid that its use will become dominated by a small core group of researchers, able to fly large proposals for intensive use through the review process. This is the situation with the IRIS/PASSCAL facility, for example. I suggest that NSF instead put such funds into the regular grants programs, so individual researchers can write small proposals to buy, say, 10,000 CPU-hours on a commercial facility like SunGrid. It's possible that the current $1 per CPU-hour cost of SunGrid will fall once a competitive marketplace for parallel computation is established, over the next few years.

In addition to the number of processors the amount of memory is an extremely important parameter that affects programming models. In some cases months worth of programming time can be saved by using OpenMP on shared memory machines with a limited number of processors (<=64). For codes that won't always be run in production mode, development time is often more important than the run time. In this case I think it also important to have access to large memory machines where more accessible programming techniques might be applied. Of course this is not to the exclusion of distributed memory codes. Of course these two may merge as Intel now offers a compiler with a so-called "cluster" version of OpenMP. Thanks for your time.

In my opinion, increase in computational resources is not an immediate condition for significant scientific progress in computational geosciences.

I have serious concerns about NSF funding a national facility for geoscience computing. My reasoning for this is that it will most likely impact the ability of individual PIs to acquire NSF funds to support local Beowulf clusters that satisfy all to most of their computing needs.

My system is fairly new, so I am currently pretty happy with my setup. However, in the ~3-5 year timeframe, my computational needs will increase and my cluster will be starting to age. At that point it would be nice to have some community-wide facilities available.

Money has been so tight in recent years that it has been difficult to upgrade our facilities.

We need access to big clusters. Long queue wait times are the biggest problems we face with regards to use of shared computing clusters.

I anticipate that my computational needs will increase significantly over the next few years, and my estimate for question 4 may be low.

In my answer to question 5, I think it would be fine to use a facility operated by someone like NSF, as long as there are not significant administrative hurdles to obtain time, and as long as the queue times are not excessive. Otherwise, something more local would be preferred.

Some queue systems are not very useful for long-duration geodynamical simulations -- it's a bit tricky to satisfy different needs, i.e. long-term several nodes vs. short-term but vastly parallel. That’s where group operated resources are great. However, computational

resources are always welcome.

Can folks outside the U.S. apply to use CIG resources in a straightforward way?

A notable point is that in the future I'll be using a wide range of programs for modeling, and it is very foreseeable that I'll need access to a teraflops-type computing center.

At this stage, we are working on small sizes of models to understand basic scientific questions such as effects of nonelastic off-fault response on earthquake dynamic rupture and ground motion. In the next few years, we are planning to incorporate as much physics as possible into models to perform large-scale simulations for ground motion prediction at high frequencies, which require petascale computation ability.

So far the UC's San Diego Supercomputing Facility has given us adequate amounts of computing time on up to 32 processors. The OMP code (J. Wicht's Magic2.0) we are using does not scale beyond 32 processors on their main shared memory machine. When this code becomes dual OMP/MPI capable, then we will definitely run it on more than 32 processors.

It is very tough for a PI at a state university to stay scientifically competitive. I have projects that require significantly more computer power than I have at my disposal, and this is the greatest barrier to doing some really exciting research; other major needs are better access to modern software and a need for trained and capable graduate students.

Our department suffers from an antiquated and heterogeneous computing environment. A departmental computational facility would not only increase our computational abilities and but also stimulate interaction between groups. There is also a significant benefit in department-wide system in terms of computer support.

For your information: In Australia we face difficulties with obtaining discipline dedicated (locally run) computing resources such as that proposed by NSF. Perceptions of university managers and funding agencies are that funds should be dedicated toward centralized and often remote multi-disciplinary facilities. The perception is that this is more efficient. In truth I think the opposite is true. The result is that they are dominated by `institutionalized' users like computational chemists. Remoteness also severely inhibits (non expert) new users, who often have problems that could take advantage of such resources, e.g., large data processing, but view such `toys' as the domain of the computational scientist arm of their discipline (intent largely on numerical simulation of natural phenomena). An `activation energy' toward utilization is created that many do not overcome. Operations like CIG make such resources more visible and accessible by all and

I am very supportive. Keeping in touch with the community is a vital aspect.

I would strongly support the recommendations made at the Petascale workshop that our community certainly does have important problems that would benefit from such a machine. However, to put one large machine to good use requires an infrastructure of smaller multiprocessor machines on which investigators can develop code and run simulations to justify working at the Petascale. Petascale applications would also benefit from user support in various forms.

Scheduling a mix of job sizes that permits balanced and fair access to a large facility seems to be one of the biggest problems when parallel computing is managed by a large central facility. In particular it becomes very difficult to get jobs that require large numbers of processors out of the queue and into execution.

Fast interconnect is also a critical issue for most of my work: essential.

Information on currently community-accessible computers would be most welcome

For the past decade, NSF HPC systems significantly lagged those available to DOE researchers (and academic researchers in some other countries). This is being rectified in a major way with NSF/OCI's deployment of Track 1 and Track 2 systems over the next 5 years, allocatable through the TeraGrid.

The attention now shifts to the ***computational science*** that these systems will support. To fully realize the promise of these new supercomputers, NSF needs to make parallel investments in research to advance the models, discretizations, algorithms, applications software, etc., underlying computational geodynamics. I think it would be a mistake to divert GEO funds that would ordinarily support this sort of work to finance additional hardware procurements. The hardware is only as good as the applications software that runs on it, and for now the gap between the capacity and capability of forthcoming parallel systems, and the state of modeling and simulation codes, is ever-widening and has to be addressed. This would be the best use of GEO divisional budgets, in my opinion.

I fail to see the value in a large, expensive computational facility that will no doubt be monopolized by just a few groups. This will simply drain funds from worthy scientific projects, and for what purpose? This sounds like a technology driven rather than a science driven initiative. If people in the geoscience community are going to push this forward, they better make sure that it is driven by science and is not just another way to concentrate more resources into the hands of a few at the expense of the many.

We need more resources for supporting the tools we created under an ITR award. As of

now the support is essentially pro-bono.

I think we need a balance between large institutional facilities, department facilities, and individual facilities. My group can do much of its work on our desktops, and on our 8 and 24 processor SUN SMPs, but we have larger needs as well. Right now our biggest hurdle is transitioning our software to clusters.

Although working in Germany the needs of our group (we use the TERRA-Code written by John Baumgardner and parallelized by Peter Bunge) may be interesting to you. We are a small group of scientific staff who implements geophysical and geochemical (that means integrated) models of Mars and Earth into the existing code while 2 Ph.D. students and 1 Post-Doc are working on numerical refinements. For production runs the national supercomputing centres are a good choice although they lack instantaneous access which is indeed a disadvantage. For numerical research a small cluster with e.g. 16 procs would be useful to observe runtime and scaling behavior of the code and of the changes we made.

The largest challenge I face is installing/maintaining underlying software packages. This is due to a lack of personnel at my institution to provide infrastructure support for computing systems. For example, even making use of the CIG codes requires that I or my graduate students install and maintain the underlying software packages (Pyre, etc.). Although this is manageable, it does take time.

Establishing such support at my institution would be an inefficient use of resources. A better model would be for me to have easy regular access to a system on which such software is maintained, allowing me to make use of the base level software when developing my own codes, and make use of codes developed by CIG and others without the need to install/maintain them on my home machines. This would be an excellent model for an NSF-supported geosciences computer facility.

The barriers to progress on facilities like the TeraGrid are many.

1. The queues get way way too long. You can wait for days to get a large number of processors unless you get someone to tweak the priority on the queues. That isn't right either and happens, one suspects, too often causing average-Joe jobs to sit forever.

2. The software issues are very complex. I've had problems porting codes I run due to dependence on software packages that cause problems. A multifaceted problem I know is well known to groups like TeraGrid, but still a concern.

I'm in charge of maintaining the 17-node cluster of my group. The overhead is too much. While I admit that through maintaining the cluster one can learn a lot, I think it's best to have the university (or at least the department) operate the clusters in order to give us

more time to focus on the research rather than worrying about why some nodes are dead. Dedicated technicians could do much better administration work of both hardware and software.

P.S. 2006 CIG meeting in St. Louis was very helpful to me. Thanks a lot for all the effort!

The capabilities of single processor workstations have kept pace with my computational requirements. I build both forward and inverse analyses around 3D FEMs for various quasi-static processes.

Sorry, but I don't know if I am necessary in this type of report. I'm from Brazil. However, we can think about a base in Brazil, to improve the resolution of the CIG for region of the South America.

As a student in Jeroen Tromp's group, my computational needs are met well by CITerra, the cluster in GPS at Caltech.