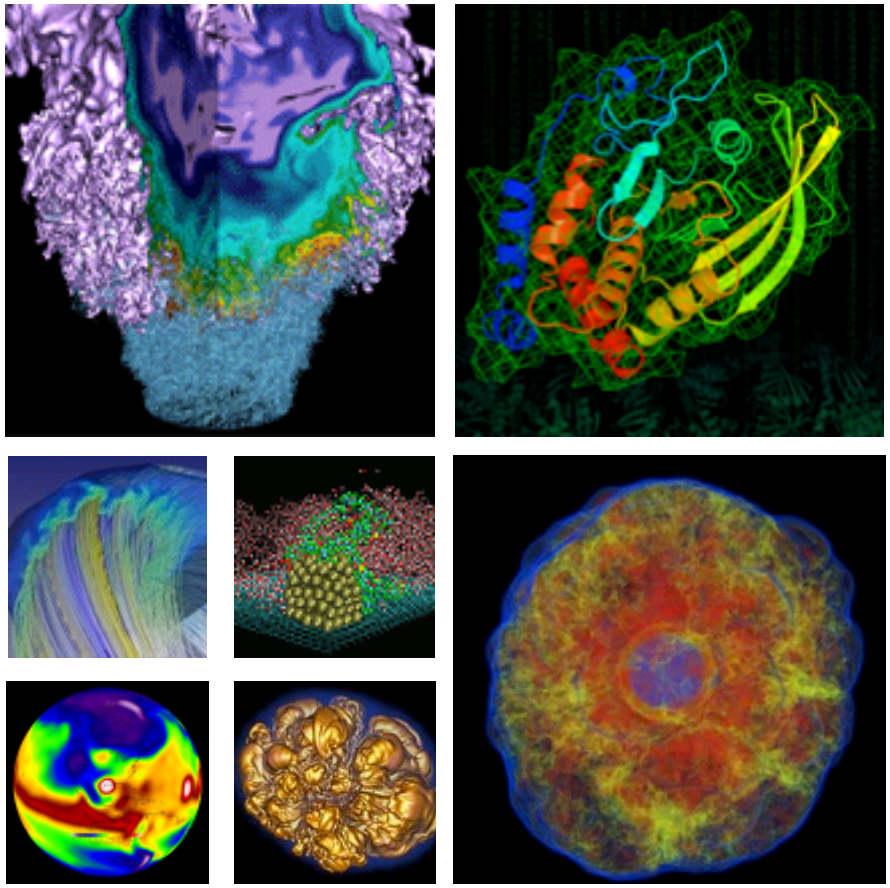
NUG Teleconference May 2013



Richard Gerber
NERSC User Services

May 2, 2013

Connection Info

Connection Info
Topic: NUG Web Conference
Date and Time:
Thursday, May 2, 2013 11:00 am, Pacific Daylight Time (San Francisco, GMT-07:00)
Event number: 664 294 540
Event password: edison

https://nersc-training.webex.com/ and chose from the list of events.

Teleconference information

1-866-740-1260
PIN: 4866820

Agenda

- **NUG Committees**
- **Report on Edison Fair Share**
- **Reserved Nodes for Interactive & Debug**
- **"Perftools lite" performance tool**
- **Math Library Performance on Edison**

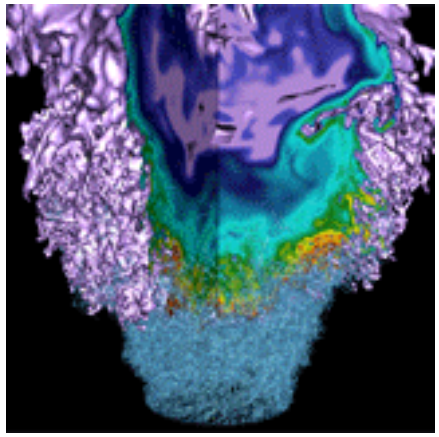
NERSC Brown Bag Seminar Broadcast at Noon Today
The Materials Project: Combining density functional theory calculations with supercomputing centers for new materials discovery

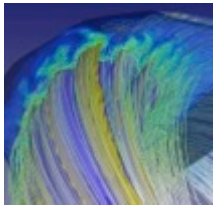
Same teleconference number and PIN as this meeting.
New WebEx connection needed, password: **science**
https://nersc-training.webex.com/

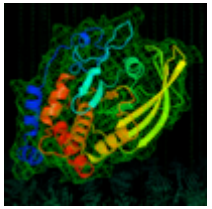
NUG Committees

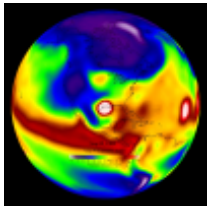
- **No action since April 11 NUG teleconference**
  - We'll start organizing this month
  - If you want to volunteer contact S. Ethier, F. Tsung, or R. Gerber
- **NERSC Achievement Awards**
  - Stephane Ethier
  - Cameron Geddes
- **NUG 2014 Meeting Planning**
  - Frank Tsung
- **Queue Advisory Committee**
  - Anubhav Jain
  - Stephen Bailey
  - Adrianne Middleton

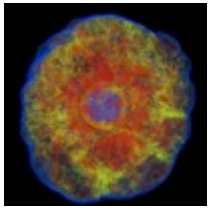
Edison Fair Share Experiment

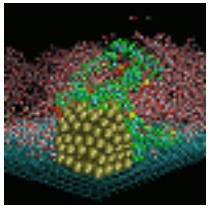












Fair Share Scheduling Experiment on Edison

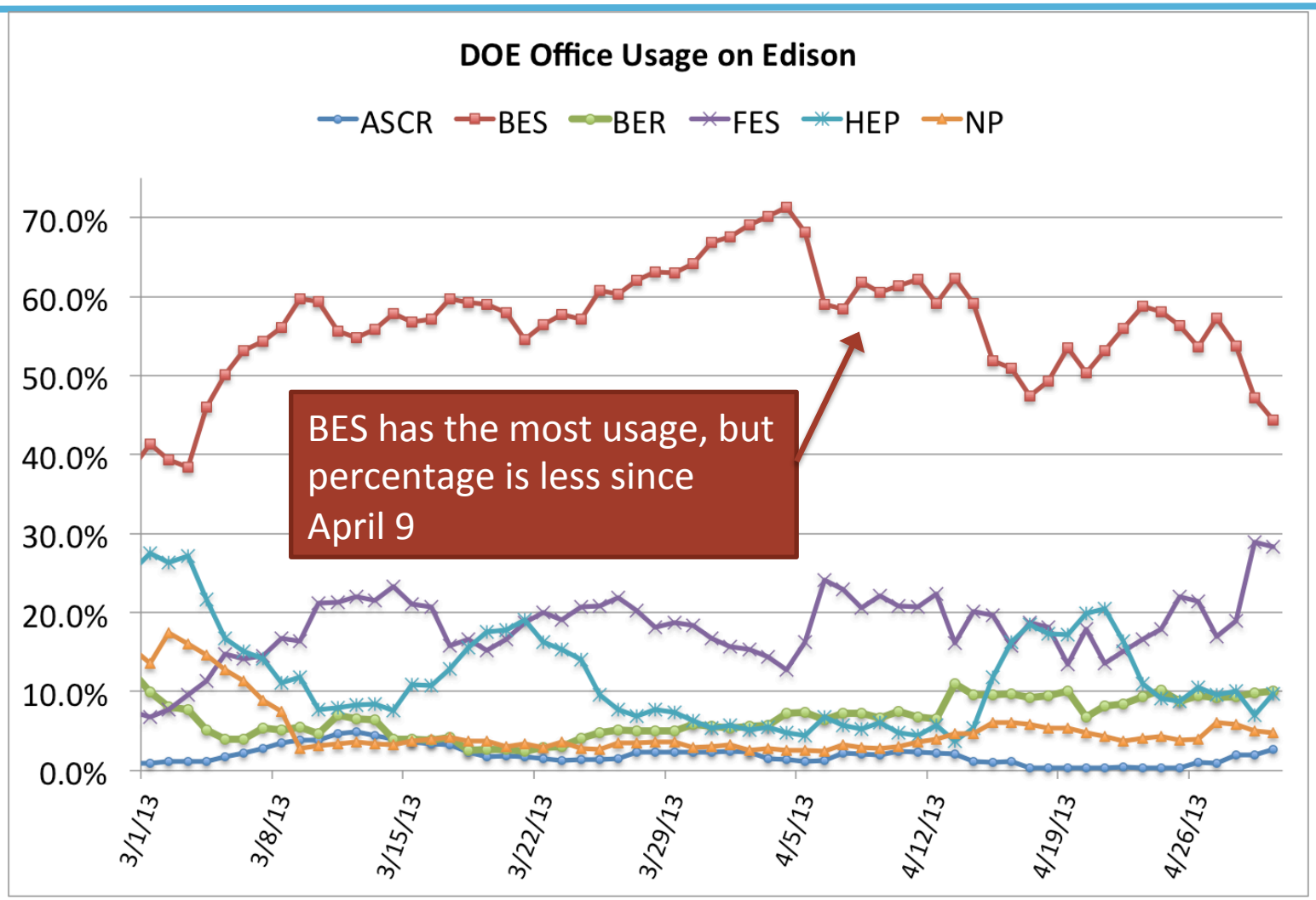
- See last month's slides on the web at for background.
  https://www.nersc.gov/users/NUG/teleconferences/april-2013/
- We implemented shares on April 9 based on DOE Office allocation percentages

Office	Share
ASCR	5 %
BER	18 %
BES	32 %
FES	18 %
HEP	14 %
NP	12 %

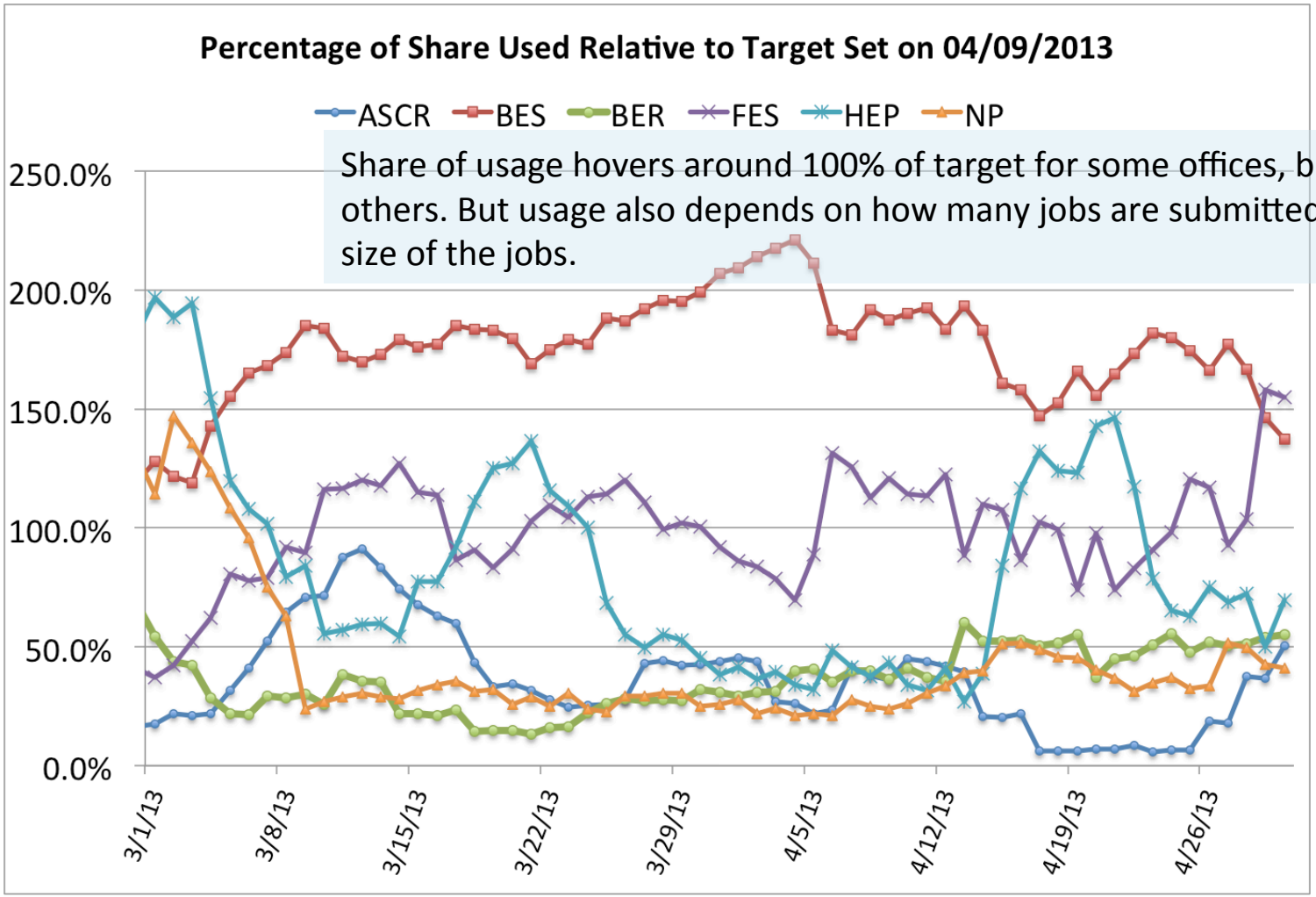
Fair Share Scheduling Experiment on Edison

- **Observations**
  - Usage per DOE Office is now closer to targets
  - Some users have complained about jobs "jumping in line"
  - DARPA mission partner usage is increasing, so the NERSC usage is decreasing somewhat
  - Review: Job priority has three components
    - Fair share
    - Wait time in queue
    - Queue priority (big job boost)
  - There are a number of ways to look at this; some usage plots follow. All data is smoothed over one week.

Percent of NERSC Usage by Office

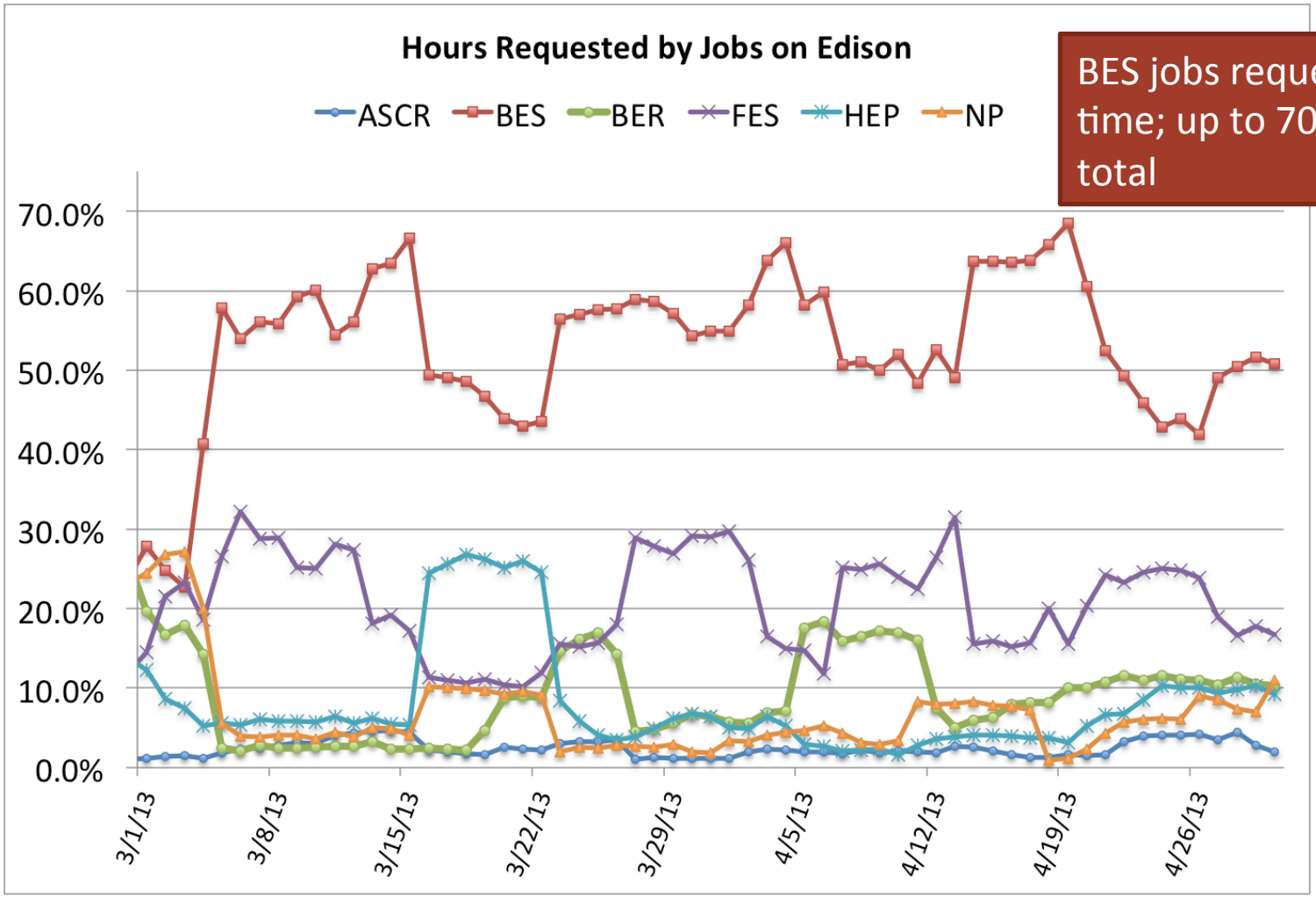


Percent of Share Target Used



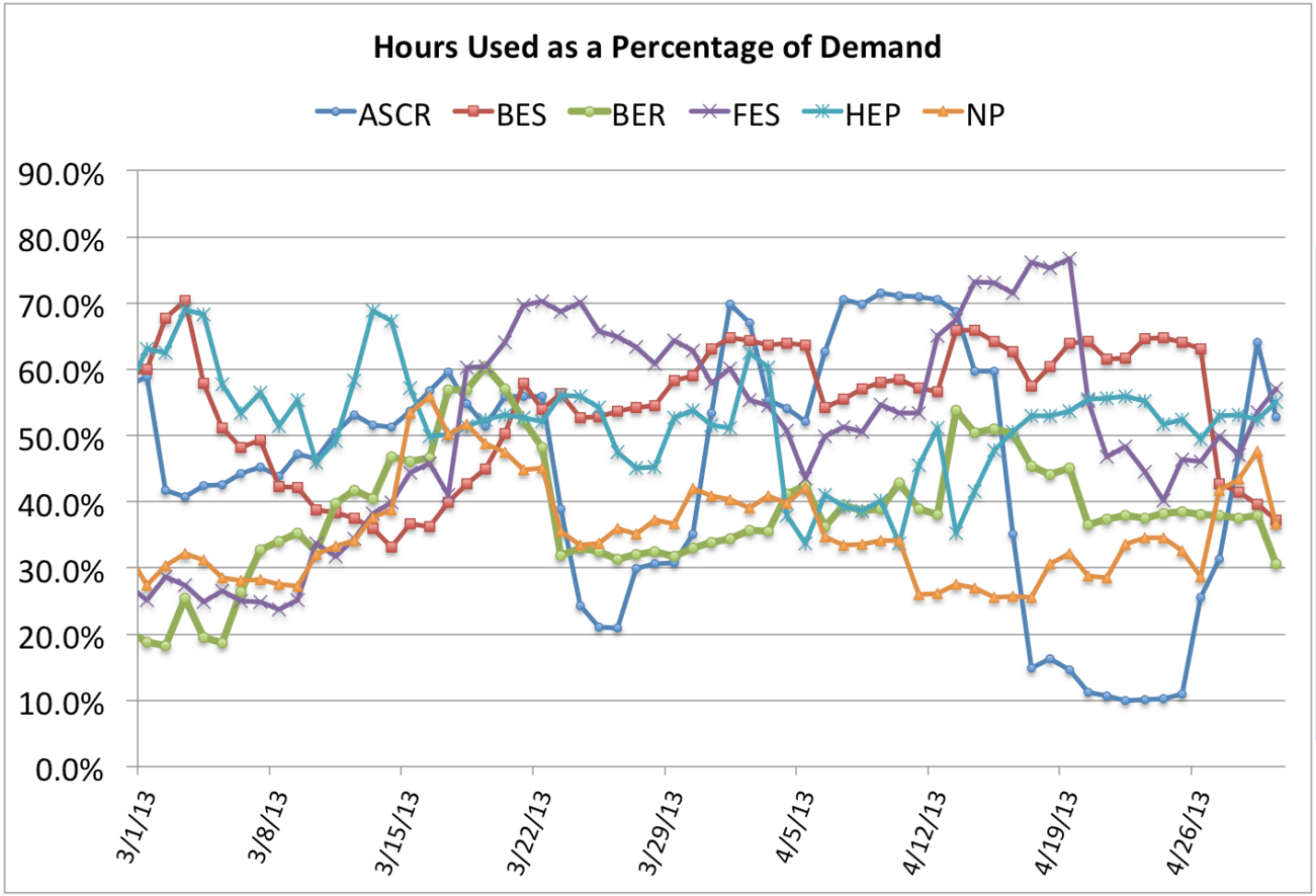
Share of usage hovers around 100% of target for some offices, but not others. But usage also depends on how many jobs are submitted and the size of the jobs.

Demand by Office



BES jobs request more time; up to 70% of total

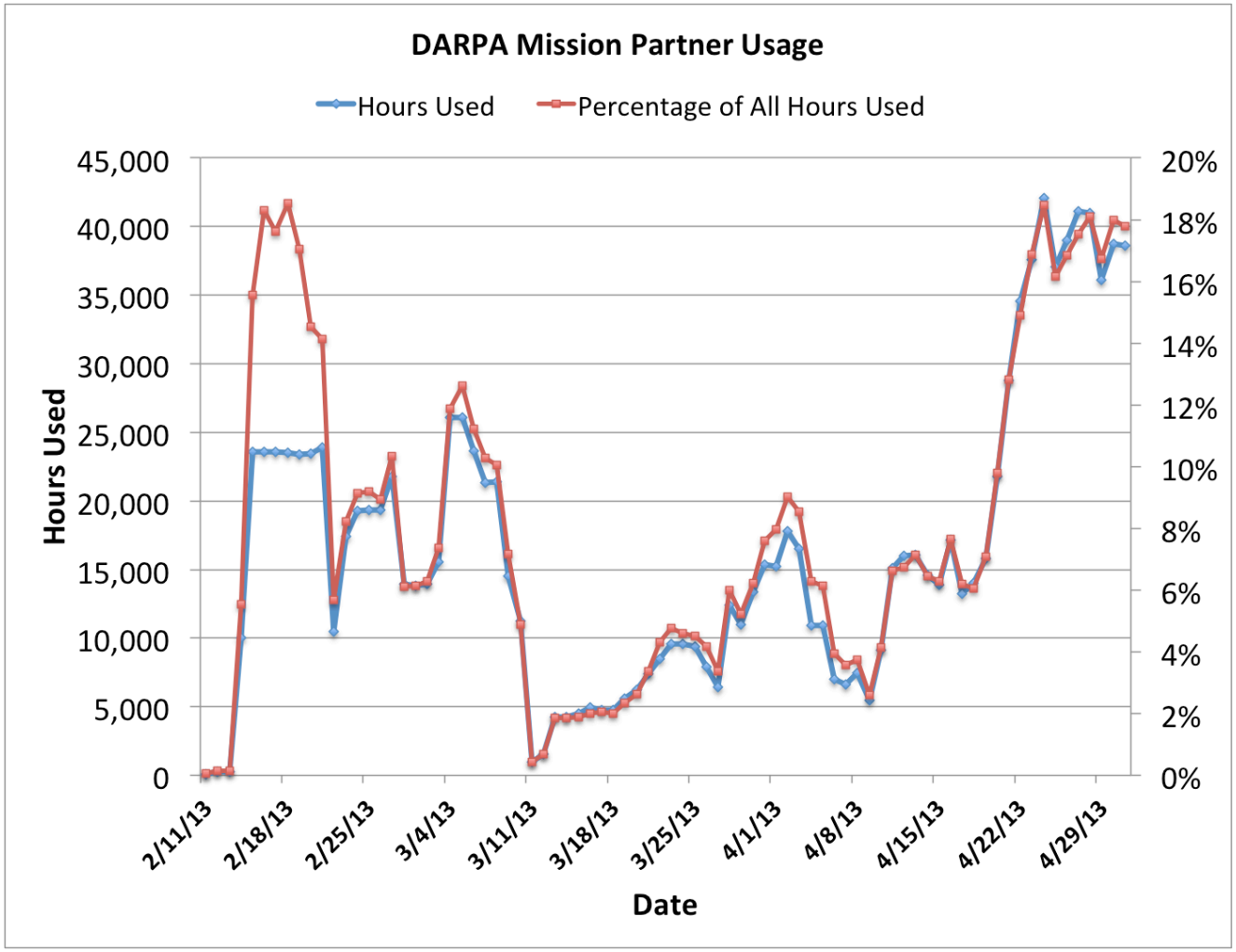
Usage vs. Demand



As a percentage of demand, most offices are actually getting similar throughput.

(The percentages are all low because jobs don't actually run for a long as they request.)

DARPA Mission Partner Usage



DARPA usage is increasing.

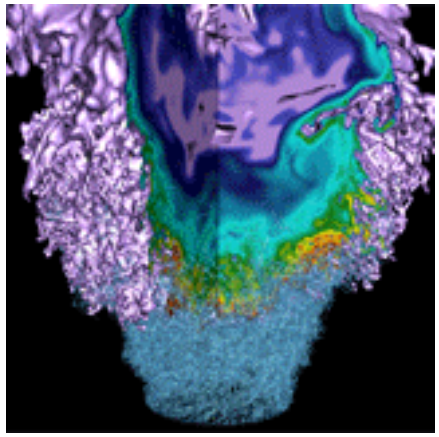
But still below 25% target because of low demand.

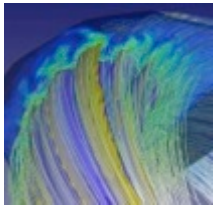
Used ~1.5 M of 20 M allocated

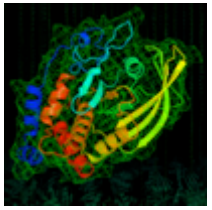
Early Conclusions

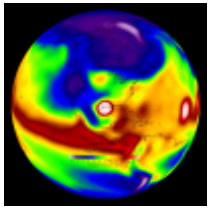
- **We think fair share is technically working as we expected, but do we have the scheduler configured the way we want it?**
- **We plan to add a user component so a single user can't continuously use up an office's share**
- **We may somewhat reduce the influence of the office's share relative to the queue wait time factor**
- **The algorithm is disconcerting and confusing to some users**
  - We plan to display the contributions from the three factors (fair share, wait time, queue) on the web queue look "very soon now"
- **Let us know what you think**

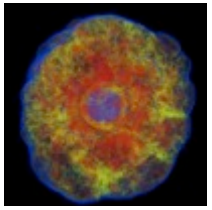
Interactive & Debug Reservations

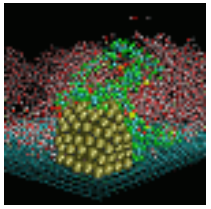








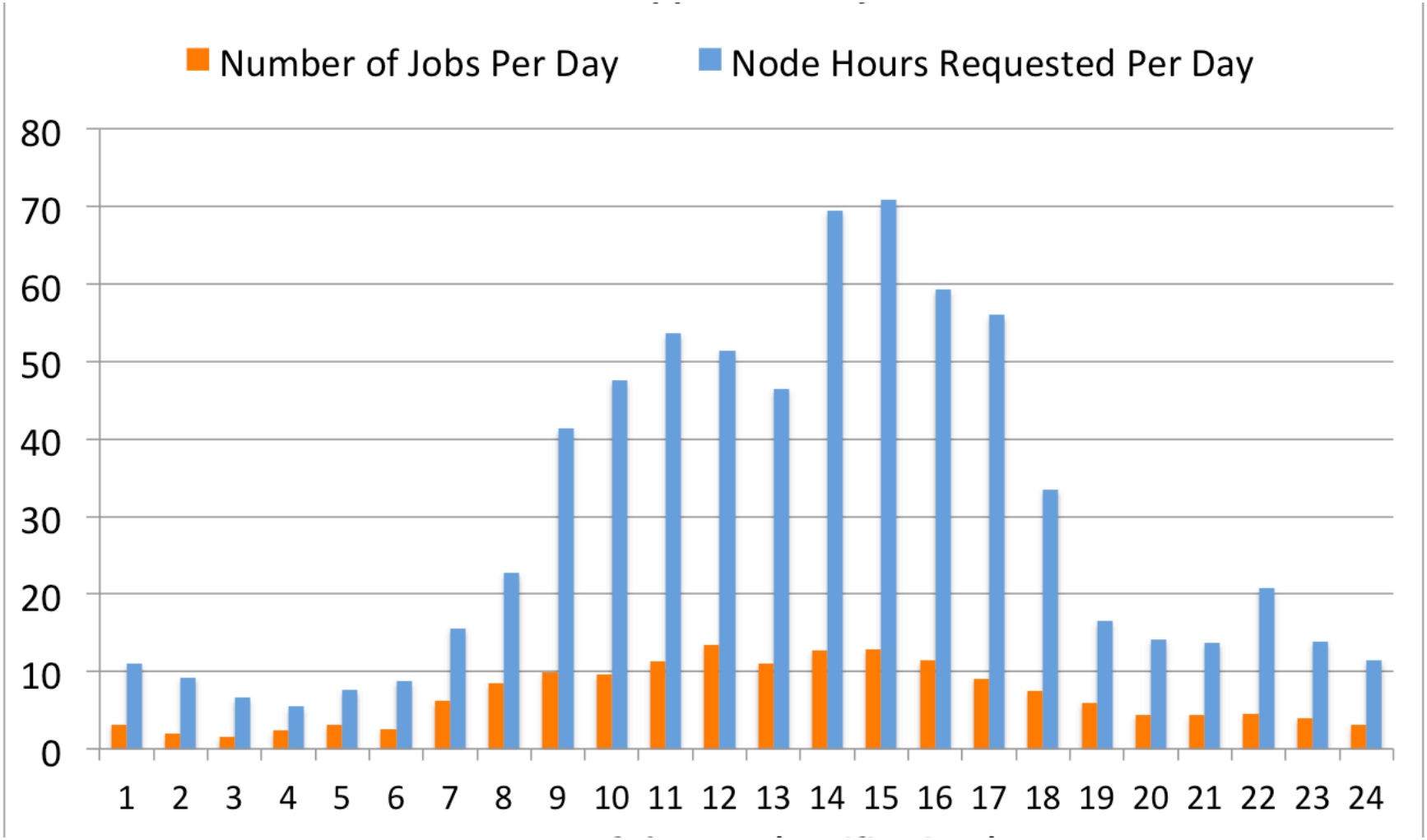




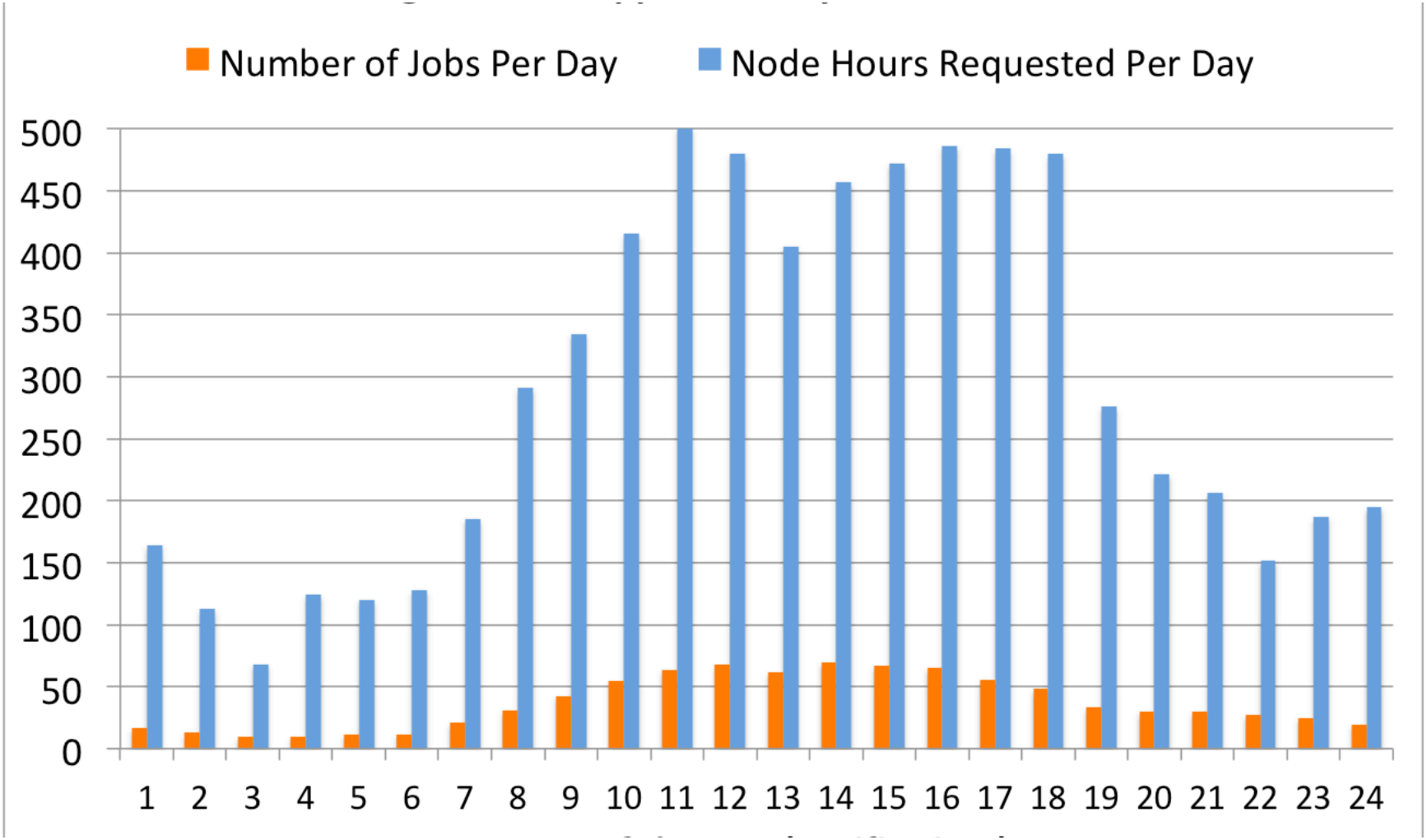
Reserved Nodes for Code Development

- We reserve 512 nodes on Hopper from 5:00 to 18:00 Pacific Time
- Is this adequate?
- We want to accommodate your interactive needs, but not have idle nodes
- We are considering setting aside some nodes outside the 5:00-18:00 time slot.
- Proposal: reserve some nodes from 18:00-24:00.
- How many? What do you think?

Interactive Queue

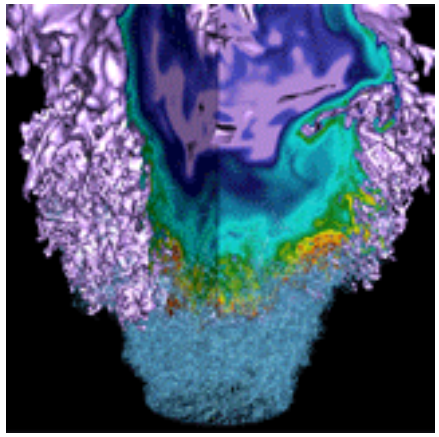


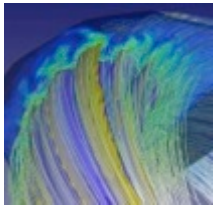
Debug Queue

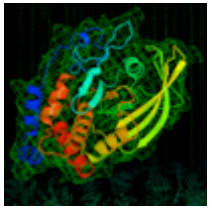

Number of Jobs Per Day
Node Hours Requested Per Day
500
450
400
350
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

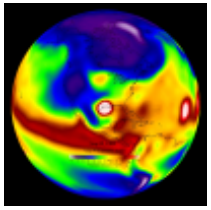
Perftools Lite

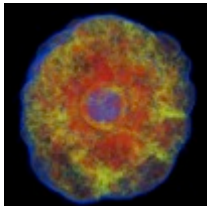
Helen He, User Services

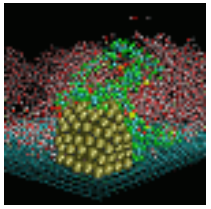












What is Perftools-lite

- A simplified and easy to use version of the CrayPat performance measurement and analysis tool.
- Released end of March, currently installed on Hopper.
- Provides basic performance analysis info automatically with simple steps.
- Users can decide whether to use full Perftools version afterwards.

Outputs from Perftools-lite

- **In stdout, basic information from the default "sample_profile" option:**
  - execution time,
  - memory high water mark,
  - aggregate FLOPS rate,
  - top time-consuming user functions
  - MPI information, etc.
- **A *.rpt text file with the same info as above**
- **A *.ap2 file that can be used with:**
  - "pat_report" for more detailed information
  - "app2" for graphic visualization
- **Possible one or more suggested MPICH_RANK_ORDER_FILE files.**

```
Experiment data directory written:
/scratch2/scratchdirs/yunhe/n6impl_20110622/GTC_1.0/run_opt/GTC-3092599.sdb/gtcmpi+16881887-5444s
########################################################################
#                                                                      #
#            CrayPat-lite Performance Statistics                       #
#                                                                      #
########################################################################
CrayPat/X:  Version 6.1.0 Revision 11030 (xf 10658)  03/20/13 16:42:24
Experiment:                  lite  sample_profile
Number of PEs (MPI ranks):   2048
Numbers of PEs per Node:       24  PEs on each of  85  Nodes
                                8  PEs on   1  Node
Numbers of Threads per PE:      1
Number of Cores per Socket:    12
Execution start time:  Fri Apr 26 16:12:34 2013
System name and speed:  nid04219 2100 MHz

Wall Clock Time:    1392.188696 secs
High Memory:              44.69 MBytes
MFLOPS (aggregate):  1747392.06 M/sec

Table 1:  Profile by Function Group and Function (top 7 functions shown)

 Samp%  |    Samp  |  Imb.  |  Imb.  |Group
        |          |  Samp  | Samp%  | Function
        |          |        |        |  PE=HIDE

 100.0% | 138825.3 |     -- |     -- |Total
|-----------------------------------------------------
|  83.3% | 115589.2 |     -- |     -- |USER
||----------------------------------------------------
||  42.4% |  58924.6 | 2280.4 |   3.7% |chargei_
||  34.7% |  48236.0 | 2863.0 |   5.6% |pushi_
||   4.5% |   6193.0 |  631.0 |   9.3% |shifti_
||   1.2% |   1669.6 | 2353.4 |  58.5% |poisson_
||====================================================
|  13.1% |  18121.2 |     -- |     -- |MPI
||----------------------------------------------------
||   7.9% |  10900.1 | 8247.9 |  43.1% |MPI_ALLREDUCE
||   5.1% |   7084.7 | 6554.3 |  48.1% |MPI_SENDRECV
||====================================================
|   3.7% |   5114.6 |     -- |     -- |ETC
||----------------------------------------------------
|   1.7% |   2348.4 |  200.6 |   7.9% | _HCOSS_V
|====================================================

Program invocation:  ./gtcmpi

For more detailed performance reports, run:
  pat_report /scratch2/scratchdirs/yunhe/n6impl_20110622/GTC_1.0/run_opt/GTC-3092599.sdb/gtcmpi+16881887-5444s.ap2

For interactive performance analysis, run:
  app2 /scratch2/scratchdirs/yunhe/n6impl_20110622/GTC_1.0/run_opt/GTC-3092599.sdb/gtcmpi+16881887-5444s.ap2

End of CrayPat output.
```

Steps to Use Perftools-lite

1. module load perftools-lite/6.1.0
2. module unload darshan (note: it has conflict with darshan)
3. Build as normal
4. In batch script, set CRAY_ROOTFS to DSL (note: the pat_report binary is dynamically linked)
5. If the job is submitted from a GPFS system, then also in the batch script:

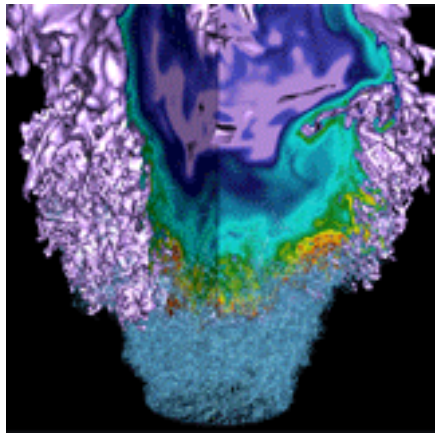
   set PAT_RT_EXPFILE_DIR to a directory in Lustre file system

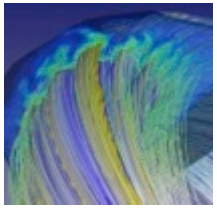
   or

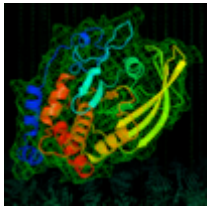
   set PAT_RT_EXPFILE_MAX >= the number of PEs or -1
6. Run as normal. Performance data is summarized at the end of the job stdout.
7. More detailed info can also be gathered with pat_report or apprentice2 after the run.

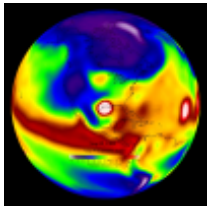
Math Library Performance on Edison

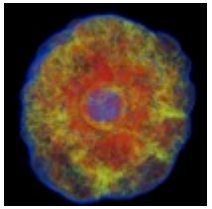
Jack Deslippe, User Services

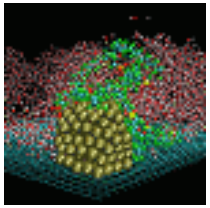












Next NUG Teleconference

- **Next scheduled: Thu. June 6, 2013**
- **Send suggested topics and comments to ragerber@lbl.gov**

National Energy Research Scientific Computing Center