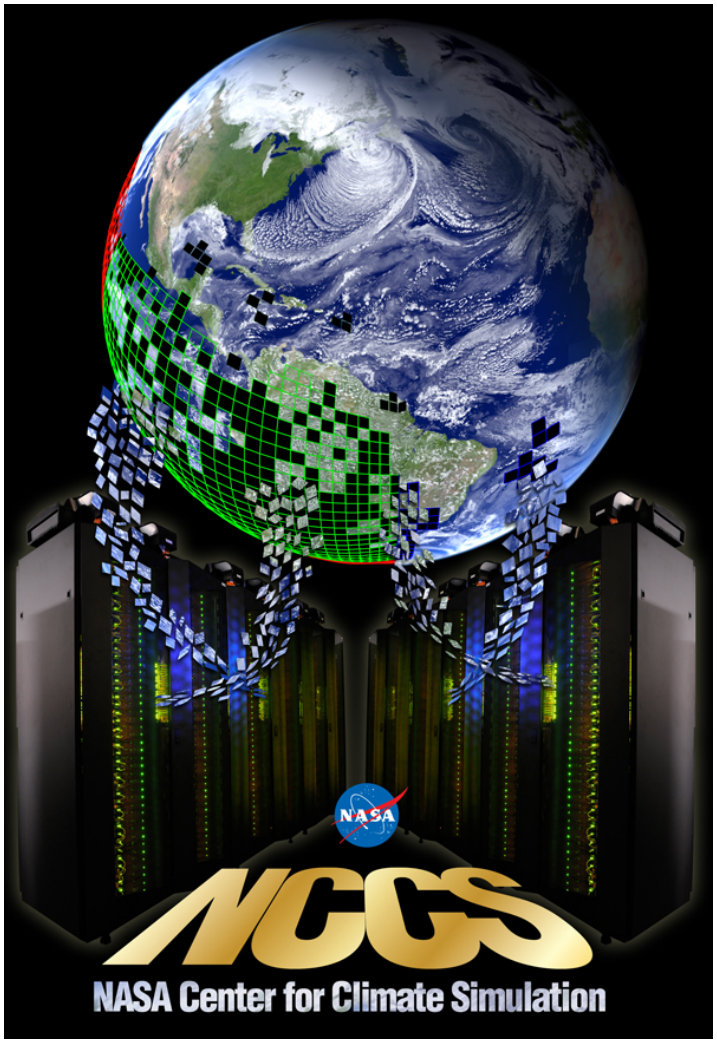


NCCS User Forum

June 25, 2013

Agenda

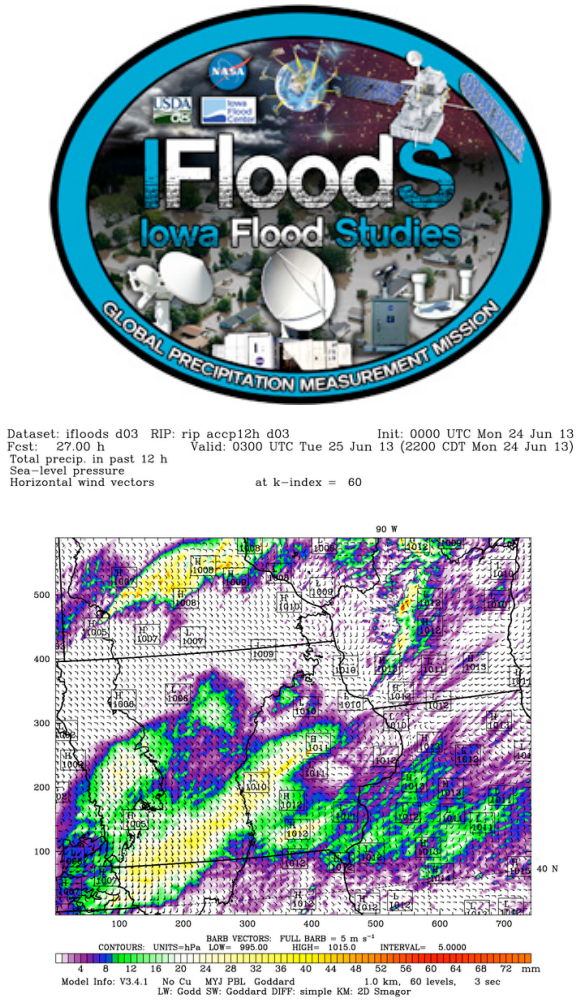
- Introduction
  - Recent Accomplishments
  - Increase in Capacity
  - Staffing
- NCCS Updates
  - Discover Updates (Including Intel Phi)
  - Remote Visualization
  - Data Portal
  - Update on NCCS Response to User Survey
  - Resource Manager Analysis of Alternatives
- NCCS Operations & User Services Updates
  - Upcoming and Status
  - Ongoing Investigations
  - Brown Bag and SSSO Seminars
- Questions & Answers

Recent Accomplishments

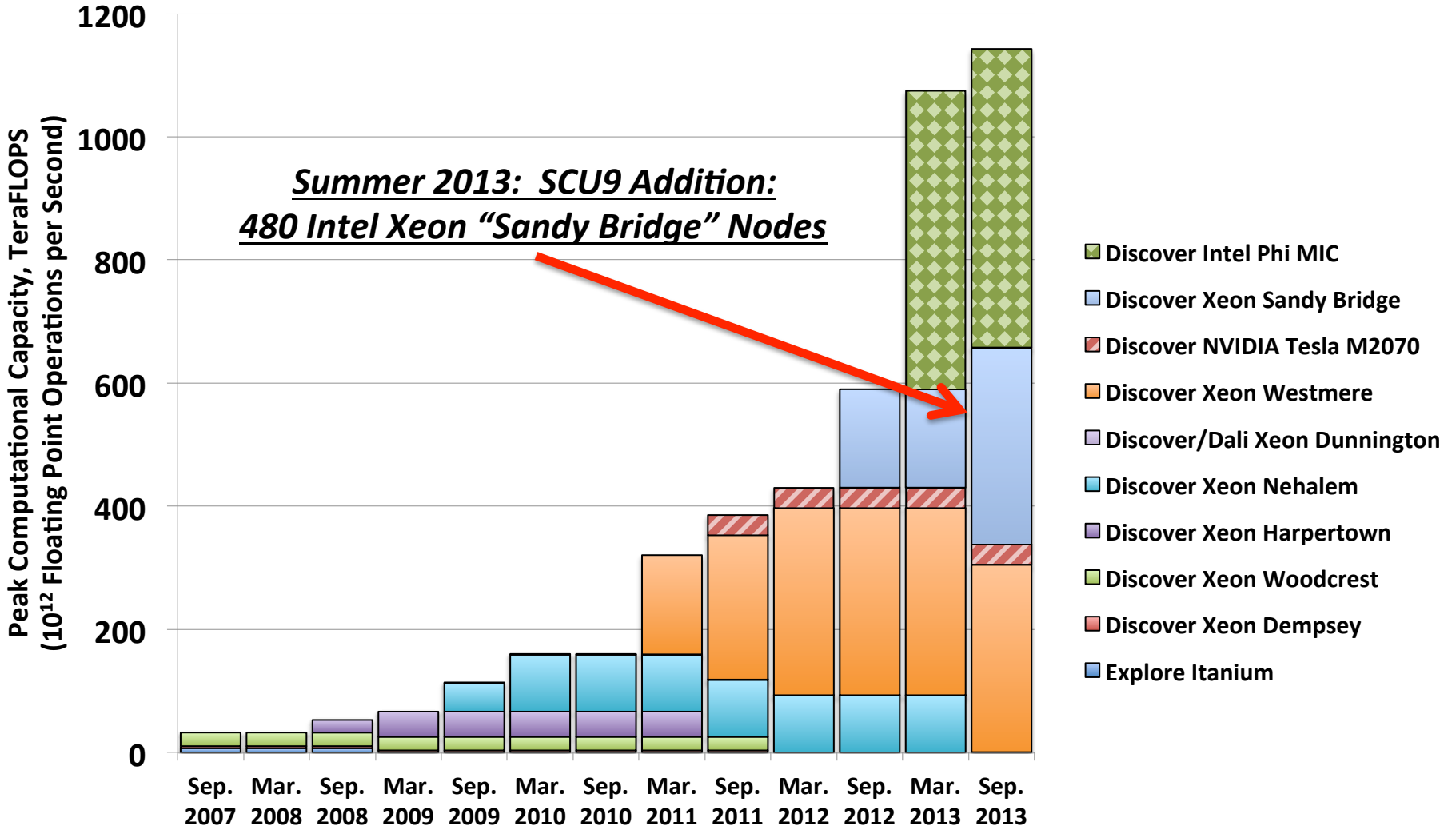
- 3 Months of Increasing Utilization
  - Rising to 74% utilization in May (based on PBS)
- 3 Months of Very High Availability
  - At or near 100% availability (does not include scheduled maintenance)
- Discover SCU9 (more later)
  - Intel Xeon SandyBridge – 480 nodes
- In-depth Intel Training and Brown Bag seminars (more to come)
- Earth System Grid Federation (ESGF) downloads
  - Over 326 TB and 10 million data sets, April 2011 – May 2013

Iowa Flood Studies Support

- IFloodS GPM Ground Validation Field Campaign support
  - Assisted Christa Peters-Lidard and her team to run NU-WRF forecasts
  - Two forecasts per day during the field campaign (ran on the SCU8 SandyBridge)
  - Tailored services (compute queues and storage) to meet the requirements for the campaign
  - No downtime during the campaign and no forecasts were missed



NCCS Compute Capacity Evolution September 2007- September 2013



Staff Additions

Welcome to New Members of the NCCS Team:

Garrison Vaughn

Julien Peters

Lyn Gerner (consultant)

Welcome to Summer Interns:

Jordan Robertson

Dennis Lazar

Winston Zhou

NCCS Updates

Dan Duffy,
HPC Lead and NCCS Lead Architect

Discover Updates

- SCU5/SCU6 Decommissioned
  - Nehalem processors, 8 cores per node, 24 GB of RAM
  - Space, power, and cooling used for SCU9
  - Part of the system will be reused internally and part will go to UMBC
- SCU8
  - SandyBridge processors, 16 cores per node, 32 GB of RAM
  - One Intel Phi accelerator per node
  - Available for general access
  - Special queue for native access by request (more on this later)

SCU9 Status

- SCU9
  - 480 SandyBridge Nodes
  - 4 GB of RAM per core; total of 64 GB per node
  - Does NOT contain any, but there is room for additional accelerators (Intel Phi or Nvidia GPUs)
  - Upgrades to all the Discover I/O nodes
  - Additional Discover Service nodes with SandyBridge processors
- Schedule
  - Integration and testing for the next 2 to 3 weeks
  - Pioneer usage during the month of July
  - General access in August

GPFS Metadata

- Discover GPFS Metadata Storage Upgrade
  - Goal is to dramatically increase the aggregate metadata performance for GPFS
  - Cannot speed up a single metadata query, but can speed up the combination of all queries
  - Acquisition in progress for solid state disks
  - Responses are being evaluated now
- Installation later this year, and the NCCS will coordinate closely with the user community during the integration of this new capability

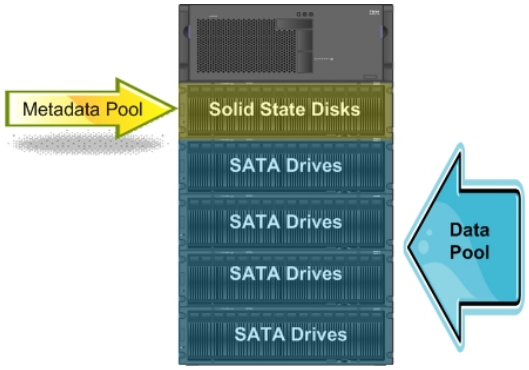


Image source: IBM

Data Portal Upgrade

- Disk Capacity
  - Adding ~100 TB of usable disk
- Additional Servers
  - Servers with 10 GbE capability
  - Support higher speed access between Discover and the Data Portal

Update on NCCS Response to User Survey

1. External data transfer performance and convenience
   - Focus has been on upgrading the Data Portal servers with 10 GbE capabilities to facility faster transfer performance
   - Analysis of the GISS to NCCS network and recommendations for upgrades
   - Upgrade of the SEN to CNE link to 10 GbE
2. More timely notifications of problems or unplanned outages
   - Web dashboard for system status is under development
3. "Architecting for More Resiliency," especially the Discover storage file systems
   - Initial architecture thoughts and requirements have been captured
   - Creation of a tiger team of NCCS and non-NCCS team members to look at how to architect for a higher resiliency
   - Evaluation of alternative computing platforms, including cloud

Resource Manager Analysis of Alternatives

- Is there an alternative to PBS that better meets NCCS and user community requirements?
- NCCS, with the support of Lyn Gerner, has generated an Analysis of Alternatives for the resource management software.
  - Includes a mapping of capabilities to requirements, and a cost/ benefit analysis.
  - A recommendation has been made to the NCCS management.
  - A decision will be made in a short time period (weeks).
  - Users will be notified of any changes as time goes on.
- The goal is to make any change as transparent to the users as possible.
- Stay tuned!

Discover Intel Phi
Many Integrated Core (MIC) Coprocessor

- Discover's 480 SCU8 nodes have one Intel Phi coprocessor per node.
  - Direct use of Phi is now available via "native" queue.
  - Offload use available on all other SCU8 nodes.
    - Coming soon: method to specify SCU8 nodes for Intel Phi Offload use.
- Training:
  - A number of NCCS Brown Bags so far.
    - Content available on NCCS web site.
    - Training will be repeated upon request.
  - Contact support@nccs.nasa.gov

NCCS Code Porting Efforts for Discover Intel Phi Many Integrated Core (MIC)

- NCCS staff, SSSO, vendor, and external community members are currently working on the following codes:
  - GEOS-5 components (GMAO, NOAA/GFDL, et al.)
  - GRAIL (high degree/high order solutions of lunar gravity field)
  - Ice Melt code (Kwo-Sen Kuo et al.)
  - WRF (with NOAA/ESL, NCAR, et al.)
- Contact support@nccs.nasa.gov .

Remote Visualization Prototype

- A prototype Remote Visualization platform is being investigated for the UVCDAT advanced visualization application.
  - Goal: applications such as UVCDAT would run on NCCS resources, with displays *and controls* on user desktop.
  - Should dramatically speed up remote visualization for users.
- Requires careful network and security configuration to safeguard NCCS resources and user desktop.
- Tests of supporting technologies are underway.
- Following evaluation of alternatives, will move into deployment and "pioneer" phase; stay tuned.

NCCS Operations & User Services Update

Ellen Salmon

- Upcoming & Status
- Ongoing Investigations
- NCCS Brown-Bag and SSSO Seminars

Upcoming (1 of 2)

- Discover resources:
  - SCU8 Sandy Bridge and SCU8 Intel Phi Coprocessors (MICs):
    - Intel Phi native use now available via "native" queue.
    - Want help porting code to the Intel Phis (either "offload" or "native")?
      - Contact support@nccs.nasa.gov.
  - SCU9 Sandy Bridge :
    - Somewhat reduced total Discover compute cores until late July/August (when pilot usage starts).
    - Targeting August for general availability.
  - Discover GPFS nobackup:
    - Following May's GPFS parameter changes, continuing to deploy additional "NetApp" nobackup disk space in a measured fashion.
      - Moving nobackup directories to disk array types best suited for their workloads.
    - Watch for more info on GPFS metadata storage upgrade as acquisition progresses.

Upcoming (2 of 2)

- Discover InfiniBand OFED (software stack) changes:
  - Continuing the rolling migration to required, upgraded, InfiniBand OFED (software stack).
    - 2/3 of Discover is already on the new OFED release
      - All computational nodes on InfiniBand Fabric 2: SCU7 Westmere nodes, and all SCU8 and SCU9 Sandy Bridge nodes
    - Rolling, announced, gradual changeovers of other parts of Discover (e.g., via PBS queues or properties).
      - SCUs 1 through 4, handful of remaining interactive and Dali nodes
    - ***Recompile is recommended***.
      - Some codes work fine without a recompile.
      - Other codes require a recompile to take advantage of some advanced features.

- Planned Outages (to date):
  - Discover downtime (full day) sometime during Field Campaign hiatus (July 16 – August 4)
    - Upgrade remaining (InfiniBand Fabric 1) I/O nodes to Intel Xeon Sandy Bridge nodes
    - Other NCCS systems may also "piggyback" on this downtime, stay tuned.

Discover Compiler / Library Recommendations

- Use **current libraries and compilers** to get many benefits:
  - Executables created with older versions can experience problems running on Discover's newest hardware.
  - Often, simply rebuilding with current compilers and libraries ***(especially Intel MPI 4.x and later)*** fixes these issues.
  - Current versions can enable use of new features like Sandy Bridge nodes' advanced vector extensions (AVX) for improved performance.
  - Use of current versions greatly increases NCCS staff's abilities to track down other problems…
    - Especially when seeking vendor support to fix problems.

Archive Tape Media Issue

- Crinkled/damaged archive tapes caused a number of "Please examine/replace these archive files" tickets in the last several months.
- ***Damage is no longer occurring.***
- Oracle identified **faulty tape motor on a single tape drive** as the cause, and:
  - Replaced that tape drive and 11 others to proactively remediate the problem, prior to pinpointing the cause.
  - Providing data recovery services to extract usable data from damaged tapes.
  - Replacing all tape media affected by the problem.
- ~12 tapes sent, so far, to Oracle Tape Recovery Services.
  - So far 5 have been returned.
    - Recovered all data on 2 of the tapes and much of the data on the other 3 tapes.
- Larger list of tapes was damaged, but NCCS staff was able to recover files from those because second copies of files still existed on separate (unaffected) tapes.
- Reminder: **dmtag –t 2 <filename>** to get two tape copies of archive files, where needed.

Ongoing Discover Investigations

GPFS and I/O

- GPFS slowness due to heavy parallel I/O activity.
  - Significant GPFS parameter changes made in May to help address issues, but much additional work remains.
    - E.g., many-month effort: background "rebalancing" of data among filesystems to better accommodate workloads.
  - New Sandy Bridge I/O nodes' capabilities will help.
    - More cores per I/O node—16 cores, rather than 8—improved concurrency.
    - More total memory channels—4, rather than 3, per "socket"—better for data moving.
    - More total I/O "lanes" per I/O node.
- Heavy GPFS metadata workloads.
  - Acquisition in progress for new metadata storage.
  - Target: improve responsiveness in handling many concurrent small, random I/O actions (e.g., for directories, filenames, etc.).

PBS "Ghost Jobs"

- Extremely rare due to successful mitigation strategy–report jobid if you see one!

NCCS Brown Bag Seminars

- ~Twice monthly in GSFC Building 33 (as available).
- Content is available on the NCCS web site following seminar:

  https://www.nccs.nasa.gov/list_brown_bags.html

- Current emphasis:

  Using Intel Phi (MIC) Coprocessors

- Current/potential Intel Phi Brown Bag topics:
  - ✓ Intro to Intel Phi (MIC) Programming Models
  - ✓ Programming on the Intel MIC Part 2 – How to run MPI applications
  - Advanced Offload Techniques for Intel Phi
  - Maximum Vectorization
  - Performance Analysis via the VTune™ Amplifier
  - Performance Tuning for the Intel Phi

Questions & Answers

NCCS User Services:

support@nccs.nasa.gov

301-286-9120

https://www.nccs.nasa.gov

Contact Information

NCCS User Services:

support@nccs.nasa.gov

301-286-9120

https://www.nccs.nasa.gov

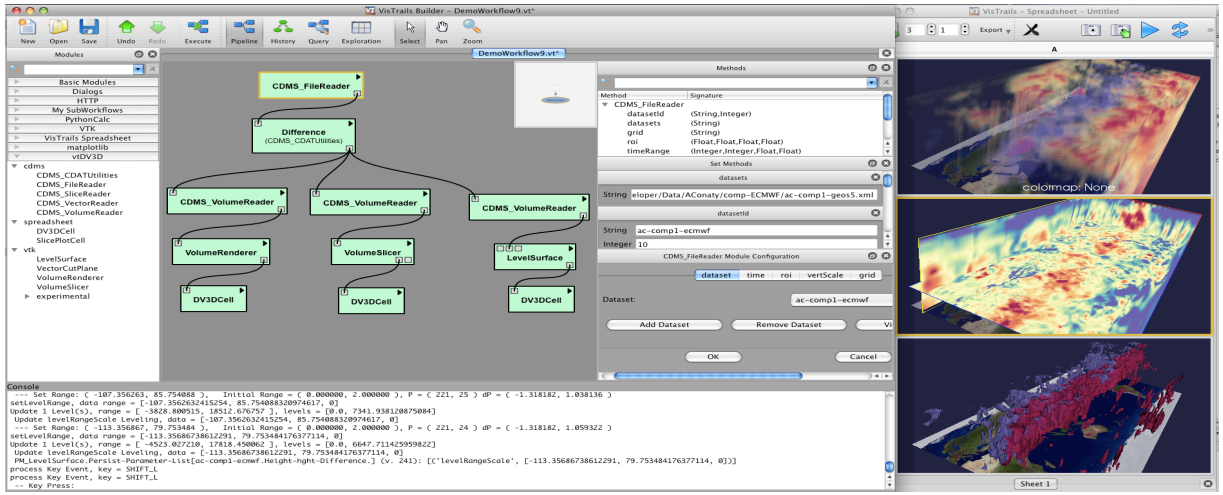
http://twitter.com/NASA_NCCS

Thank you

Supporting Slides

NCCS Brown Bag: Climate Data Analysis and Visualization using UVCDAT

- Climate scientist Jerry Potter (606.2) and analysis/ visualization expert Tom Maxwell (606.2) demonstrated how climate scientists can use the open-source Ultrascale Visualization Climate Data Analysis Tools (UVCDAT) to explore and analyze climate model output, such as the NetCDF-formatted model output files produced by GEOS-5 system and MERRA.



A UVCDAT Vistrails demonstration screenshot, displaying (clockwise from top center): the VisTrails workflow builder, a VisTrails "spreadsheet" of visualizations, console window, and UVCDAT modules in use (e.g., vtDV3D and the matplotlib Python module). PoC: Thomas.Maxwell@nasa.gov

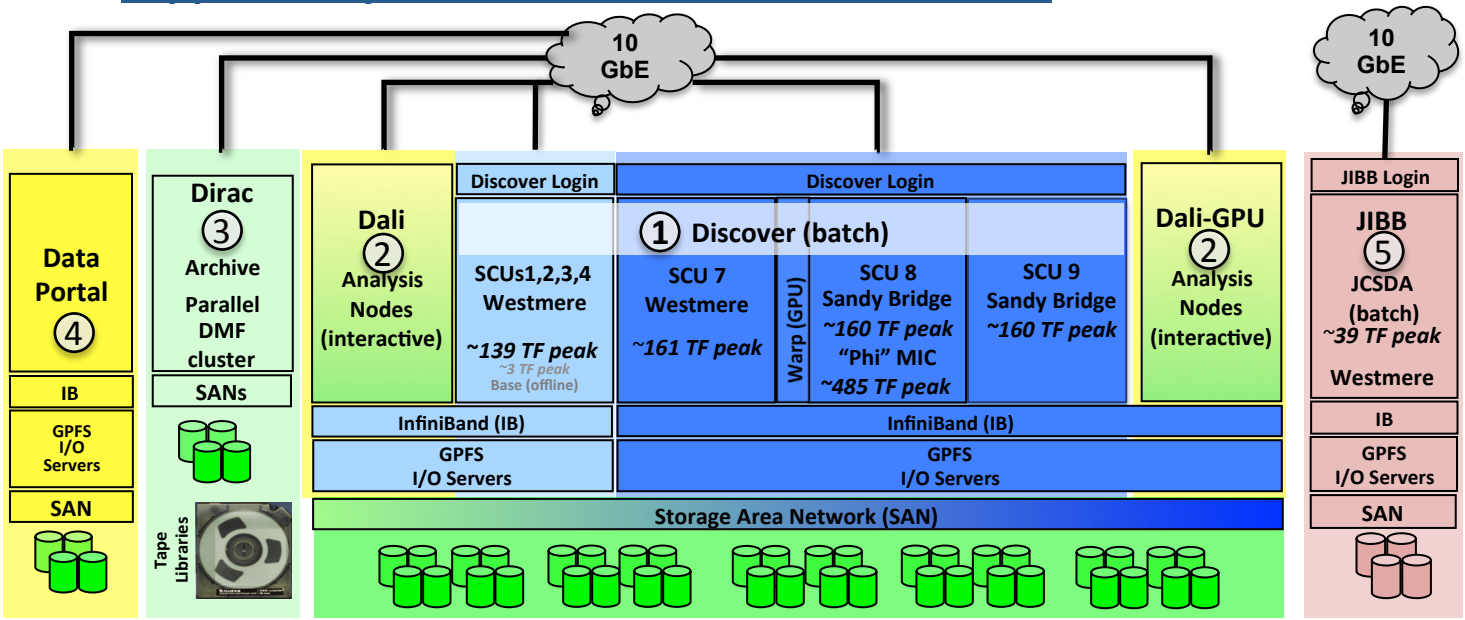
- The UVCDAT tools feature workflow interfaces, interactive 3D data exploration, automated provenance generation, parallel task execution, and streaming data parallel pipelines, and can enable hyperwall and stereo visualization.
- UVCDAT is the new Earth System Grid analysis framework designed for climate data analysis, and it combines VisTrails, CDAT and ParaView.
- Tom Maxwell developed vtDV3D, a new module included with recent UVCDAT and VisTrails releases, which provides user-friendly workflow interfaces for advanced visualization and analysis of climate data via a simple GUI interface designed for scientists who have little time to invest in learning complex visualization frameworks.

NASA Center for Climate Simulation Supercomputing Environment

Supported by HQ's Science Mission Directorate

① ***Discover*** Linux Supercomputer, June 2013:

- Intel Xeon nodes
  - ~3,200 nodes
  - ~43,000 cores
  - Peak ~624 TFLOPS general purpose
  - 97 TB memory (2 or 4 GB per core)
- Coprocessors:
  - Intel Phi MIC
    - 480 units
    - ~485 TFLOPS
  - NVIDIA GPUs
    - 64 units
    - ~33 TFLOPS
- Shared disk: **7.2** PB



② ***Dali*** and ***Dali-GPU*** Analysis

- 12- and 16-core nodes
- 16 GB memory per core
- Dali-GPU has NVIDIA GPUs

③ ***Dirac*** Archive

- 0.9 PB disk
- ~70 PB robotic tape library
- Data Management Facility (DMF) space management

④ ***Data Portal*** Data Sharing Services

- Earth System Grid
- OPeNDAP
- Data download: http, https, ftp
- Web Mapping Services (WMS)server

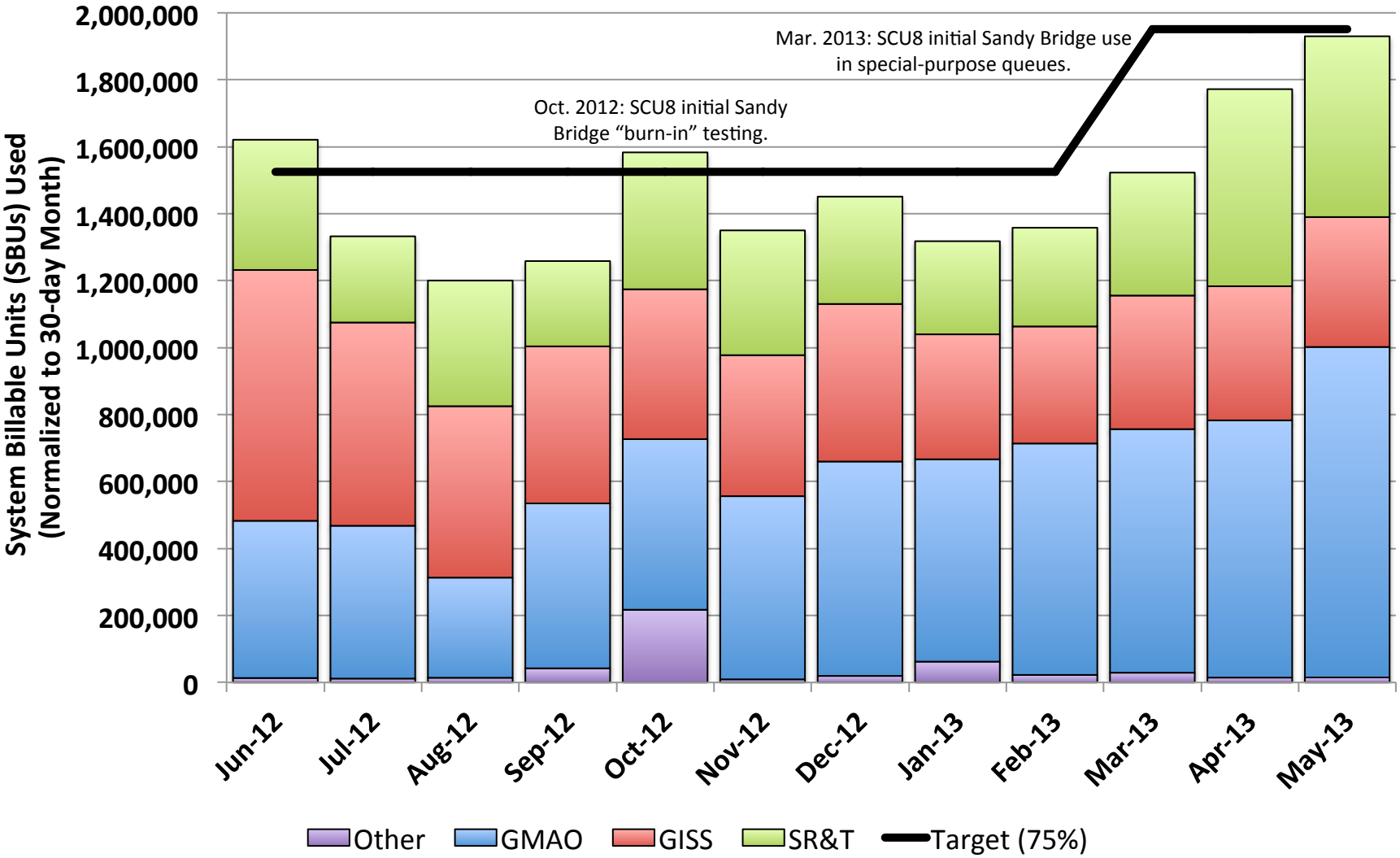
⑤ ***JIBB***

- Linux cluster for Joint Center for Satellite Data Assimilation community

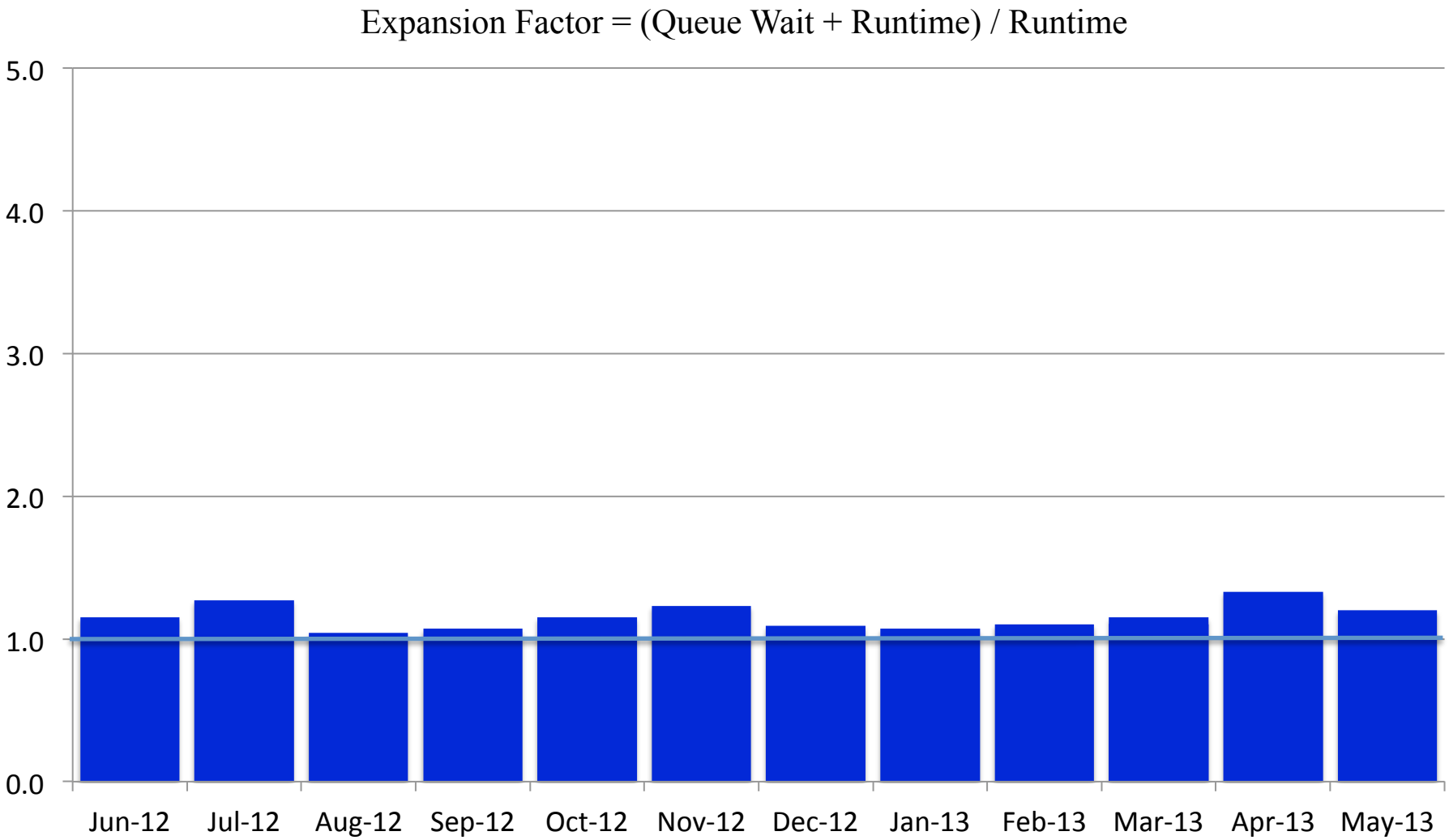
March 1, 2013

NCCS Metrics Slides
(Through May 31, 2013)

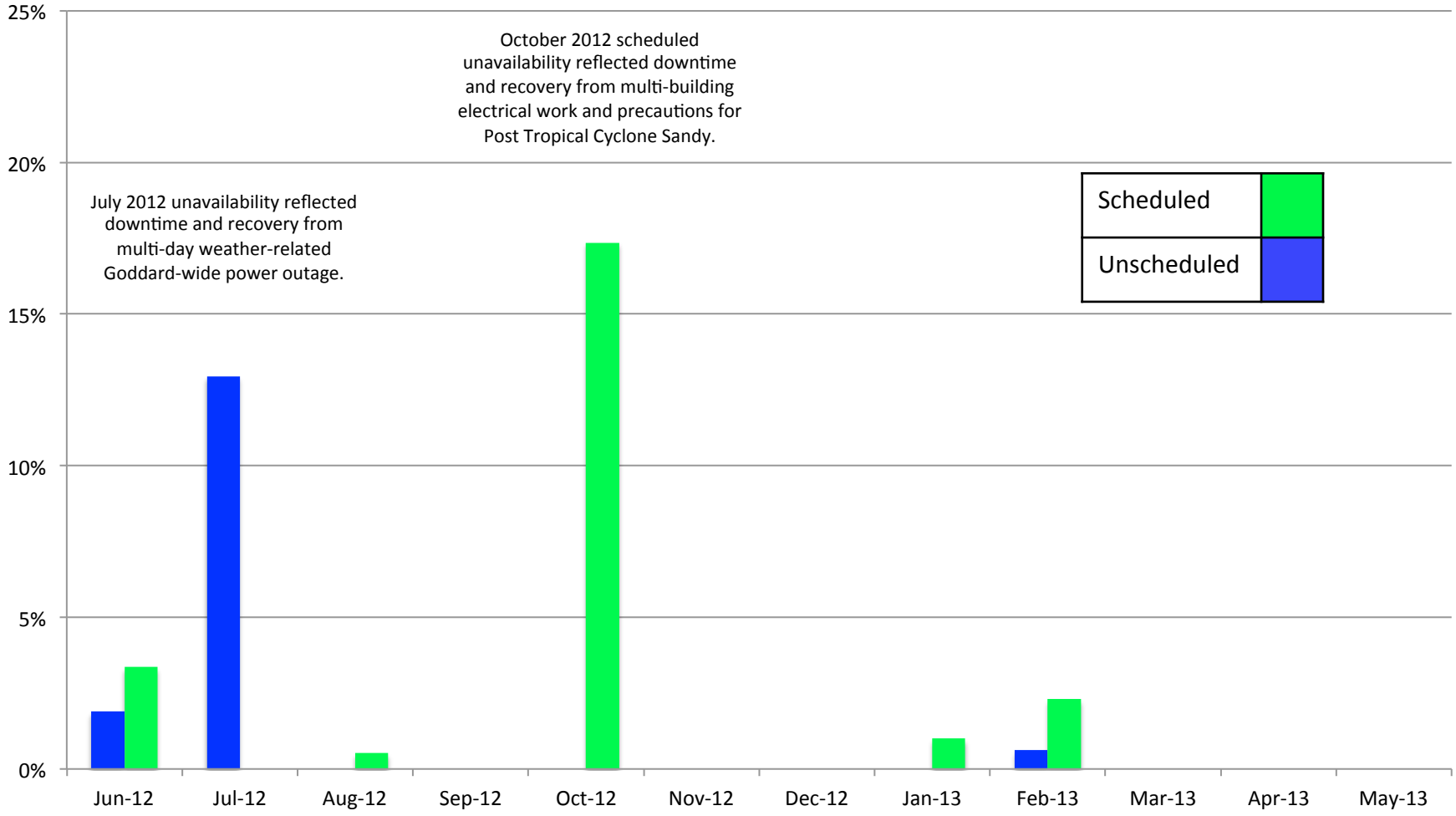
NCCS Discover Linux Cluster Utilization Normalized to 30-Day Month



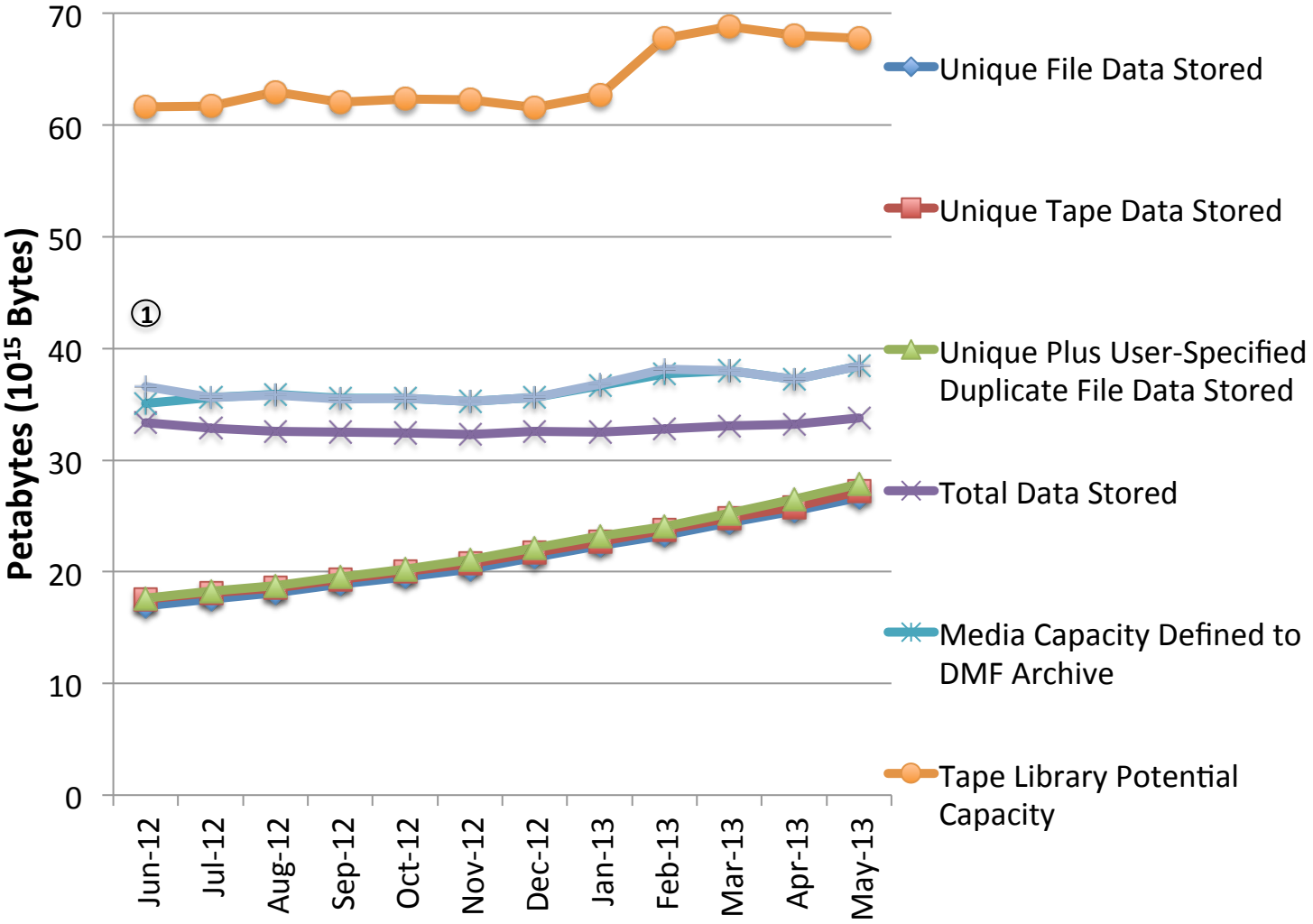
Discover Linux Cluster Expansion Factor



Discover Linux Cluster Downtime



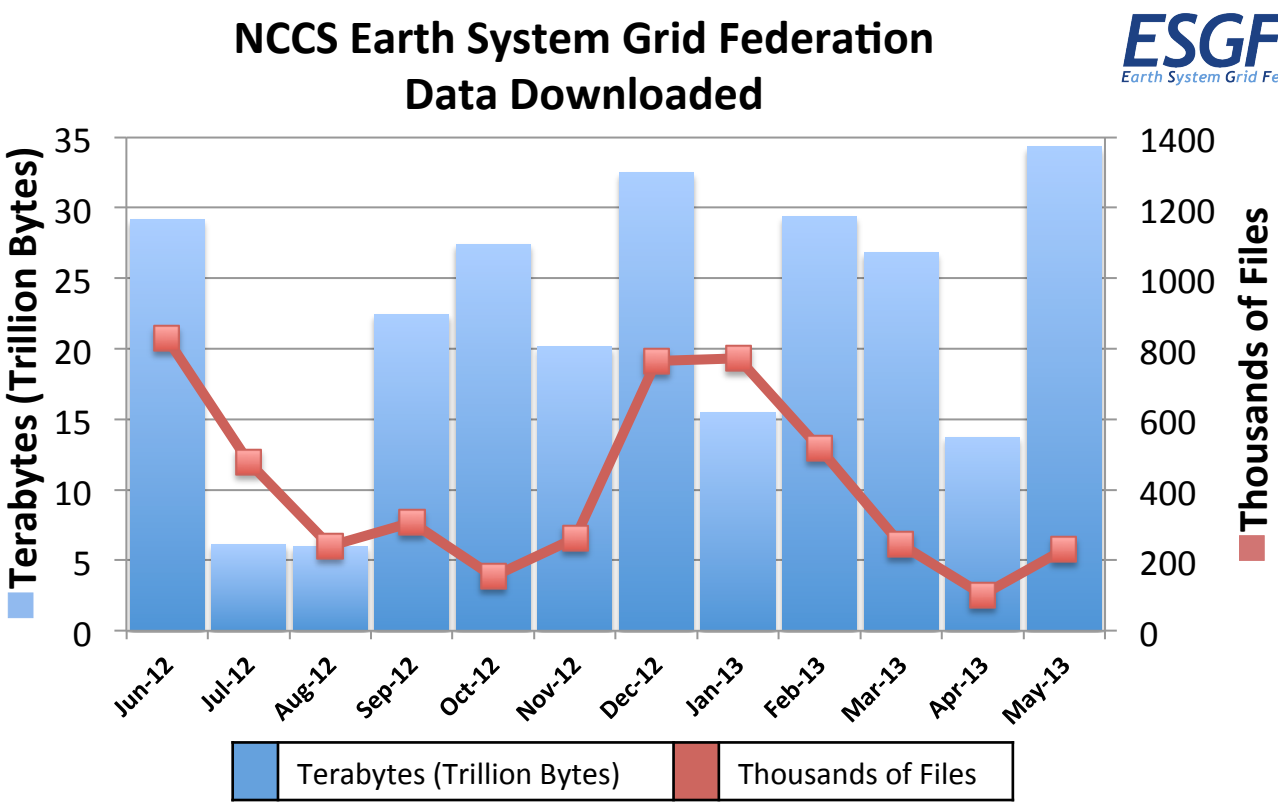
NCCS Mass Storage



① As of late May, 2012, NCCS changed the Mass Storage default so that two tape copies are made only for files for which two copies have been explicitly requested. NCCS is gradually reclaiming second-copy tape space from legacy files for which two copies have not been requested.

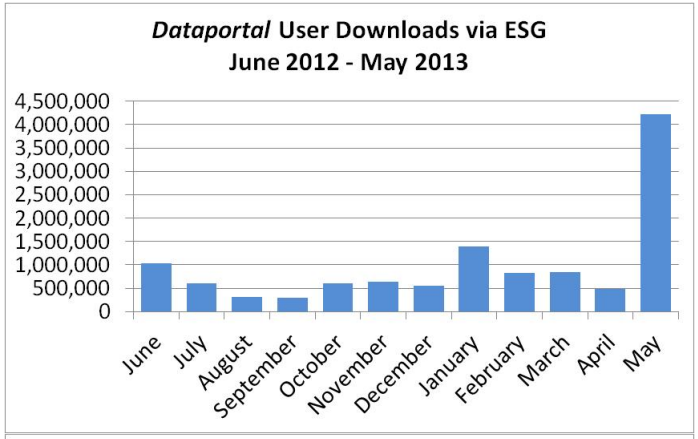
NCCS Earth System Grid Federation Services for NASA's and Peer Organizations' Climate Simulations, Selected NASA Observations, and Selected Analyses

- **GISS and GMAO** researchers are using the NCCS Discover cluster for simulations in support of the fifth phase of the Coupled Model Intercomparison Project (**CMIP5**), which supports the **Intergovernmental Panel on Climate Change's Fifth Climate Assessment (IPCC AR5)** and related research.
- The research community accesses data via the NCCS's **Earth System Grid Federation (ESGF)** node http://esgf.nccs.nasa.gov/ .

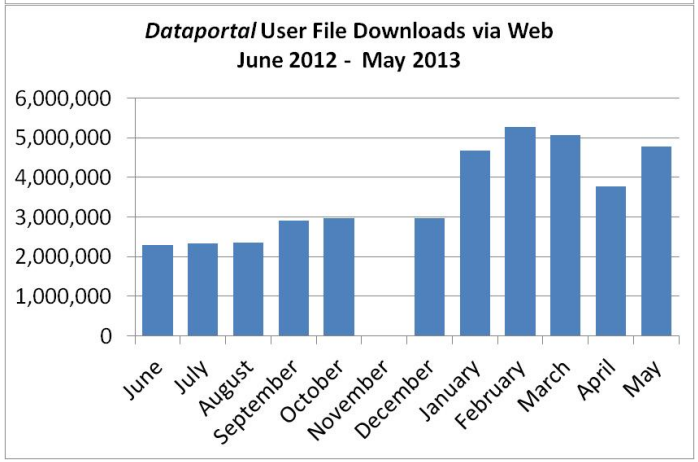


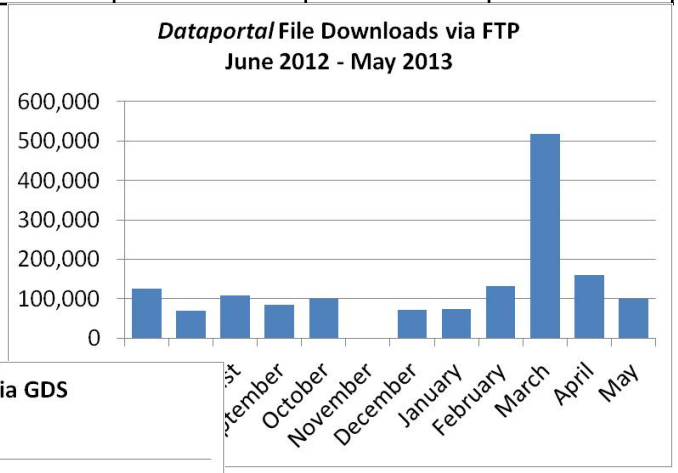
- The **NCCS Data Portal** serves data in CF-compliant format to support these Earth System Grid Federation Projects:
  - **CMIP5**: Long-term NASA GISS simulations, and decadal simulations from NASA's **GMAO**; **NOAA NCEP**; and **COLA** (Center for Ocean-Land-Atmosphere Studies).
  - **Obs4MIPs**: selected satellite observations from NASA's **GPCP**, **TRMM**, **CERES-EBAF,** and **Terra MODIS**.
  - **Ana4MIPs**: analyses from NASA/GMAO's Modern Era Retrospective-Analysis for Research and Applications (**MERRA**).
  - **NEX-DCP30**: bias-corrected, statistically downscaled (0.8-km) CMIP5 climate scenarios for the conterminous United States.

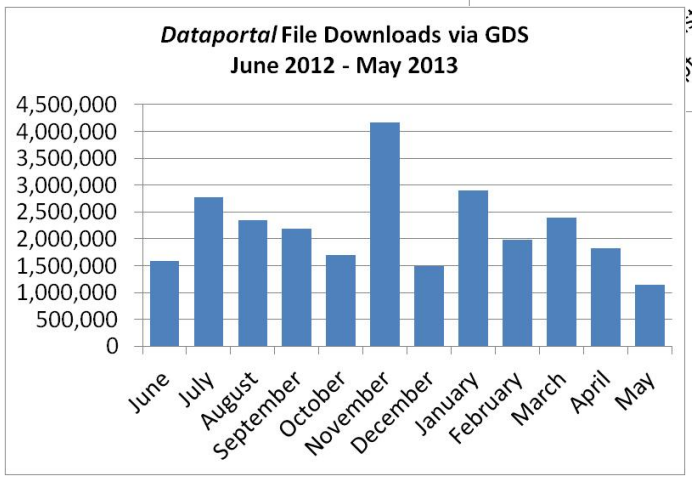
Dataportal Utilization – File Downloads



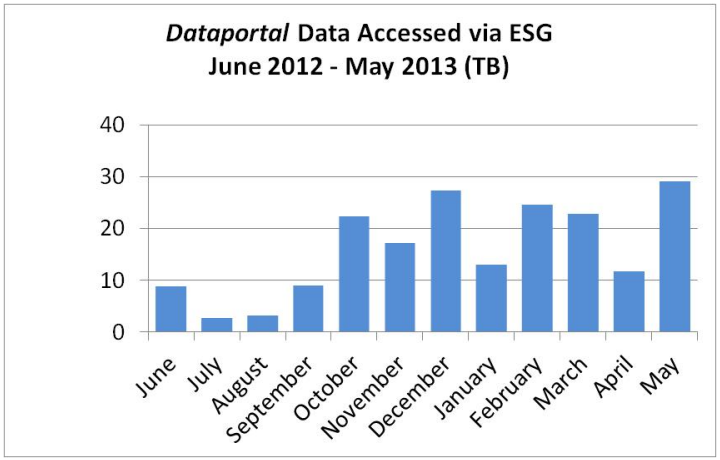
Download Mechanism	Downloaded Data Files			Available Data (TB)
	March	April	May	
Web	5,068,450	3,776,964	4,771,425	89
ESGF	838,721	493,281	4,219,467	153
FTP	517,941	160,465	98,902	143
GDS	2,396,618	1,827,154	1,143,213	91



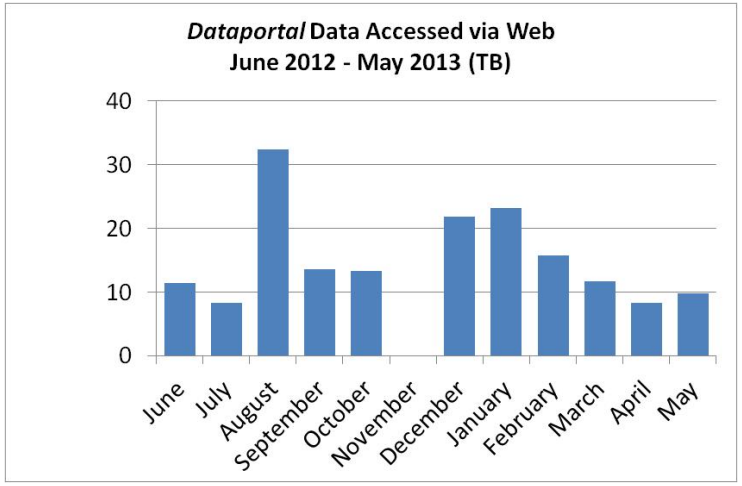


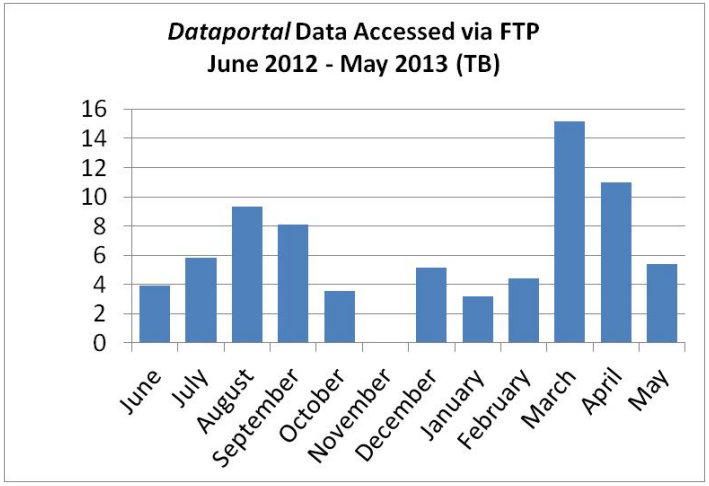


Dataportal Utilization – Data Accessed



Download Mechanism	Data Accessed (TB)			Available Data (TB)
	March	April	May	
Web	11.7	8.3	9.8	89
ESGF	22.7	11.6	29.1	153
FTP	15.1	11.0	5.4	143





Some Discover Updates Slides (Intel Sandy Bridge and Intel Phi MIC) from September 25, 2012 NCCS User Forum

Discover SCU8 Sandy Bridge: AVX

- The Sandy Bridge processor family features:

  Intel Advanced Vector eXtensions

- Intel AVX is a wider, new 256-bit instruction set extension to Intel SSE (Streaming 128-bit SIMD Extensions), hence higher peak FLOPS with good power efficiency.

- Designed for applications that are floating point intensive.

Discover SCU8 Sandy Bridge: User Changes

- Compiler flags to take advantage of Intel AVX (for Intel compilers 11.1 and up)

  -xavx:

  - Generate an optimized executable that runs on the Sandy Bridge processors ONLY

  -axavx –xsse4.2:

  - Generate an executable that runs on any SSE4.2 compatible processors but with additional specialized code path optimized for AVX compatible processors (i.e., run on all Discover processors)
  - Application performance is affected slightly compared to with "-xavx" due to the run-time checks needed to determine which code path to use

Sandy Bridge *vs.* Westmere: Application Performance Comparison – Preliminary

Sandy Bridge Execution Speedup Compared to Westmere				
WRF NMM 4km	**Same executable**		**Different executable (compiled with –xavx on Sandy Bridge)**	
	Core to Core	**Node to Node**	**Core to Core**	**Node to Node**
	1.15	**1.50**	**1.35**	**1.80**
GEOS5 GCM half degree	**Same executable**		**Different executable (compiled with –xavx on Sandy Bridge)**	
	Core to Core	**Node to Node**	**Core to Core**	**Node to Node**
	1.23	**1.64**	**1.26**	**1.68**

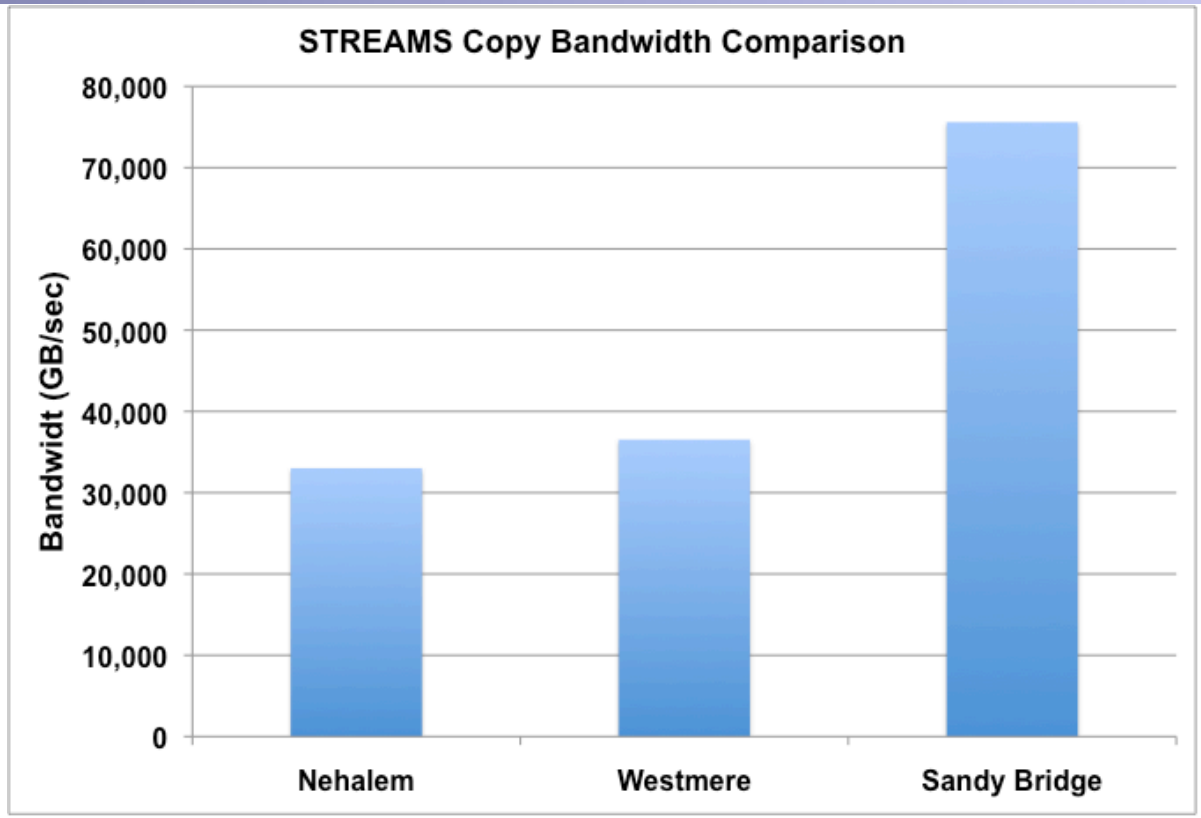
Discover SCU8 – Sandy Bridge Nodes

- 480 IBM iDataPlex Nodes, each configured with
  - Dual Intel SandyBridge 2.6 GHz processors (E5-2670) 20 MB Cache
  - 16 cores per node (8 cores per socket)
  - 32 GB of RAM (maintain ratio of 2 GB/core)
  - 8 floating point operations per clock cycle
  - Quad Data Rate Infiniband
  - SLES11 SP1
- Advanced Vector Extensions (AVX)
  - New instruction set (http://software.intel.com/en-us/avx/)
  - Just have to recompile

Discover SCU8 – Many Integrated Cores (MIC)

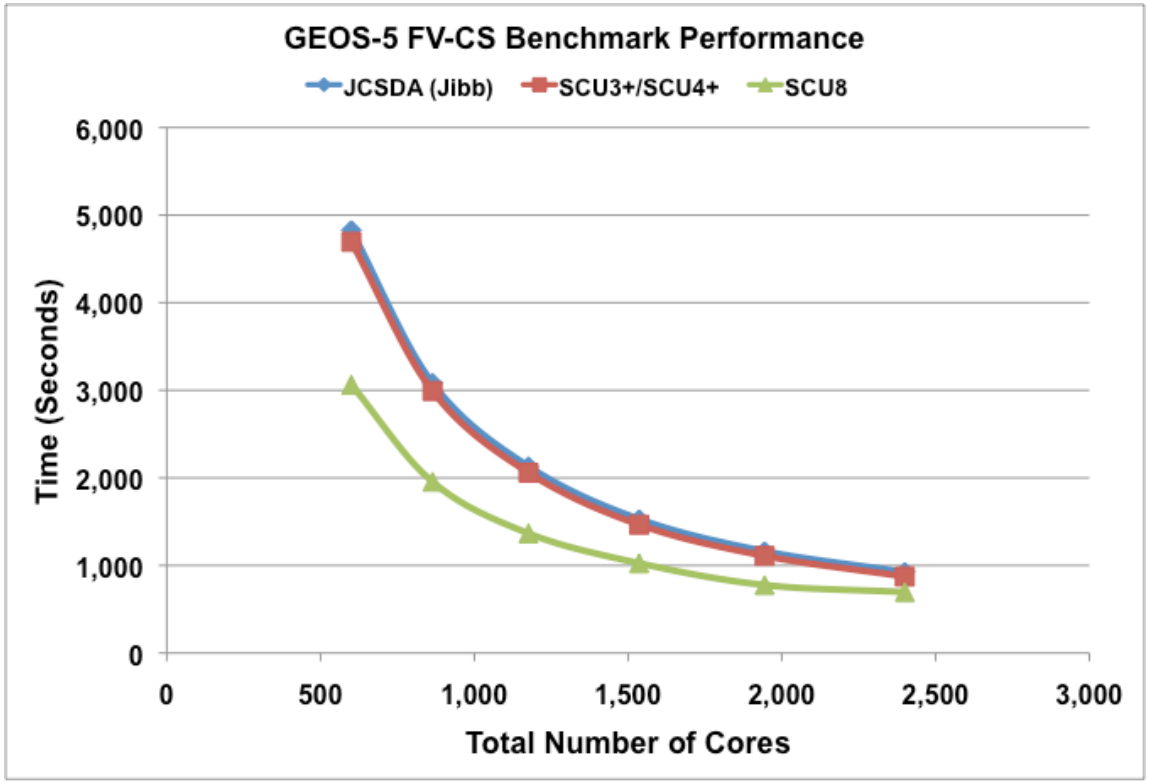
- The NCCS will be integrating 240 Intel MIC Processors later this year (October)
  - ~1 TFLOP per co-processor unit
  - PCI-E Gen3 connected
  - Will start with 1 per node in half of SCU8
- How do you program for the MIC?
  - Full suite of Intel Compilers
  - Doris Pan and Hamid Oloso have access to a prototype version and have developed experience over the past 6 months or so
  - Different usage modes; common ones are "offload" and "native"
  - Expectation: Significant performance gain for highly parallel, highly vectorizable applications
  - Easier code porting using native mode, but potential for better performance using offload mode
  - NCCS/SSSO will host Brown Bags and training sessions soon!

Sandy Bridge Memory Bandwidth Performance



- STREAMS Copy Benchmark comparison of the last three processors
  - Nehalem (8 cores/node)
  - Westmere (12 cores/node)
  - SandyBridge (16 cores/node)

SCU8 Finite Volume Cubed-Sphere Performance



JCSDA (Jibb): Westmere

Discover SCU3+/SCU4+: Westmere

Discover SCU8: Sandy Bridge

- Comparison of the performance of the GEOS-5 FV-CS Benchmark 4 shows an improvement of 1.3x to 1.5x over the previous systems' processors.

Discover: Large "nobackup" augmentation

- Discover NOBACKUP Disk Expansion
  - 5.4 Petabytes RAW (about 4 Petabytes usable)
  - Doubles the disk capacity in Discover NOBACKUP
  - NetApp 5400
    - http://www.netapp.com/us/products/storage-systems/e5400/
    - 3 racks and 6 controller pairs (2 per rack)
    - 1,800 by 3 TB disk drives (near line SAS)
    - 48 by 8 GB FC connections
- Have performed a significant amount of performance testing on these systems
- First file systems to go live this week
- If you need some space or have an outstanding request waiting, please let us know (email support@nccs.nasa.gov).