INDIANA UNIVERSITY

16c – Optimizing Research Support

Action Item Template Response

General Action Item Information

Lead Division/Office: Research Technologies
Action Item Number: 16c
Action Item Short Name: Optimizing Research Support
Dependencies with other EP Action Items: 3, 4
Implementation leader (name & email): Craig Stewart (stewart@indiana.edu)

I. DESCRIBE YOUR PLANS FOR IMPLEMENTING THIS ACTION.

Indiana University has a long and proud tradition of open access to its research computing facilities, going back to policies set in the 1950s by the first director appointed to head up research computing at IU - Marshall Wrubel. We have historically avoided "nuisance fees" and have provided the best possible services to all who asked. We have also worked aggressively with IU faculty to bring grant funding for hardware and staff resources to IU. As a result, we have been able to expand our cyberinfrastructure resources (hardware and personnel) more rapidly than many of our peers and competitors. IU has a particularly large and diverse group of users of its advanced cyberinfrastructure and the university's research and creative activities have been broadly enhanced by its investment in cyberinfrastructure. However, one additional result of this situation is that UITS faces more demand for computing resources and consulting and programming staff than it can possibly meet, and the current economic situation makes it impossible to expand resources to meet demand. Furthermore, our consistent history is that whenever we expand resources this spurs even greater demand.

UITS is at present each year producing millions of dollars worth of CPU time on its HPC systems that are consumed by the IU community (and under contract, users outside IU). UITS is also managing the delivery of millions of dollars of consulting and programming staff time per year.

In the past we have taken an approach to management of access to computational resources based on use of so-called "fair share" algorithms, in which a researcher's priority in a high performance system (HPC) queue is decreased in proportion to the amount of resource used in some proceeding time period (typically a month). This approach does not work in practice with a cyberinfrastructure as large and complex as IU's current systems, particularly given the very sophisticated users we have who are now making use of Big Red and Quarry.

As regards consulting and programming requests, we have subdivided consulting into two types of interactions - short-term consultations and long-term consultations. Short-term consultations are just what they sound like - quick questions and answers, typically taking less than 4 hours of staff time. Long-term consultations take longer than 4 hours of staff time. UITS has typically offered programming services of up to one person month of time for free, and when good cases can be made for the value of the work being done, up to 6 person months of consulting/programming time. As demand has exceeded supply, we have used a simple FIFO algorithm for requests of more than 1 month of staff time. This seems fair at first blush. However, what has happened in practice is that there have been two suboptimal side effects of this approach

because the demand so greatly exceeds supply. One is that people who have work that is closely aligned with university priorities get frustrated in waiting and abandon important projects. Another outcome is that at times UITS staff are at work supporting activities of relatively great distance from university priorities than work of obviously high priority that sits in the queue waiting to be attended to.

At the same time, there are people within the university community who are doing highly meritorious research, who could make good use of additional resources from UITS, who do not know to ask for those resources because we do not have a well-advertised process for lodging resource requests.

A side effect of this situation is that it appears to the members of the IU community to be easier to wait in the queue for IU resources than to use resources available outside IU (for example, resources via the TeraGrid) because there is a fairly formal and involved allocation request process. This process was in past years badly administered and some allocation decisions were made capriciously. As a result, within the IU community, there is somewhat of a negative impression of the TeraGrid as compared to IU resources. IU has led efforts within the TeraGrid to make this process easier for people requesting resources and better for the scholarly community in general.

UITS supports all scholarly and creative work done by the IU community. To each member of the community, their work is the most important. That's why they do it. But when demand for resources exceeds supply, there are two options. One is that we can fail to put in place practices that aim to optimize use of resources. The other is that we can put in place practices designed to optimize use of resources. The Office of the Vice Provost for Research established a precedent for lightweight peer review processes (internal to IU) for allocating funds in the Faculty Research Support Program. This model has now been used in allocation of METACyt funds and some PTI funds.

We recommend the implementation, carefully and over time, of an allocation process for distribution of large amounts of computational resources and consulting/programming resources. At present there is not a need to implement this sort of approach with storage, although we want to set the policy basis for doing that when needed. We recognize three key factors in dealing with resources to support research and creative activities at IU:

- Current demand exceeds local supply for extended consultations (interactions that often involve weeks of effort by a UITS professional staff member) and supercomputer resources. At the same time, lack of a well advertised process prevents some researchers from knowing that they can and should ask for additional resources.

- Some sort of "first in, first out" algorithm is not in keeping with the general traditions of resource allocation with academia; rather faculty peer review is the time-tested method for allocation of limited resources.

Adoption of an allocation process, based on IU faculty peer review, for staff time in extended consultations will ensure that in any given year, the projects that get the attention of staff resources or very large amounts of supercomputer time are those recognized as most meritorious by a review panel of IU faculty. Adoption of an allocation process will result in UITS being more responsive, in any given year, to highly meritorious research work done by the IU community. The allocation process would also serve as a way to predict demand so that we can look to resources within and outside IU in order to meet demand. Among other things, it will serve as a way to help UITS and faculty members each spring discuss needs anticipated for the following year - early enough that it becomes possible to prepare grant proposals of a variety of types to bring additional resources to bear to aid IU faculty. This process may also result in people also being told directly "No, there are not resources available within the university to support this project." If that happens, based on peer review, that seems better than seeming to say yes but having the yes be "yes, but only three years from now."

The proposed process for implementation is described below.

- Early in spring semester each year, UITS will announce an internal proposal process similar to the existing process used by the Office of the Vice Provost for the Faculty Research Support Program.

- UITS will establish allocation thresholds above which a formal request for an allocation will be required. These thresholds are as follows:

    - 500,000 CPU (wall-clock) hours per year
    - One person month per year in extended consulting/programming effort
    - 50 TB in disk storage
    - 100 TB of tape storage per year.

- Allocation thresholds will be adjusted periodically. The default allocation levels indicated above are such that fewer than 20 IU faculty members would exceed them during the current (08/09) FY. We are thus trying to influence the source of delivery of a large fraction of our total resource use, generated by a small fraction of IU researchers.

- Researchers who wish to use more than the default allocations listed above will be asked to submit a proposal requesting those resources. The proposal process will be lightweight - a 2-page proposal and standard NSF-format minibios for the PI and any Co-PIs identified. The deadlines on these proposals would be set for some time between 2 and 4 weeks past the deadline for support of FRSP proposals.

UITS staff will work with a committee of leading faculty to review proposals with one of the following outcomes:

- A request is viewed as highly meritorious, and can be fulfilled using IU resources

- A request is viewed as highly meritorious, but is beyond what can be fulfilled using IU resources alone. In such cases part, but not the entire request, will be fulfilled with IU resources. In addition the review committee will provide a recommendation that the PI work with UITS staff to submit a request for resources to the TeraGrid, Open Science Grid, DOE InCite program, Amazon Web Services Research Grants, or other national resource.

- A request is viewed as other than highly meritorious, and resources in excess of the default allocations will not be awarded. In this case faculty will be invited to apply for resources on TeraGrid or Open Science Grid and will be supported in such efforts by UITS staff. In the case of supercomputing time, however, the one thing we never want to do is have our supercomputers sit idle. So rather than saying "no" per se, researchers in this category would still be able to run on UITS systems if they wished, but only at very low priority relative to other researchers.

Recognizing that sometimes emergencies or particularly high priorities arise on an unpredictable basis, OVPIT and UITS would retain the ability to allocate CPU time (or priority access to queues of all sorts) to important projects that arise on short notice, but under an MOU via which faculty members commit to requesting allocations on the TeraGrid or other facilities.

Likewise, recognizing that new faculty come to IU each summer, and that unanticipated needs arise, we would anticipate allocating less than the total amount of resources (consulting and systems) available each year, so that we retain the flexibility to react to needs as they arise. UITS will provide assistance and advice in preparation of proposals to this internal process. The goal is to have proposals be of high quality, not simply to use the process as a way to thin out demand.

At first blush, this proposal seems to be at odds with the goal stated in *Empowering People* of creating a philosophy of abundance. However, carefully implemented, this process will have the impact of increasing the net resources available to the IU community because it will identify needs early enough in the annual cycle that it becomes possible to apply for and receive competitively awarded resources available from outside IU.

II. WHAT ARE THE POLICY AND PRACTICE IMPLICATIONS OF YOUR PLANS?

I. This is a significant change in practice, in that the proposed actions deal head on and openly with the gap between resource availability and ability of the university community to consume those resources. As pointed out above, usage is metered already, more or less through the outcomes of persistence of graduate students in a process somewhat akin to natural selection. A key point here is that the change in practice is designed to provide better support overall for IU researchers and scholars and to make best use of the resources available locally and nationally.

II. III. IDENTIFY STAKEHOLDERS.

- Research Technologies, OVPIT, PTI
- VP For Research, ORA
- All IU researchers