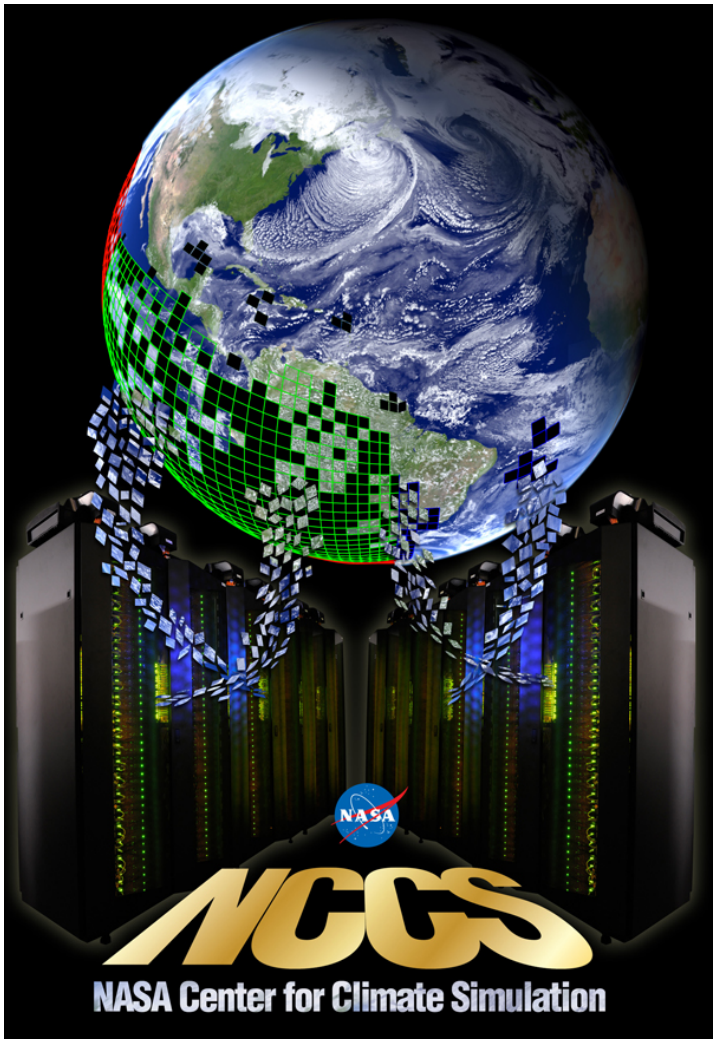


NCCS User Forum

March 5, 2013

Introduction and Special Announcement

- Lynn Parnell
  - Has served as the HPC Lead for the past 3 years
  - Will move into supporting special projects for the NCCS and CISTO
  - Thank You Lynn!

- Daniel Duffy
  - Will be the new HPC Lead and will continue serving as the NCCS lead architect (for now)

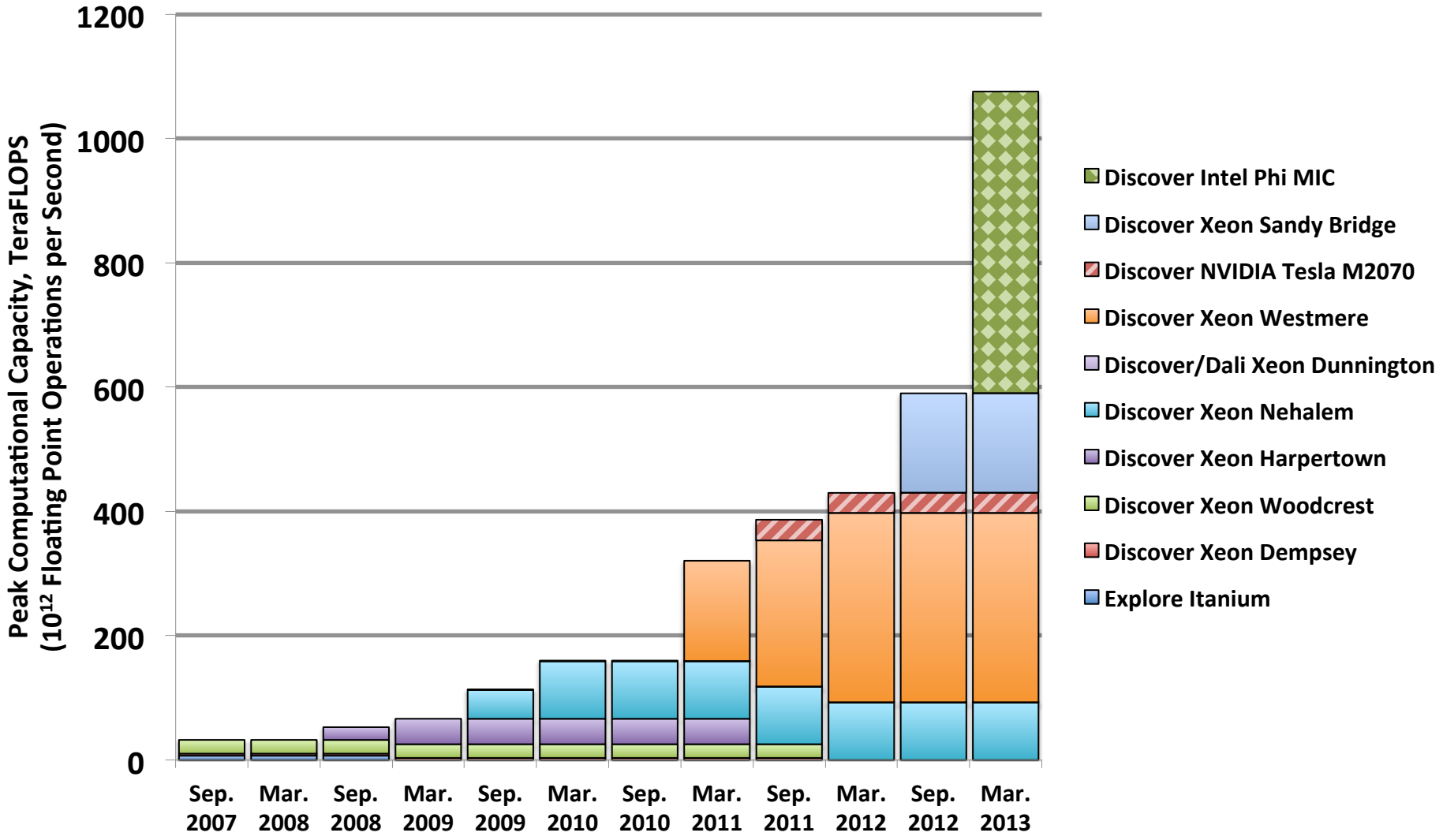
Agenda

- Introduction
- Discover Updates
- Results & Responses to 1st Annual NCCS User Survey
- NCCS Operations & User Services Updates
- Question & Answer

Recent Accomplishments

- Discover SCU8 Installation
  - Intel Xeon SandyBridge – 480 nodes each with one Intel Phi coprocessor
  - SCU8 is #31 on the Green500 (www.green500.org)
  - SCU8 is #53 on the Top500 (www.top500.com)
- Discover Disk Capacity
  - Added 4 Petabytes (usable) last year
  - Taken major strides on making this system ready for operations
- Brown Bag seminars (more to come)
- NCCS User Survey
  - First annual survey results are in and will be discussed in subsequent slides
- Earth System Grid Federation (ESGF) downloads
  - Over 275 TB and 9 million data sets, April 2011 – February 2013

NCCS Compute Capacity Evolution September 2007- March 2013



Staff Additions

Welcome to
Dave Kemeza, System Administrator

Discover Update

Dan Duffy,
HPC Lead and NCCS Lead Architect

Discover Updates

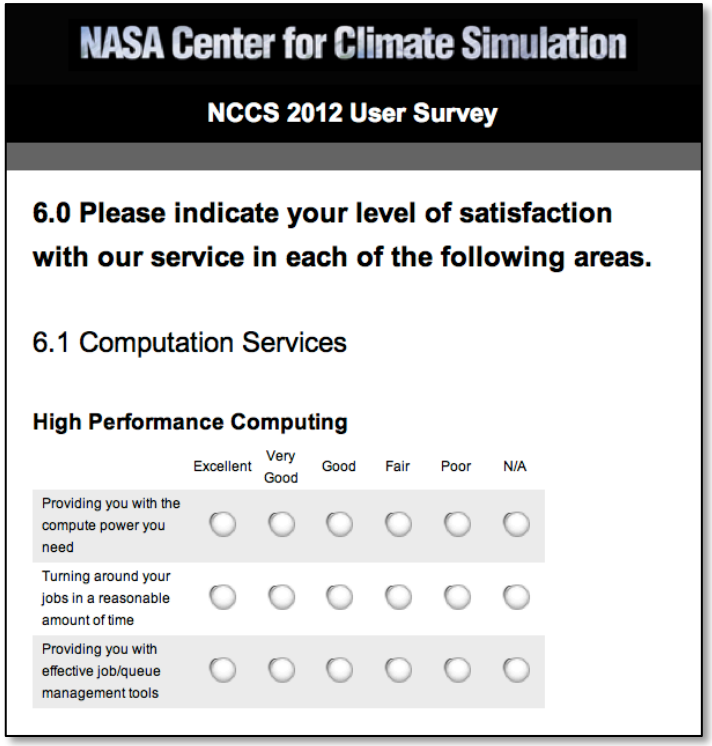
- SCU8 status
  - Pioneer users currently have access
  - General access is planned soon
  - A test queue will be set up for native application porting to the Intel Phi
- Next Compute Upgrade
  - In planning and preparation
  - Budgetary constraints
  - Possibilities include additional Intel Phi coprocessors, more compute nodes, or a combination of both
- GPFS Metadata Storage Upgrade
  - Goal is to dramatically increase the aggregate metadata performance for the GPFS metadata
  - Looking at very high speed storage subsystems to be installed later this year

Results & Responses to 1st Annual NCCS User Survey

Dan Duffy

First Annual NCCS User Survey, October 2012

- Ten-minute online survey via SurveyMonkey.
- Provides a more systematic way for NCCS to gauge what's working well for you, and what needs more work.
- We intend to repeat survey annually so we can evaluate progress.



Sample NCCS User Survey screen.

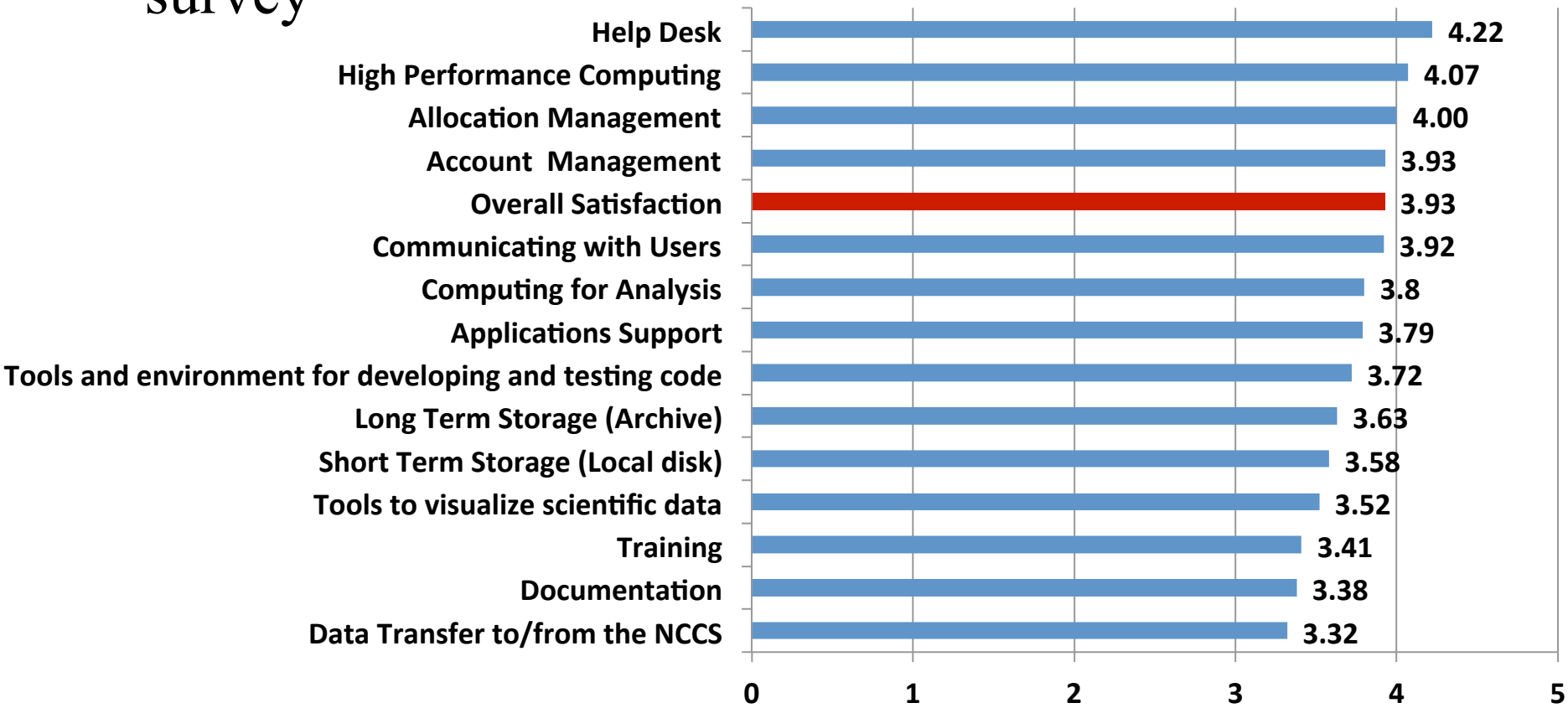
Thank you very much for your frank opinions and suggestions!

Fall 2012 NCCS User Survey Results

- Excellent feedback from the user community
  - 94 responses
- Overall Satisfaction
  - 3.93 out of 5
  - Slightly under a "Very Good"
- Download the report at
  - http://www.nccs.nasa.gov/news.html#survey

Detailed Results

- The following chart shows some of the details of the report and the rankings in order based on area of the survey



Aspects that are Outstanding

- 46 responses detailed aspects of the NCCS that were outstanding, including
  - Support/Help Desk (~30)
  - Compute Power (~10)
  - Availability/Stability
  - Disk Storage
  - Unlimited Archive
  - Training

- Also, a number of individuals were called out for going the extra mile!

Where are improvements needed?

- NCCS has decided to focus on the following three main areas based on the user survey responses.
  1. External data transfer performance and convenience
  2. More timely notifications of problems or unplanned outages
  3. "Architecting for More Resiliency," especially the Discover storage file systems

External Data Transfer

- Problem Statement
  - ***How can the NCCS improve data transfer between Discover and other systems, including the Data Portal?***
- Actions Taken So Far
  - Reached out to users for additional feedback
    - Data Portal 1 GbE limitation was a definite bottleneck
    - Provided 'proxy mode' info for connecting to systems and moving data (new info for some users)
      - http://www.nccs.nasa.gov/primer/getstarted.html#proxy
  - Identified and corrected a number of internal NCCS network issues
    - Parameter correction for several Discover systems with 10 GbE interfaces
    - Upgraded the NCCS Firewall operating systems which improved file transfer performance
    - Identified and addressed operating system and configuration issues with 10 GbE core switch

External Data Transfer (Continued)

- Pending Actions
  - Add 10 GbE interfaces to the Data Portal
  - Develop and publish best practices for moving data between systems
  - Upgrade NCCS/SEN/CNE network connectivity to 10 GbE
  - Run application performance tests between NCCS and external systems (e.g., NAS)
- Comments/Requests/Questions
  - Contact NCCS User Services: support@nccs.nasa.gov

More Timely Problem Notifications

- Problem Statement
  - ***How can the NCCS communicate problem notifications to the user community in a more timely and efficient manner?***
- Actions Taken So Far
  - Message of the Day (MOTD) elevated to high visibility on the NCCS web page
  - NCCS is more diligent on updating the MOTD
- Pending Actions
  - Modify internal processes for notifying users during issues
  - Creation of a web dashboard for system status
  - Creation of an overall NCCS Communications Plan
- Comments/Requests/Questions
  - Contact NCCS User Services: support@nccs.nasa.gov

Revise System Architecture for Increased Resiliency

- Problem Statement
  - ***How can the NCCS (specifically the Discover system) be architected for increased availability and resiliency?***
- Actions Taken So Far
  - Internal discussions and resetting of priorities
- Pending Actions
  - Creation of a special architecture group to review the current architecture and discuss options
- Comments/Requests/Questions
  - Contact NCCS User Services: support@nccs.nasa.gov

NCCS Operations & User Services Update

Ellen Salmon

- SMD Requests for HPC Allocations
- Upcoming & Status
- Ongoing Investigations
- In-Depth Training
- NCCS Brown-Bag and SSSO Seminars

Spring 2013 Call for Applications for NASA SMD HPC Resources

- Science Mission Directorate's (SMD) half-yearly call for applications for significant allocations for NASA High End Computing resources:

  Deadline: March 20, 2013.

- Apply for:
  - New allocations.
  - Significant increase in "November-Cycle" allocations.
  - Renewed allocations for "Spring-Cycle" HPC awards.
- Details coming soon to:

  http://www.hec.nasa.gov/request/announcements.html

Upcoming (1 of 2)

- Discover resources:
  - SCU8 Sandy Bridge:
    - Significant work (planar replacements; risers & Phi coprocessors installed, …).
    - Now available to pioneer users via test queue and some special-purpose queues.
    - Contact NCCS User Services (support@nccs.nasa.gov) for access to the test queue.
  - SCU8 Intel Phi Coprocessors (MICs), now 480 units, one per SCU8 node:
    - "**Offload mode**" available now on SCU8 nodes **(requires use of Intel 13 compiler)**.
    - "**Native mode**" testing coming soon, in a special queue (e.g., "warp-phi").
    - *More user training in progress...*
    - Want help porting code to the Intel Phis? Contact support@nccs.nasa.gov.
  - "NetApp" nobackup (4+ Petabytes):
    - Modifications for new architecture array, plus new massively parallel I/O workloads:
      - GPFS upgrade & configuration, new I/O servers for high-activity filesystems, …
    - Have resumed deploying additional NetApp nobackup in a measured fashion.
    - Moving nobackup directories to disk array types best suited for their workloads.

Upcoming (2 of 2)

- Discover Software Changes:
  - Rolling migration to required, upgraded, InfiniBand OFED (software stack).
    - All of SCU8 is already on the new OFED release.
    - Rolling, announced, gradual changeovers of other parts of Discover (e.g., via PBS queues or properties).
    - ***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):
  - None scheduled at the moment.

Archive & Data Portal Status

- Archive:
  - Default: single tape copy since ~June 2012.
  - Continuing migration from 2nd tape copies to reclaim & reuse those tapes.
  - Reminder: request 2 tape copies if warranted:

    dmtag –t 2 <filename>

- Data Portal:
  - Developing plans for additional storage and servers to meet demand for additional NCCS data services.

Ongoing Discover Investigations (1)

GPFS and I/O

- GPFS hangs due to new disk array and new parallel I/O activity.
  - Significant work during February 26 downtime to address issues, so expecting improvement:
    - GPFS upgrade and new parameter settings.
    - Four additional dedicated I/O servers for filesystems with heavy parallel I/O activity.
- Heavy GPFS metadata workloads.
  - Investigating hardware options.
  - Target: improve responsiveness in handling many concurrent small, random I/O actions (e.g., for directories, filenames, etc.).

Ongoing Discover Investigations (2)

- "Ghost Jobs" on PBS
  - Job "executes successfully" (return code 0), but no work was done.
  - Although your job script is copied to the job's "head node," PBS appears to execute a "null" (empty) job script and so completes immediately.
  - *Affected head nodes will repeat this behavior until "some" change occurs…*
  - NCCS staff is working closely with Altair (PBS vendor).
    - PBS does not "know" there is a problem: difficult to automate detection & capture pertinent data!
  - Please contact support@nccs.nasa.gov ***with job id number*** right away (call 301-286-9120 if during business hours) to help us track down this problem.
- Intermittent qstat, qsub delays (PBS slowness).
  - Problem severity escalated with Altair, the PBS vendor.
- Jobs exhausting available node memory, causing GPFS hangs.
  - Continuing to refine automated monitoring, PBS slowness complicates this.

In-Depth Training Opportunities

- March 19-21 (3 full days), SSSO Seminar Series:
  Advanced Scientific Programming
  - Target: scientist programmers involved in development of large, complex numerical models in FORTRAN.
  - Limited number of seats; see email from SSSO's Tom Clune for more details.

- April 11 (1 full day), potentially additional workshops (to be announced):
  Intel's Focused Workshops for NCCS Users, potential topics:
    - Parallel Studio XE, including
      - Using VTune™ Amplifier XE (Performance Analyzer)
      - Using Parallel Inspector XE (Memory and Thread Checker)
    - Additional Intel Phi Coprocessor (MIC) topics
    - Vectorization
    - Performance Libraries, including IPP (Performance Primitives)
  - Details & agenda are forthcoming.
  - Contact support@nccs.nasa.gov if you have particular areas of interest.

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
  - Intel MPI on Intel Phi
  - Advanced Offload Techniques for Intel Phi
  - Maximum Vectorization
  - Performance Analysis via the VTune™ Amplifier
  - Performance Tuning for the Intel Phi

SSSO Seminar Series

- Current series:

  Python Programming for Data Processing and Climate Analysis

- ~Every other Monday at 1:30 p.m. here in Building 33, Room H114.
- Next up, March 11, 2013:

  Session (2):

  Array and Matrix Manipulation with NumPy and Mathematical Functions with SciPy

- Downloadable content and details on upcoming sessions are available on SSSO's Modeling Guru web site, currently at:

  https://modelingguru.nasa.gov/docs/DOC-2322

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

Resolved Issues

- Dirac (/archive): occasional hangs in NFS exports to Discover.
  - Implemented an automated workaround while awaiting vendor patch.
    - Greatly reduced the problem's impact.
  - Installed vendor patch on all eNFS nodes, completed February 26.

- Data Portal: fixed problem with one of four disk arrays.
  - Temporarily moved data to separate storage.
  - Applied the necessary disk array microcode updates.
  - Performed file "sanity checking."
  - Moved the data back to its original location.

NCCS Brown Bag Seminars: ❑ Proposed Topics

Intel Phi MIC Coprocessor Tutorials:

- ✓ Intel Many Integrated Core (MIC) Prototype Experiences
- ✓ Intel Phi Coprocessors: an Introduction to Programming Models
- ❑ Intel MPI on Intel Phi
- ❑ Advanced Offload Techniques for Intel Phi
- ❑ Maximum Vectorization
- ❑ Performance Analysis with VTune™ Amplifier
- ❑ Performance Tuning for Intel Phi

Other Potential Brown Bag Topics:

- ❑ Using PODS (Portable Distributed Scripts) on Discover
- ❑ Introduction to iRODS
- ❑ Using GNU Octave
- ❑ Introduction to Using Matlab with GPUs
- ❑ Best Practices for Using Matlab with GPUs
- ❑ Using CUDA with NVIDIA GPUs
- ❑ Using OpenCL
- ❑ Using Database Filesystems for Many Small Files

NCCS Brown Bag Seminars: ✔Delivered Topics

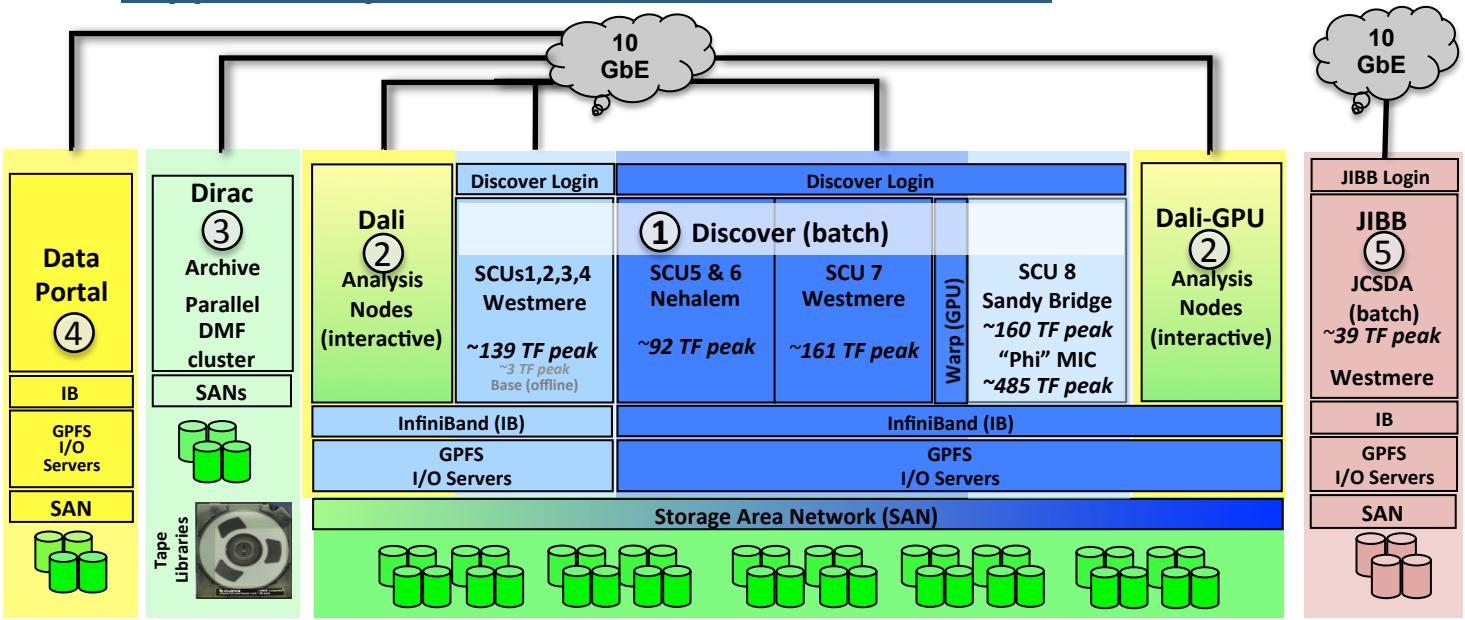
- ✓ Intel Many Integrated Core (MIC) Prototype Experiences
- ✓ Intel Phi Coprocessors: an Introduction to Programming Models
- ✓ Introduction to the NCCS Discover Environment
- ✓ Tips for Monitoring PBS Jobs and Memory
- ✓ Detecting Memory Leaks with Valgrind
- ✓ Intro. to Using Python Matplotlib with Satellite Data
- ✓ Code Debugging Using TotalView on Discover, Parts 1 and 2
- ✓ Code Optimization Using the TAU Profiling Tool
- ✓ Climate Data Analysis & Visualization using UVCDAT
- ✓ NCCS Archive Usage & Best Practices Tips, & dmtag
- ✓ Using winscp with Discover, Dali, and Dirac
- ✓ SIVO-PyD Python Distribution for Scientific Data Analysis

NASA Center for Climate Simulation Supercomputing Environment

Supported by HQ's Science Mission Directorate

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

- Intel Xeon nodes
  - ~3,900 nodes
  - ~43,600 cores
  - ~557 TFLOPS peak
  - 96 TB memory (2 or 3 GB per core)
- Coprocessors:
  - Intel Phi MIC
    - 480 units
    - ~485 TFLOPS
  - NVIDIA GPUs
    - 64 units
    - ~33 TFLOPS
- Shared disk: 7.7 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
- ~60 PB robotic tape library
- Data Management Facility (DMF) space management

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

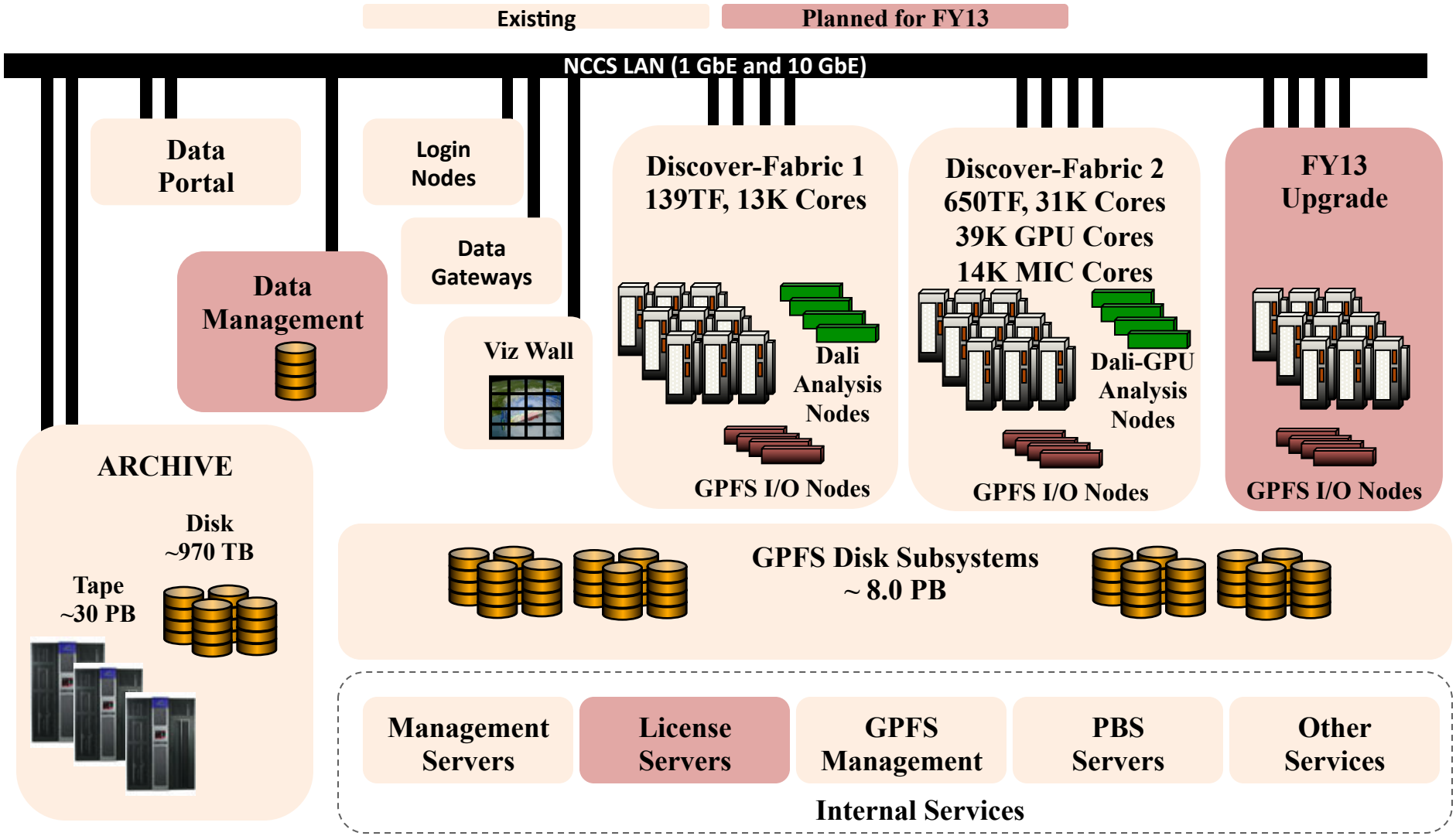
- Earth System Grid
- OPeNDAP
- Data download: http, https, ftp
- Web Mapping Services (WMS)server

⑤ ***JIBB***

- Linux cluster for Joint Center for Satellite Data Assimilation community

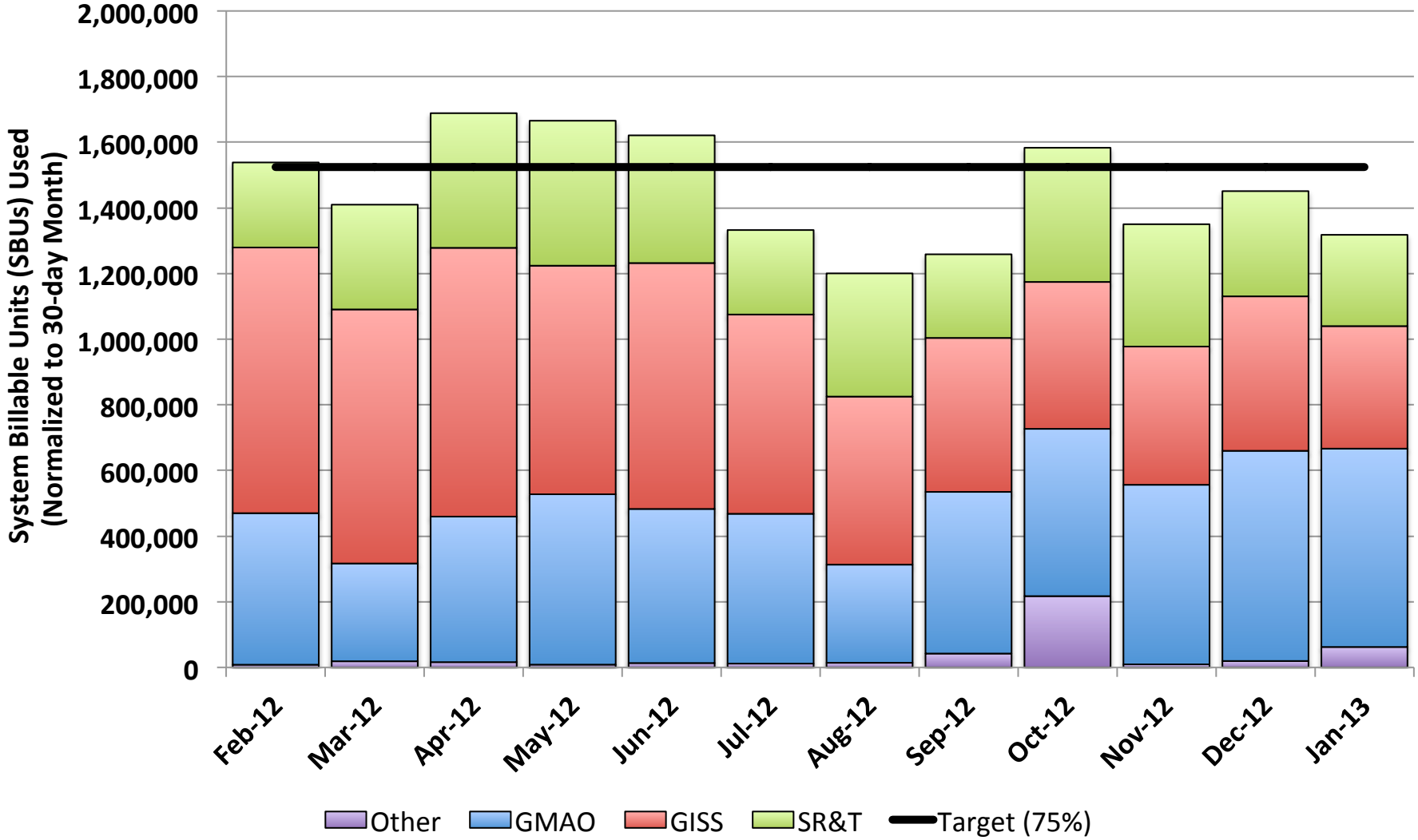
March 1, 2013

NCCS Architecture

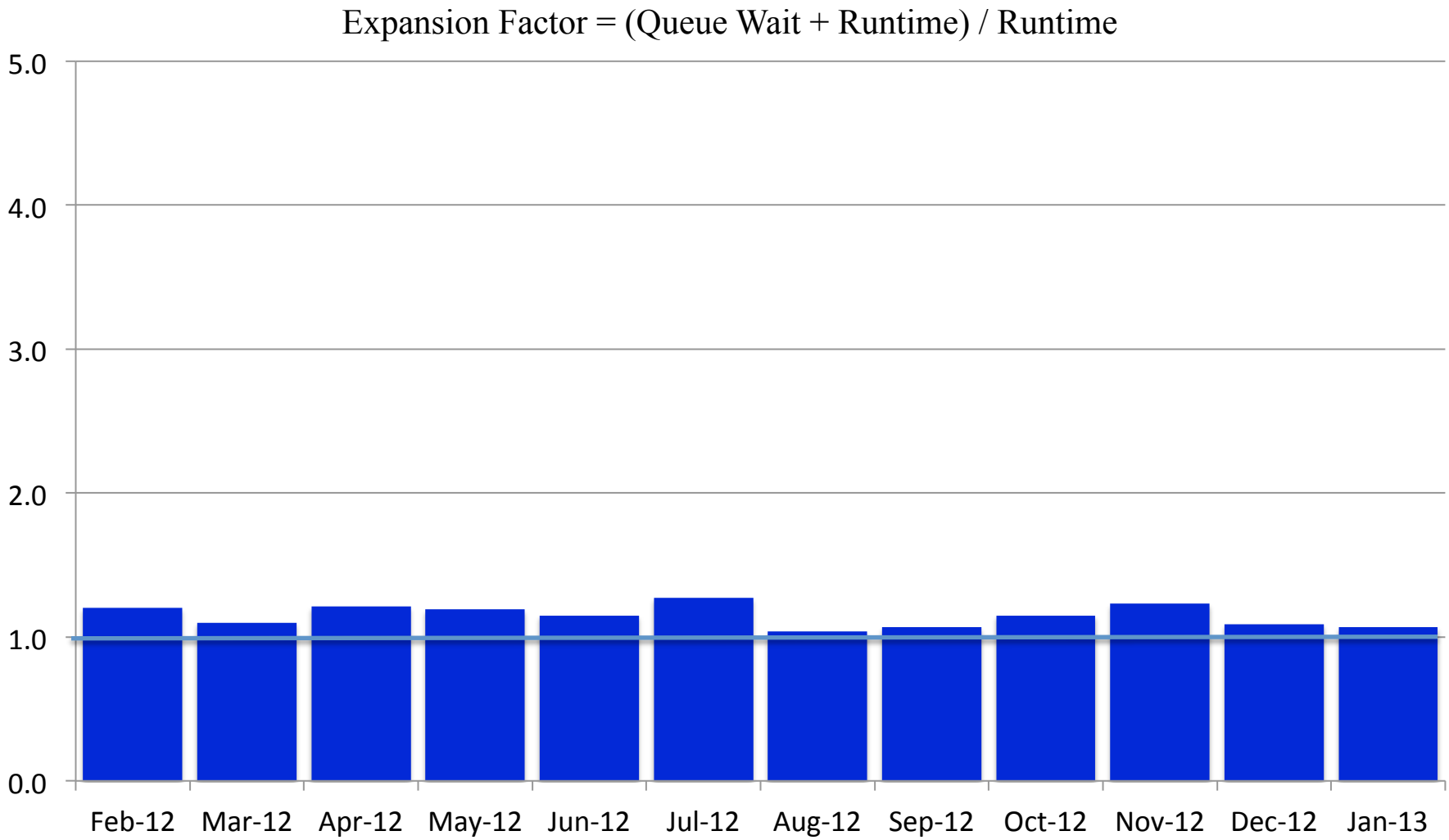


NCCS Metrics Slides
(Through January 31, 2013)

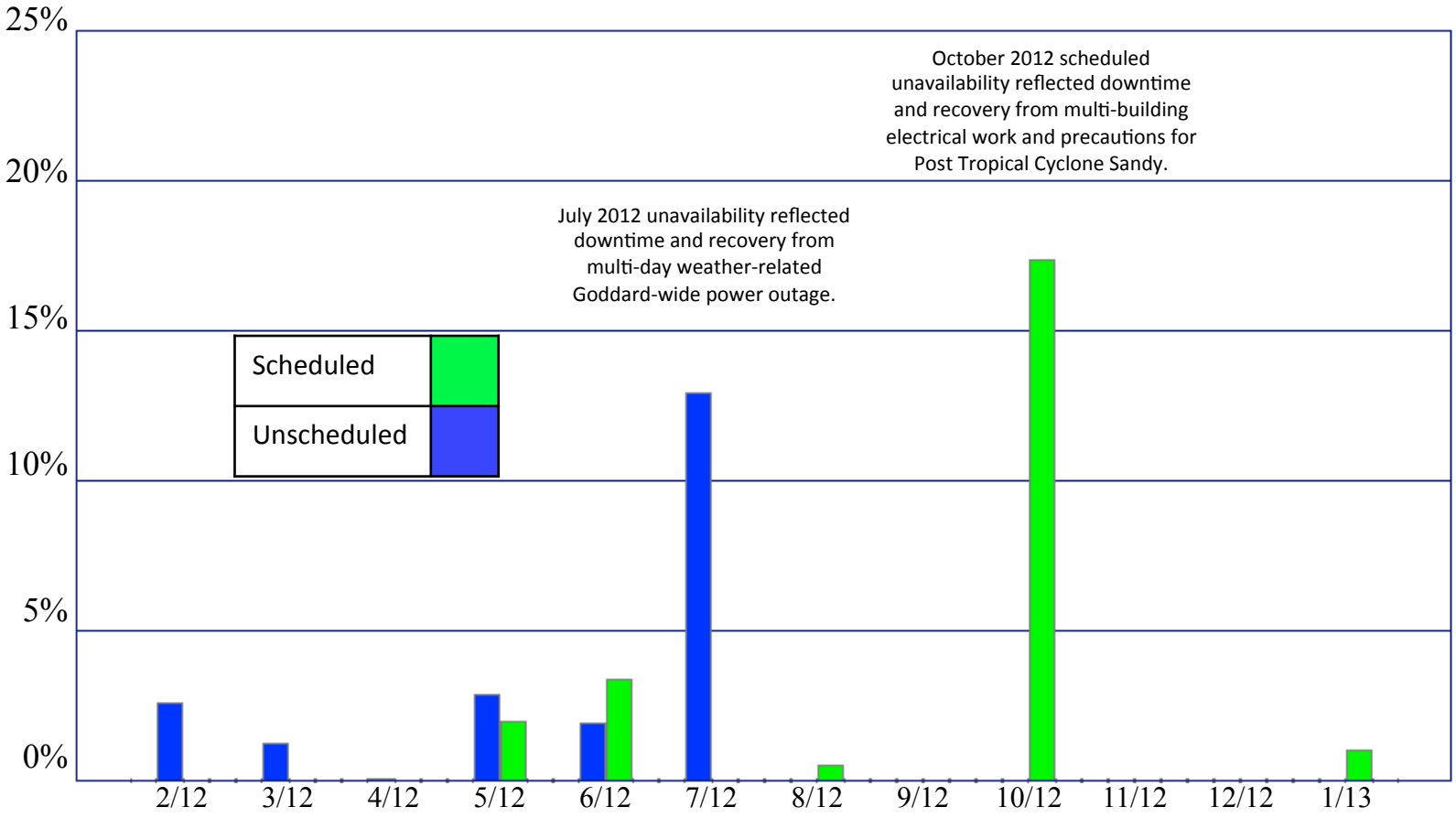
NCCS Discover Linux Cluster Utilization Normalized to 30-Day Month



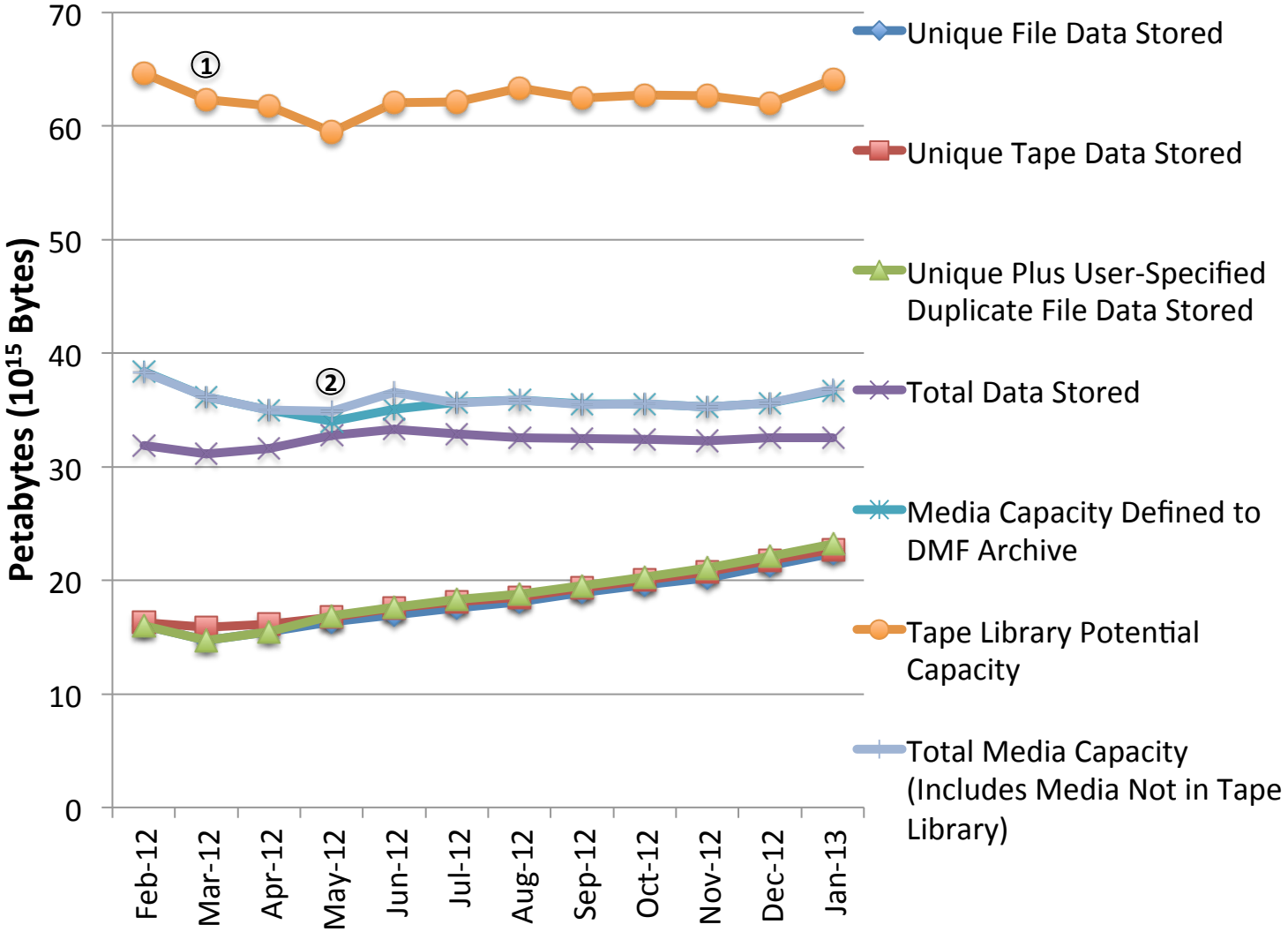
Discover Linux Cluster Expansion Factor



Discover Linux Cluster Downtime



NCCS Mass Storage



① "Tape Library Potential Capacity" decreased beginning in March 2012 because, at NCCS request, users began deleting unneeded data, and that tape capacity had not all been reclaimed so that new data could be written to the tapes.

② In 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.

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**

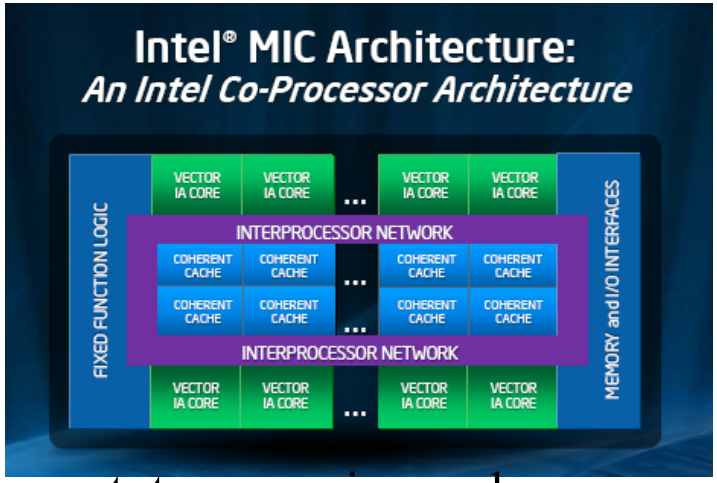
Discover SCU8 – Sandy Bridge Nodes

- It's Here!
- 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

- **Running some final system level tests**
- **Ready for pioneer users later this week**

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



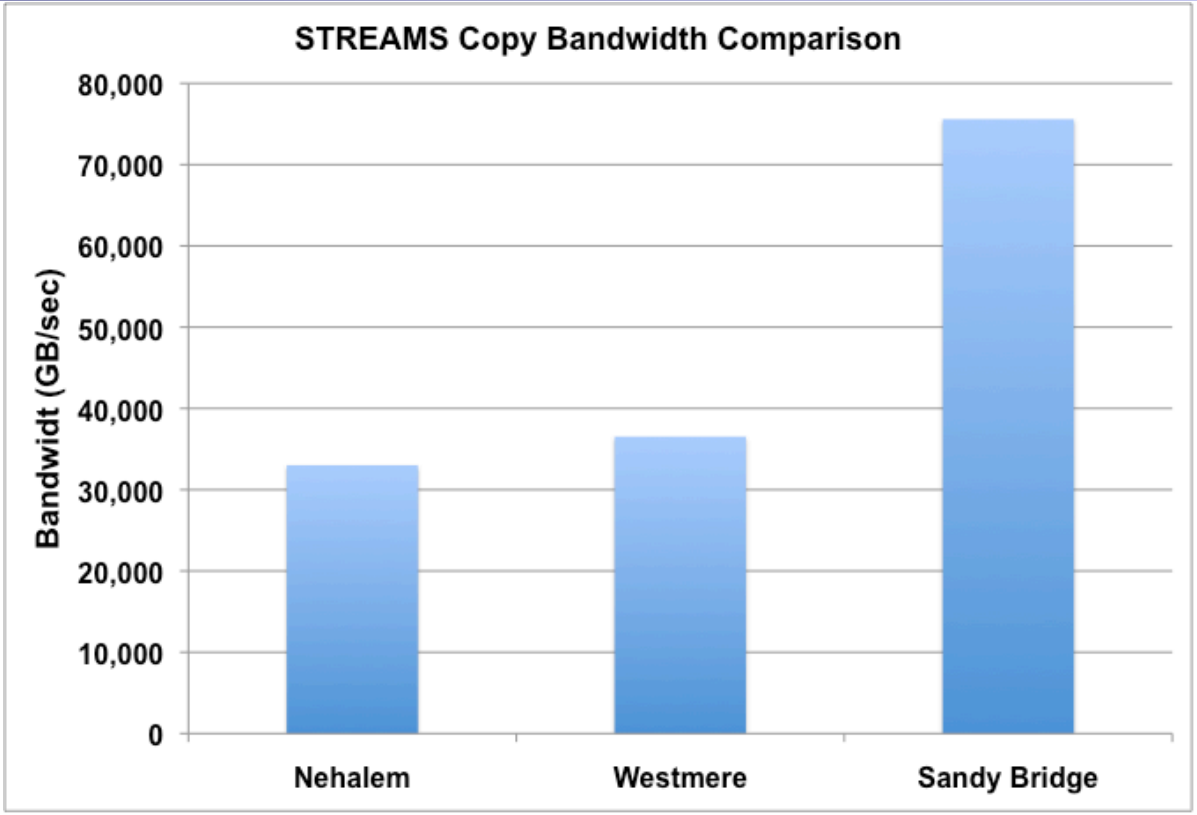
- How do you program for the MIC?
  - Full suite of Intel Compilers
  - Doris Pan and Hamid Oloso have access to a prototype version and have developed experience over the past 6 months or so
  - Different usage modes; common ones are "offload" and "native"
  - Expectation: Significant performance gain 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!

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

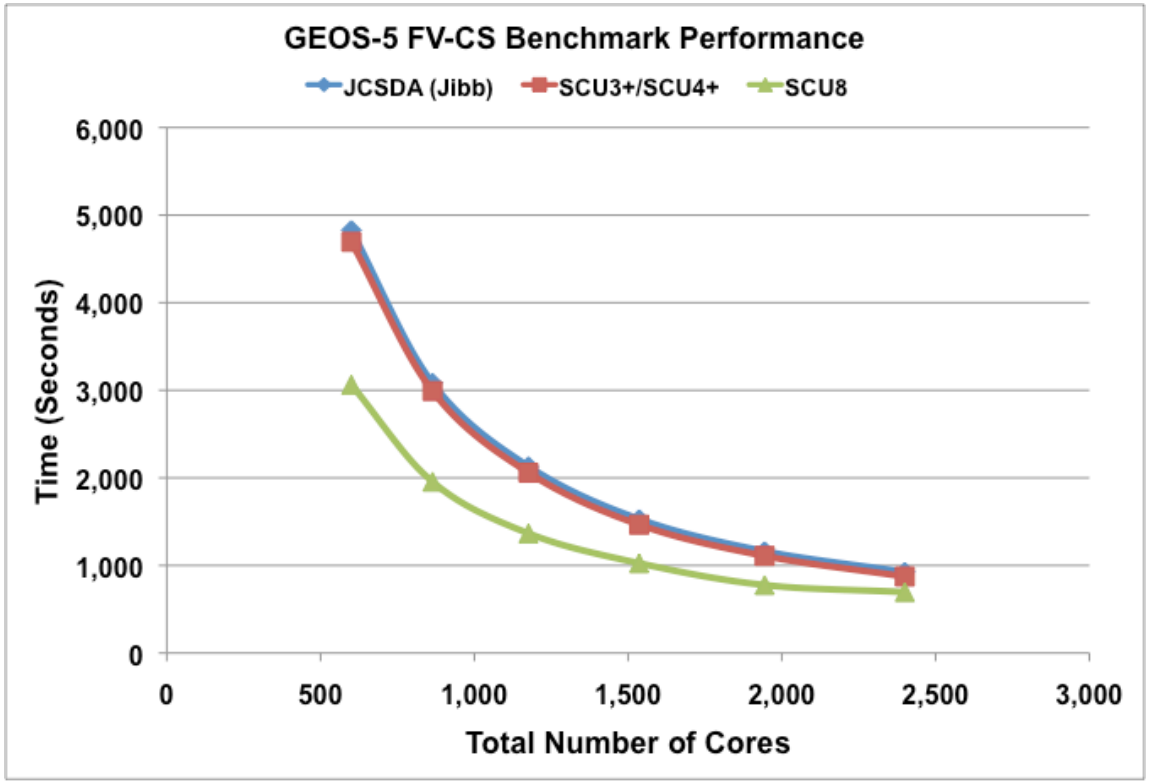
- Have performed a significant amount of performance testing on these systems
- First file systems to go live this week
- If you need some space or have an outstanding request waiting, please let us know (email support@nccs.nasa.gov).

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.