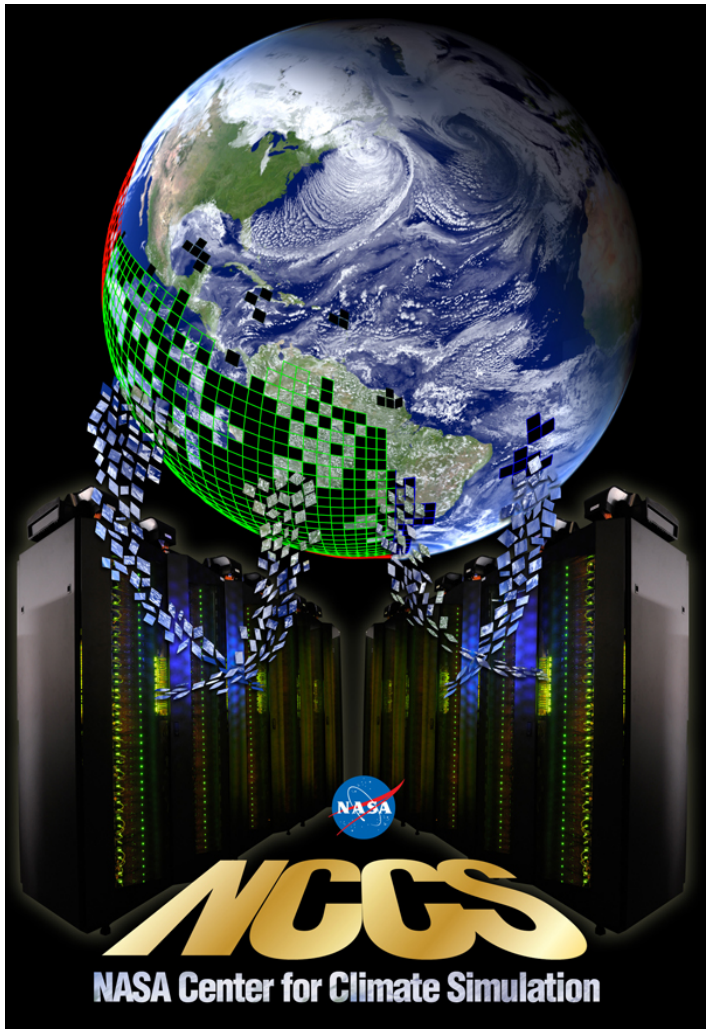


NCCS User Forum

September 25, 2012

Agenda

- Introduction
- Discover Updates
- NCCS Operations & User Services Updates
- Question & Answer

Accomplishments

- Discover SCU8 augmentation.
  - Intel Xeon Sandy Bridge: ~160 TFLOPs (480 nodes), pioneer users this week.
  - Many Integrated Core (MIC): expected October 2012 (240 units).
- ~4 Petabytes (usable) Discover nobackup disk.
  - Testing ongoing, starting to provide some project nobackup later this week.
- Ten NCCS Brown Bag seminars so far, more to come.
- NCCS Earth System Grid Federation (ESGF) Data Nodes: *over 141 TB and 6 million data sets served* (April 2011 to September 2012).
  - NASA's CMIP5/IPCC AR5 climate simulation contributions.
  - Obs4MIPS – observations formatted for use in climate/ocean/weather model intercomparison studies (e.g., CERES EBAF, TRMM, …).
  - ana4MIPS – analyses, initially MERRA monthly means.

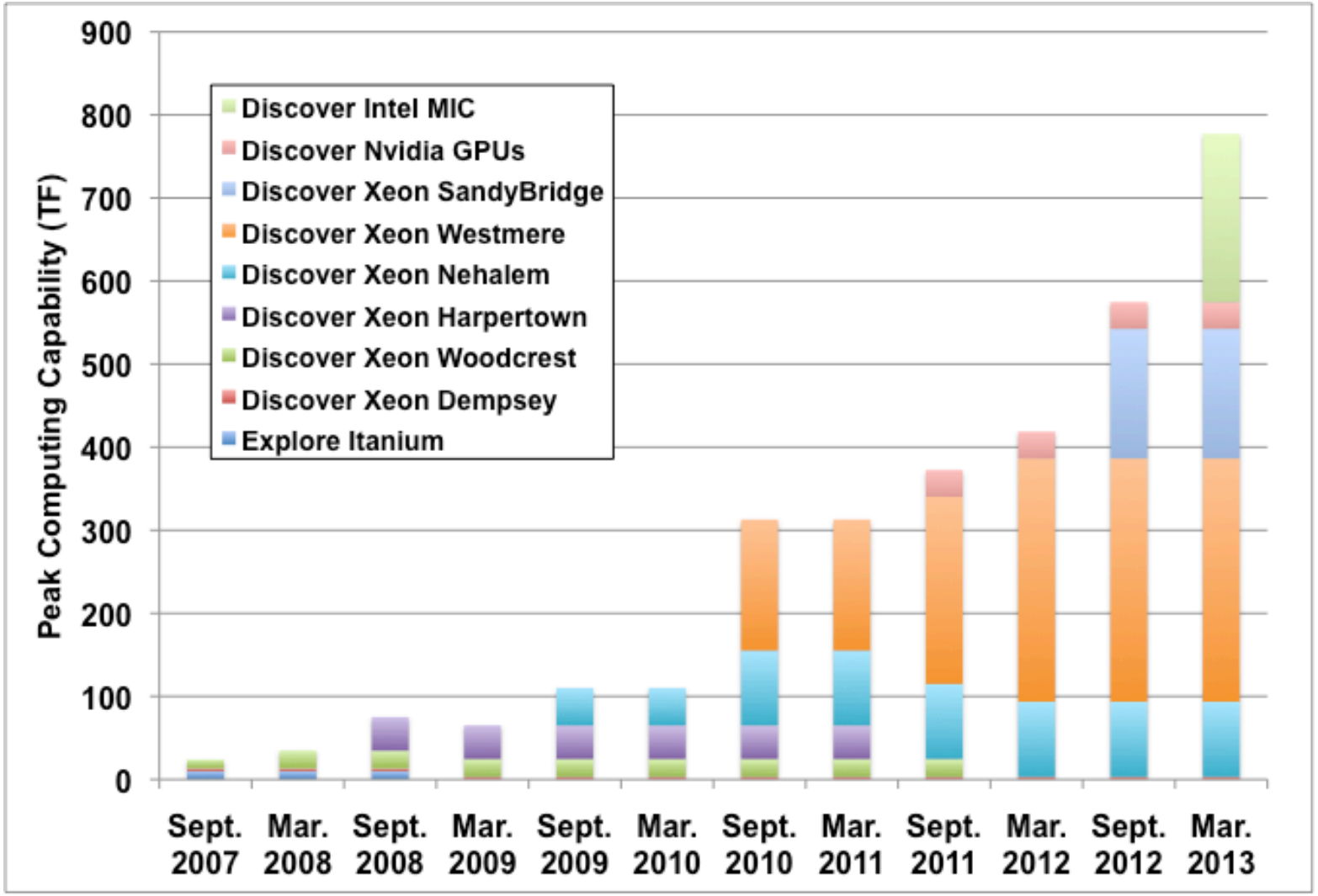
Staff Additions

Brandon Rives, Intern

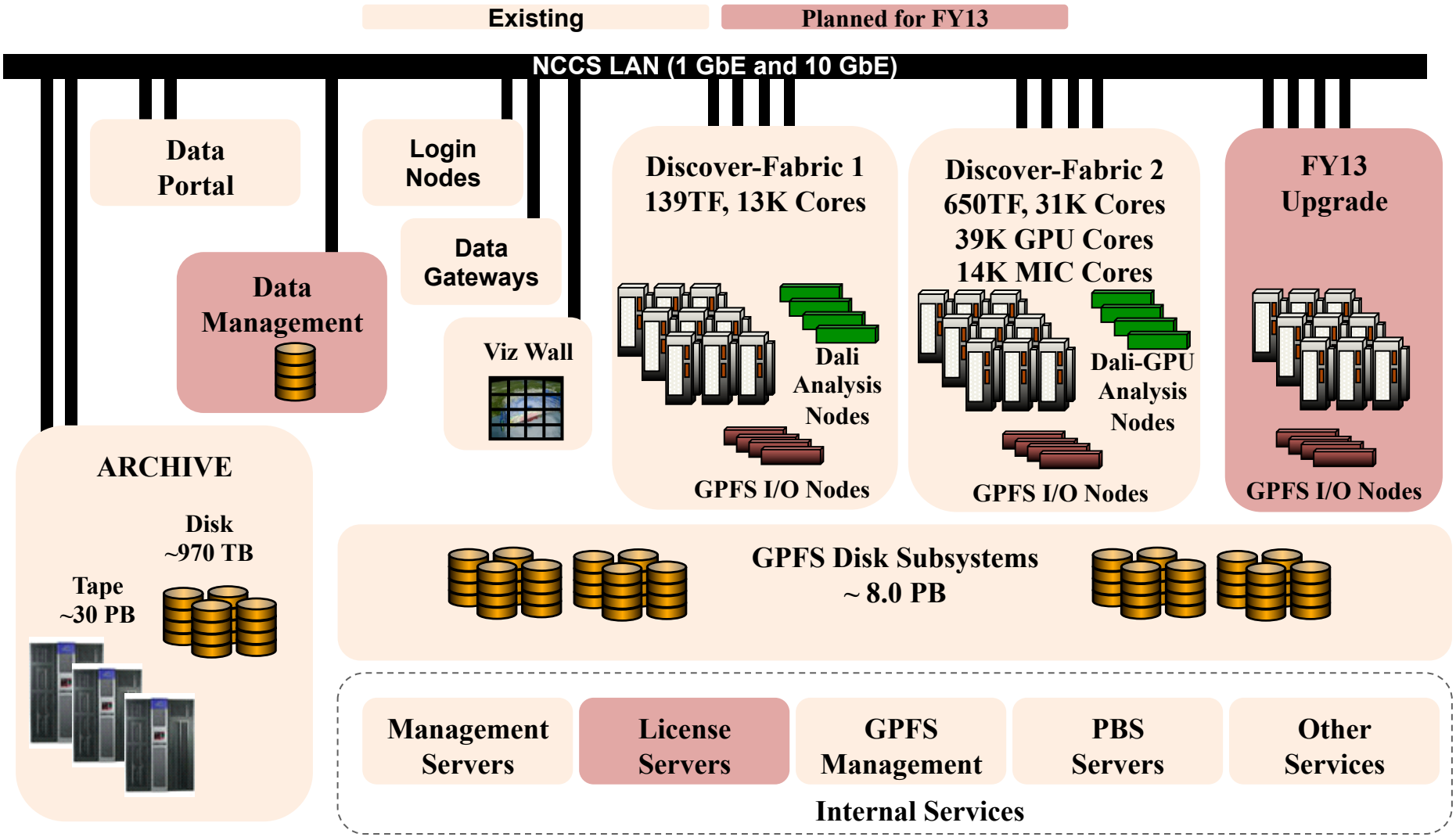
Discover Update

Dan Duffy, NCCS Lead Architect

NCCS Compute Capacity Evolution 2007-2013



NCCS Architecture



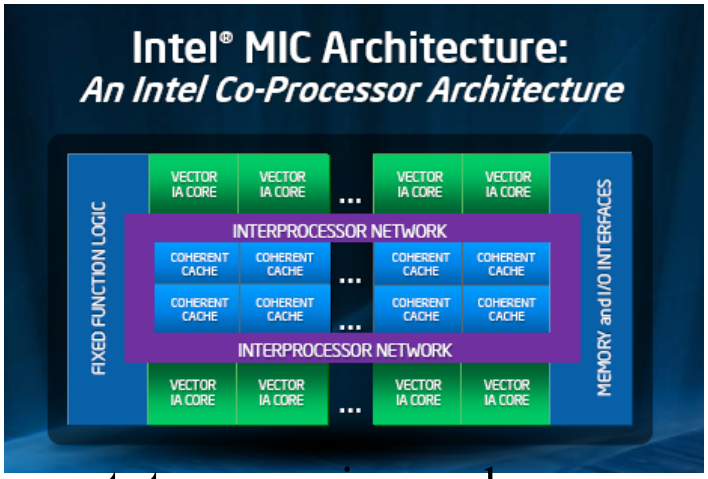
Discover SCU8 – Sandy Bridge Nodes

- 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**

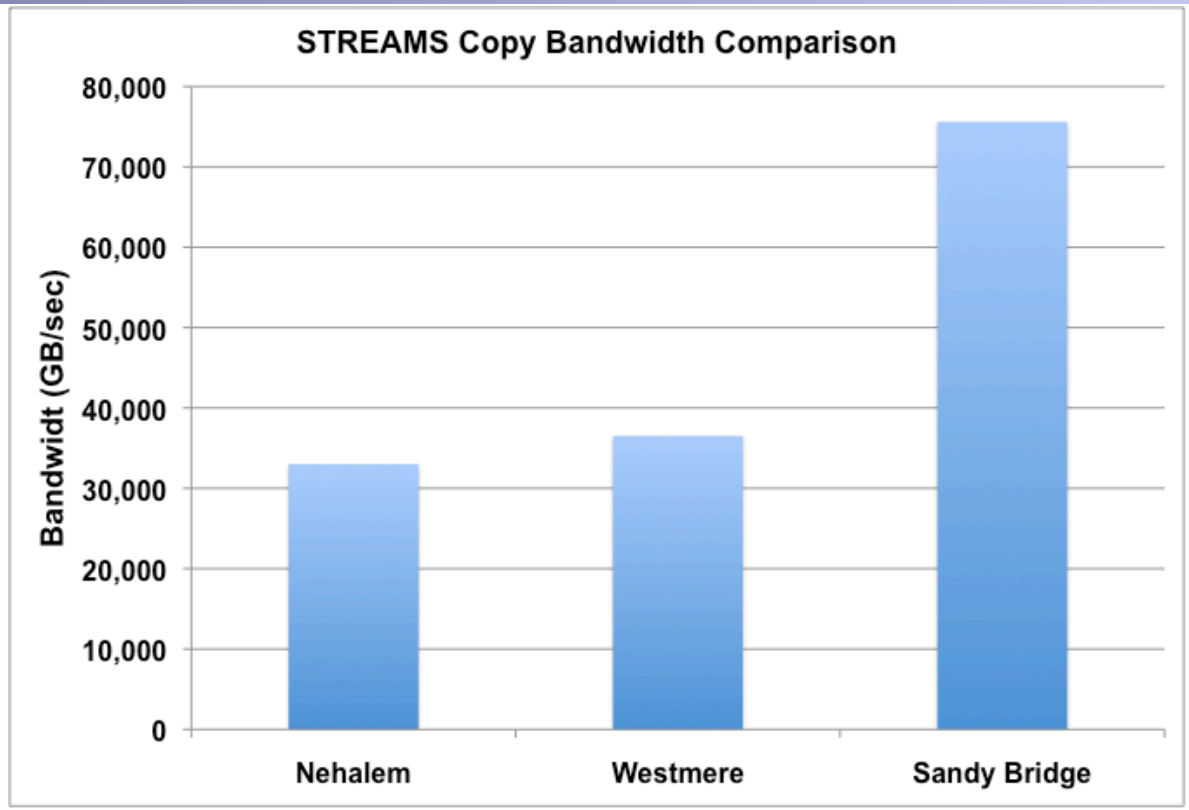
Discover SCU8 – Many Integrated Cores (MIC)

- 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



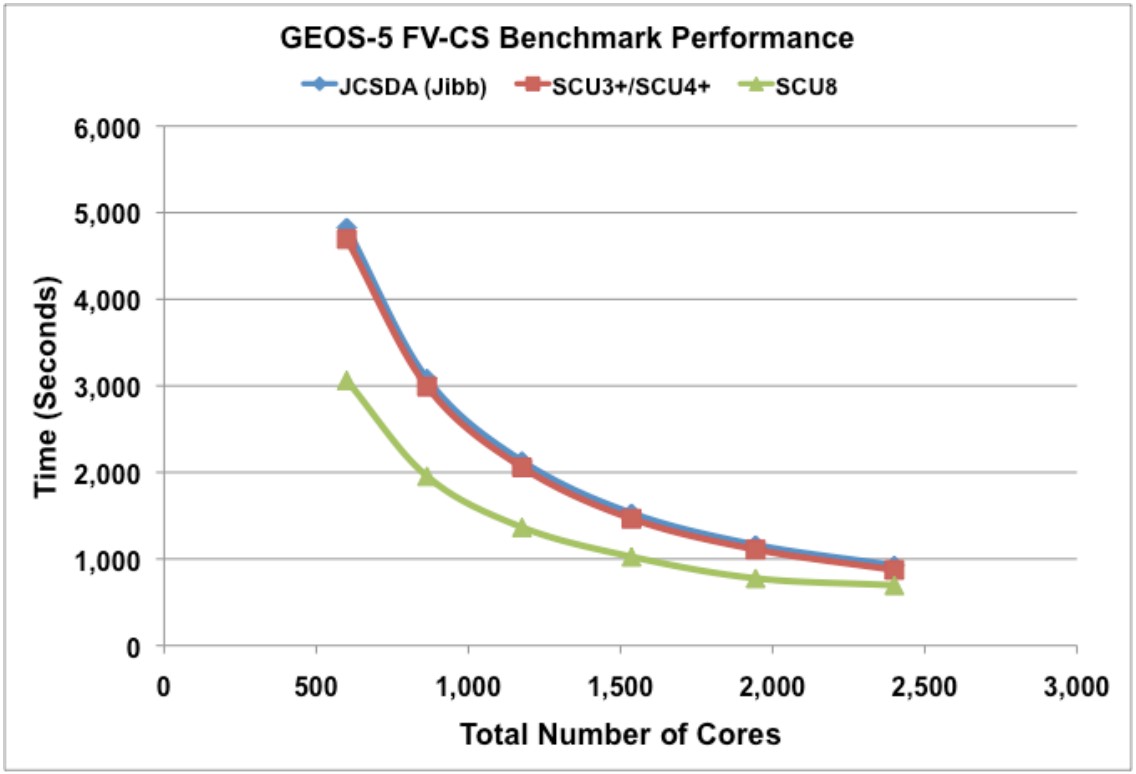
- How do you program for the MIC?
  - Full suite of Intel Compilers
  - Doris Pan and Hamid Oloso have access to a prototype version and have developed experience over the past 6 months or so
  - Different usage modes; common ones are "offload" and "native"
  - Expectation: Significant performance gain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

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).

NCCS Operations & User Services Update

Ellen Salmon

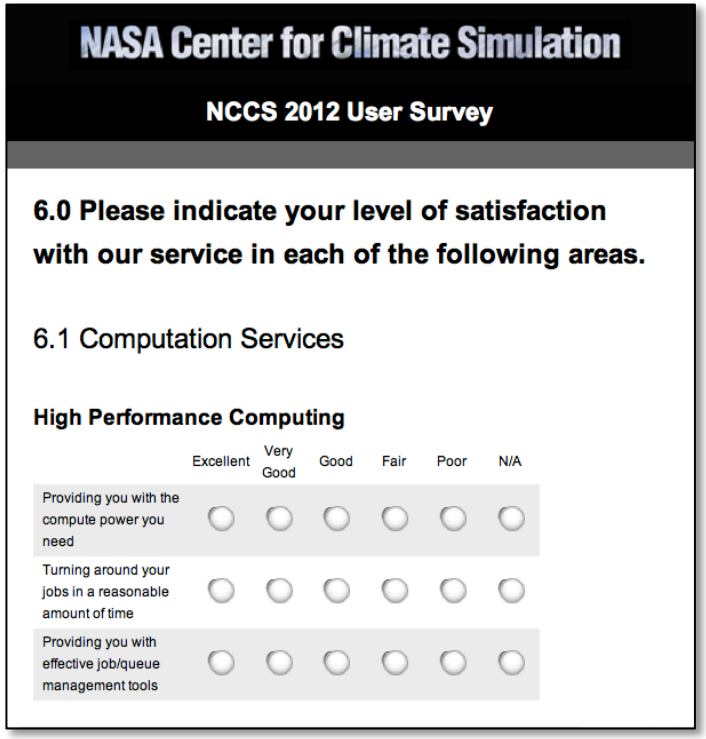
NASA Center for Climate Simulation

Upcoming

- New Resources:
  - Initial "NetApp" Discover nobackup, starting later this week.
  - Discover SCU8 Sandy Bridge: some nodes available to all for "early use" later this week.
    - Use *"sp1"* queue (specify *"ncpus=16"* to get SCU8 nodes).
  - Discover Intel MIC (240 units) installed in October; user training, testing period later fall/winter.
- Planned Outages (to date):
  - Goddard's multi-building electrical outage: NCCS and all of B28, plus Bldgs 5, 18, 19, 20, 28D, 29.
    - NCCS: *~2000 ET Friday, Oct. 12, through Sun., Oct. 14, or Mon., Oct. 15 (TBD—Oct. 28-30?).*
  - Maybe all Discover, definitely SCUs 5, 6, 7, 8: Required InfiniBand OFED (software stack) upgrade
    - *Likely mid-October*, post-HS3 Field Campaign.
  - Half of SCU8 (only) will be taken offline to install MIC units *(mid-October?)*.
  - All of SCU8 (only) downtime to meet benchmark commitments *(mid/late October weekend)*.
- Required Changes:
  - Rolling SLES11 SP1 upgrades for Discover in next few weeks! (mostly transparent)
    - Try now! Use "*sp1*" queue (PBS); *ssh discover-test*, *dali-test* from Discover or Dali (interactive).
  - ***All codes must be recompiled after required InfiniBand OFED (software stack) upgrade, likely mid-October*** (after HS3 Field Campaign finishes).

NCCS User Survey – Coming Soon

- 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.


NASA Center for Climate Simulation

NCCS 2012 User Survey

6.0 Please indicate your level of satisfaction with our service in each of the following areas.

6.1 Computation Services

High Performance Computing

	Excellent	Very Good	Good	Fair	Poor	N/A
Providing you with the compute power you need						
Turning around your jobs in a reasonable amount of time						
Providing you with effective job/queue management tools						

Sample NCCS User Survey screen.

Your frank opinions and suggestions are very much appreciated.

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**

NCCS Brown Bag Seminars

- ~Twice a month 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

- Next talk:

  "Tips for Monitoring Memory in PBS Jobs"

  Tue., 16 October, 12:30-1:30, Building 33, Room E125

- Prioritize topics of interest (and add your suggestions) on today's feedback or signup sheet, or via email to support@nccs.nasa.gov
- ***We will repeat seminars upon request.***

Ongoing Investigations

- Discover: intermittent PBS slowness (qstat, qsub delays).
  - Problem severity escalated with Altair, the PBS vendor.
- Discover GPFS hangs due to jobs exhausting available node memory.
  - Continuing to refine automated monitoring, PBS slowness complicates this.
- Dirac (/archive): occasional hangs in NFS exports to Discover, SGI investigating.
  - Workaround: use scp or sftp to dirac – /archive files are available, just not via NFS.
- Data Portal: resolving problem with one of four disk arrays.
  - Updated disk array microcode applied, file sanity checking is ongoing.
- Discover: investigating hardware options for heavy GPFS metadata workloads (many concurrent small, random I/O actions for directories, filenames, etc.).

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) Database Problem, Aug. 22-30:** Files on the Dirac DMF Archive cluster were unavailable for several days while NCCS staff worked with SGI to recover from a rare-occurrence database corruption, which did not affect the integrity of the content of files stored in the archive. SGI has identified the cause of the corruption and has created a remedy to prevent future occurrences.

- **Long PBS Startup:** The NCCS resolved a longstanding and vexing problem with very long initialization times at startup for Discover's PBS batch system by installing PBS 11 and applying a newly available bugfix patch from PBS vendor Altair (June 2012).

- **Problematic "HDE" Discover nobackup Disks:** Replaced 540+ problematic disks via staff bucket-brigade, and began migrating data back to repaired disk array (July 2012).

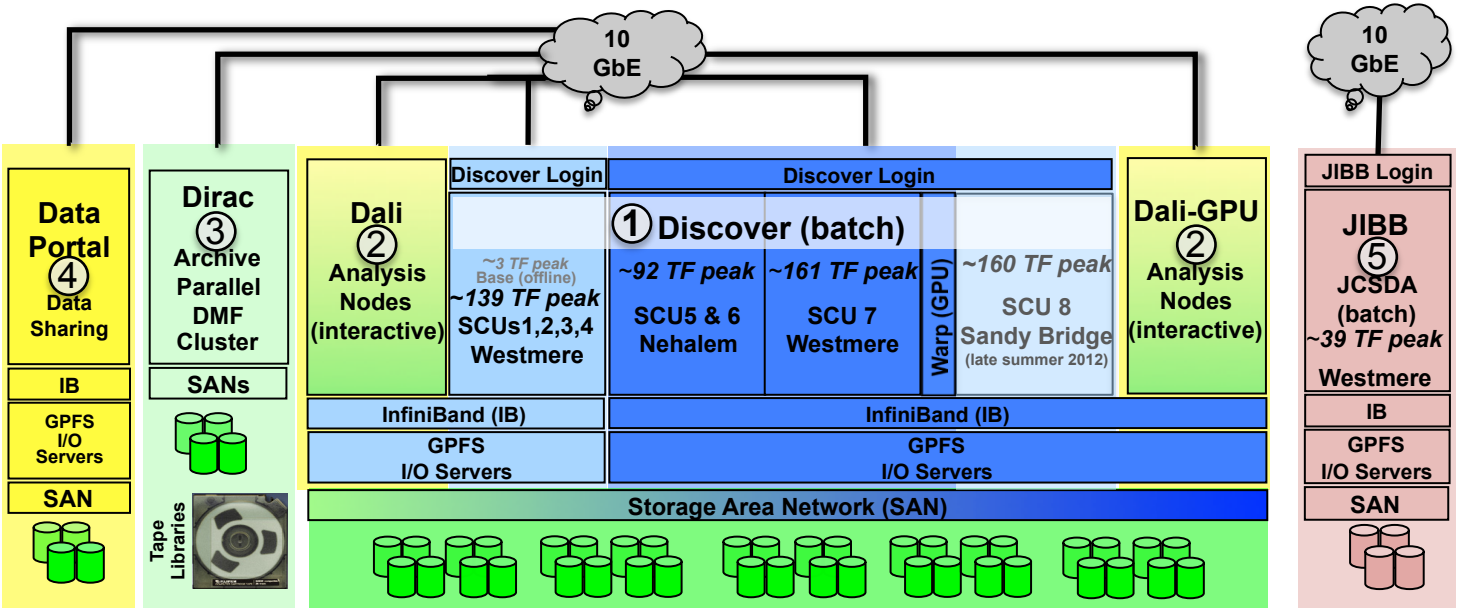
- **Post-Derecho Recovery:** Severe storms the evening of Friday, June 29 led to a Goddard-wide power outage starting Saturday, June 30. The power outage and subsequent restoration led to the failure of a breaker required to power the air conditioning units in one of the two main NCCS computer rooms as well as multiple Discover DDN disk storage hardware errors. The extensive recovery efforts included replacing the failed breaker and two DDN disk controllers, as well as performing multiple disk rebuilds on failed storage tiers. The Discover cluster was restored to service on Thursday, July 5.

NASA Center for Climate Simulation Supercomputing Environment

Supported by HQ's Science Mission Directorate

① ***Discover*** Linux Supercomputer

- Summer 2012:
  - ~3,400 nodes (35,560 cores total)
  - ~400 TFLOPS peak
  - 78 TB memory (2 or 3 GB per core)
  - 3.6 PB disk
- Fall 2012 *additions*:
  - 480 nodes (7,680 cores)
  - ~160 TFLOPS peak
  - 15 TB memory
  - ~3 PB disk



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

Twice-a-Month NCCS Brown Bag Seminars: Delivered & Proposed Topics

- Monitoring PBS Jobs and Memory
- Introduction to Using Matlab with GPUs
- Best Practices for Using Matlab with GPUs
- Using CUDA with NVIDIA GPUs
- Using OpenCL
- Introduction to the NCCS Discover Environment
- Scientific Computing with Python
- Using GNU Octave
- Using Database Filesystems for Many Small Files

- ✓ Intel Many Integrated Core (MIC) Prototype Experiences
- ✓ 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

Slides from June 19, 2012 NCCS User Forum

NASA Center for Climate Simulation

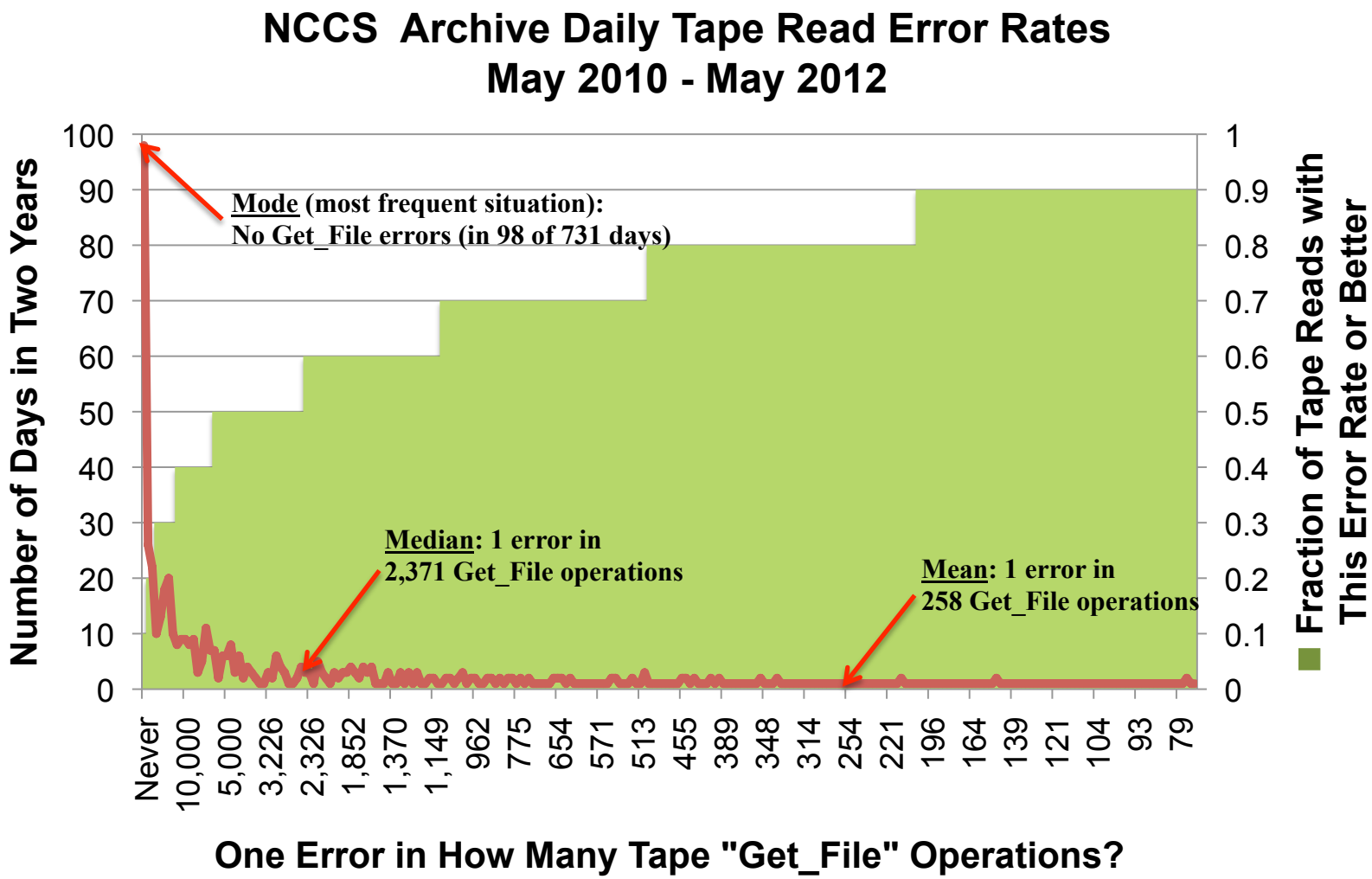
Coming Discover Changes: Linux SLES 11 SP1

- Reason for upgrades:
  - SLES 11 SP1 is required to maintain current Linux security patches.
- SLES 11 SP1:
  - NCCS staff tested & documented the few changes (updated libraries and Linux kernel).
  - Planning for phased, rolling deployments with minimal downtime.

Dirac Archive Single Tape Copy Default Started May 31, 2012

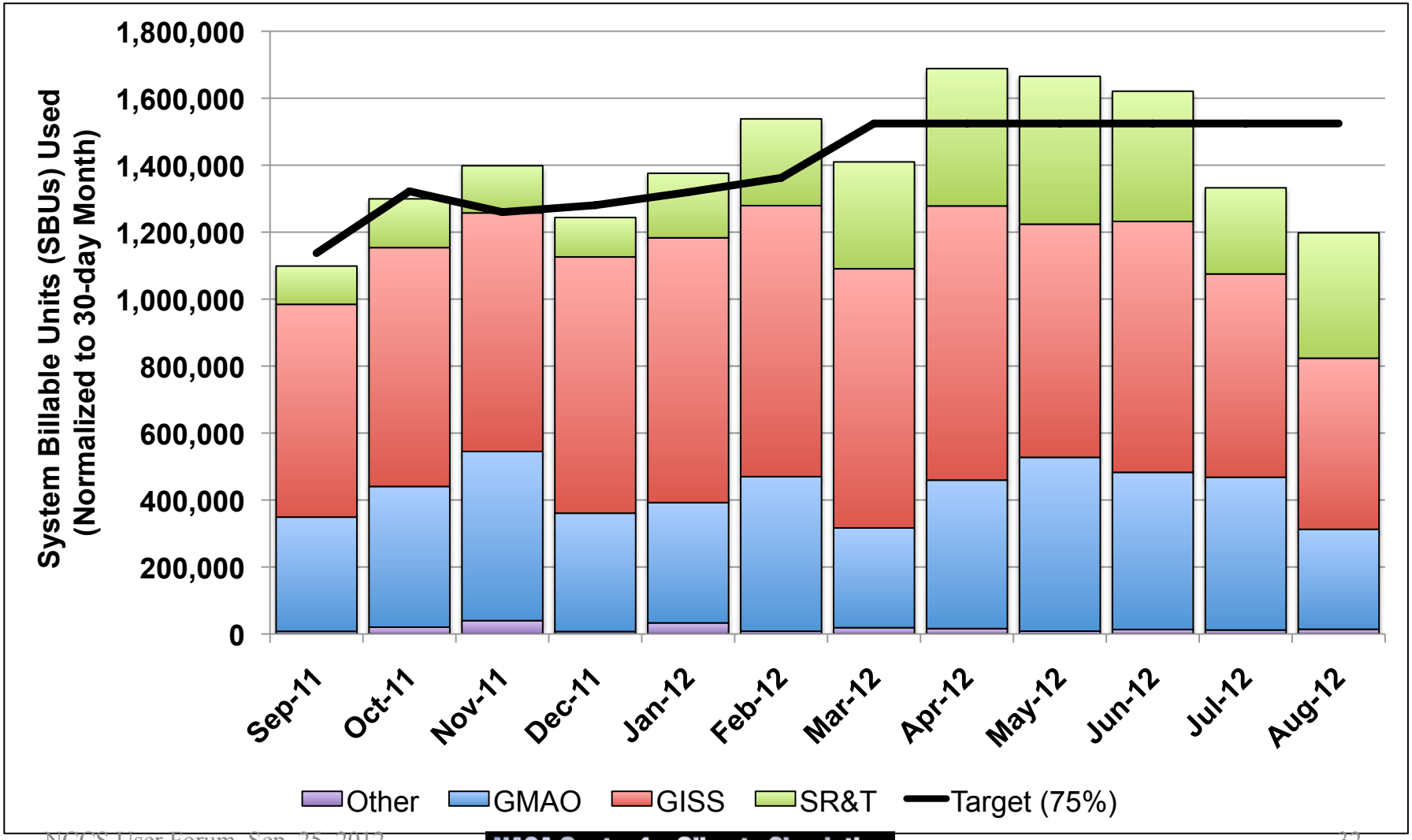
- Default became single tape copy of NCCS archive files:
  - …for all newly created files *and*
  - …all *existing* archive files.
- You can request a second tape copy of your **critical** archive files *at any time**:
  - **dmtag –t 2 <archive_file_name>**
  - *Second tape copy will be made within a few hours to a few days.
- See the following for more details:
  - https://www.nccs.nasa.gov/news.html#dmtag
  - http://www.nccs.nasa.gov/primer/data.html#secondcopy
  - https://www.nccs.nasa.gov/images/DMF-dtag.pdf
- Please contact support@nccs.nasa.gov (301-286-9120) if you have questions or concerns.

NCCS Observed "Get_File" Tape Operation Error Rates, Temporary and Permanent, May 2010 – May 2012



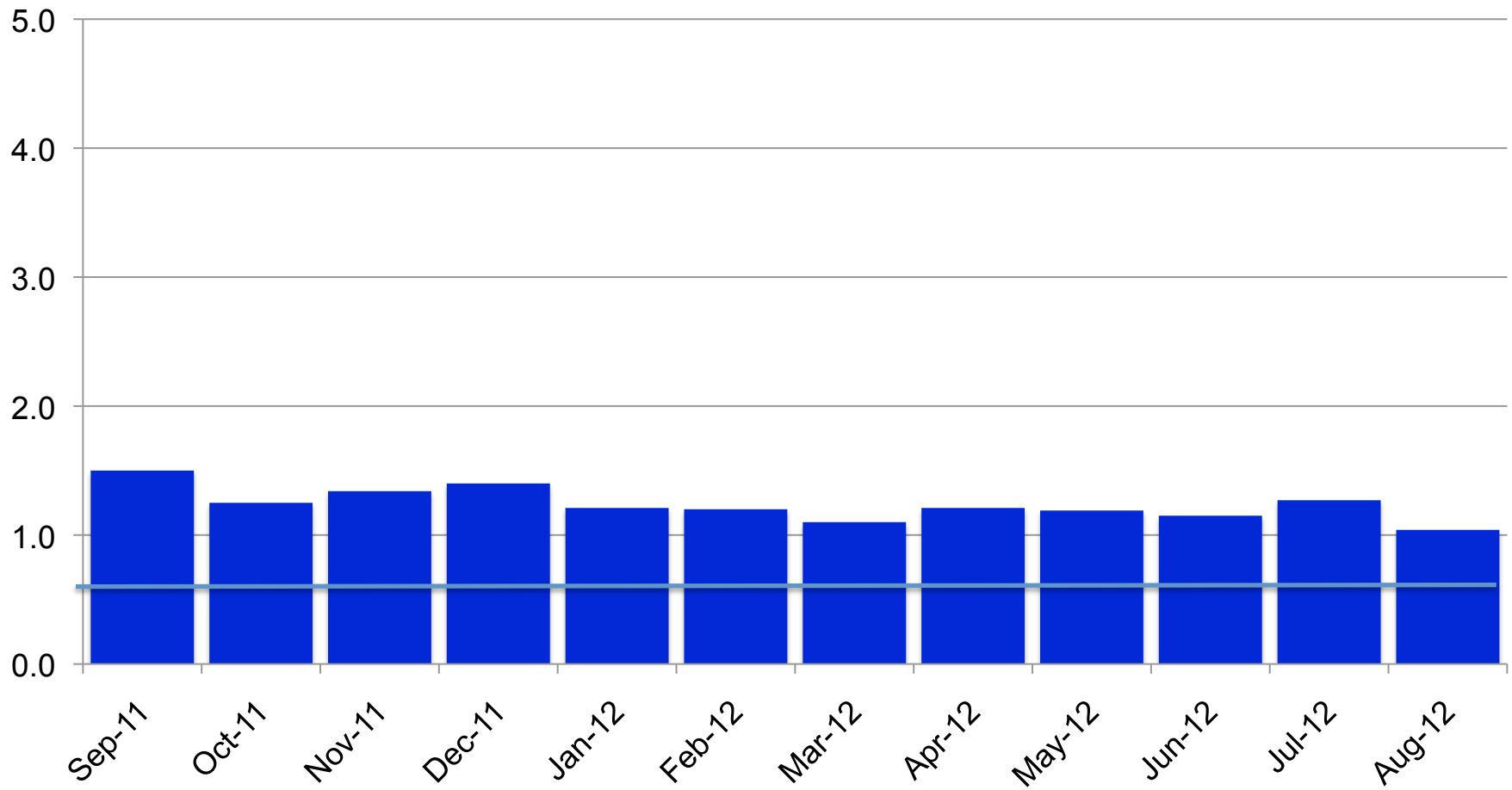
NCCS Metrics Slides
(Through August 31, 2012)

NCCS Discover Linux Cluster Utilization Normalized to 30-Day Month

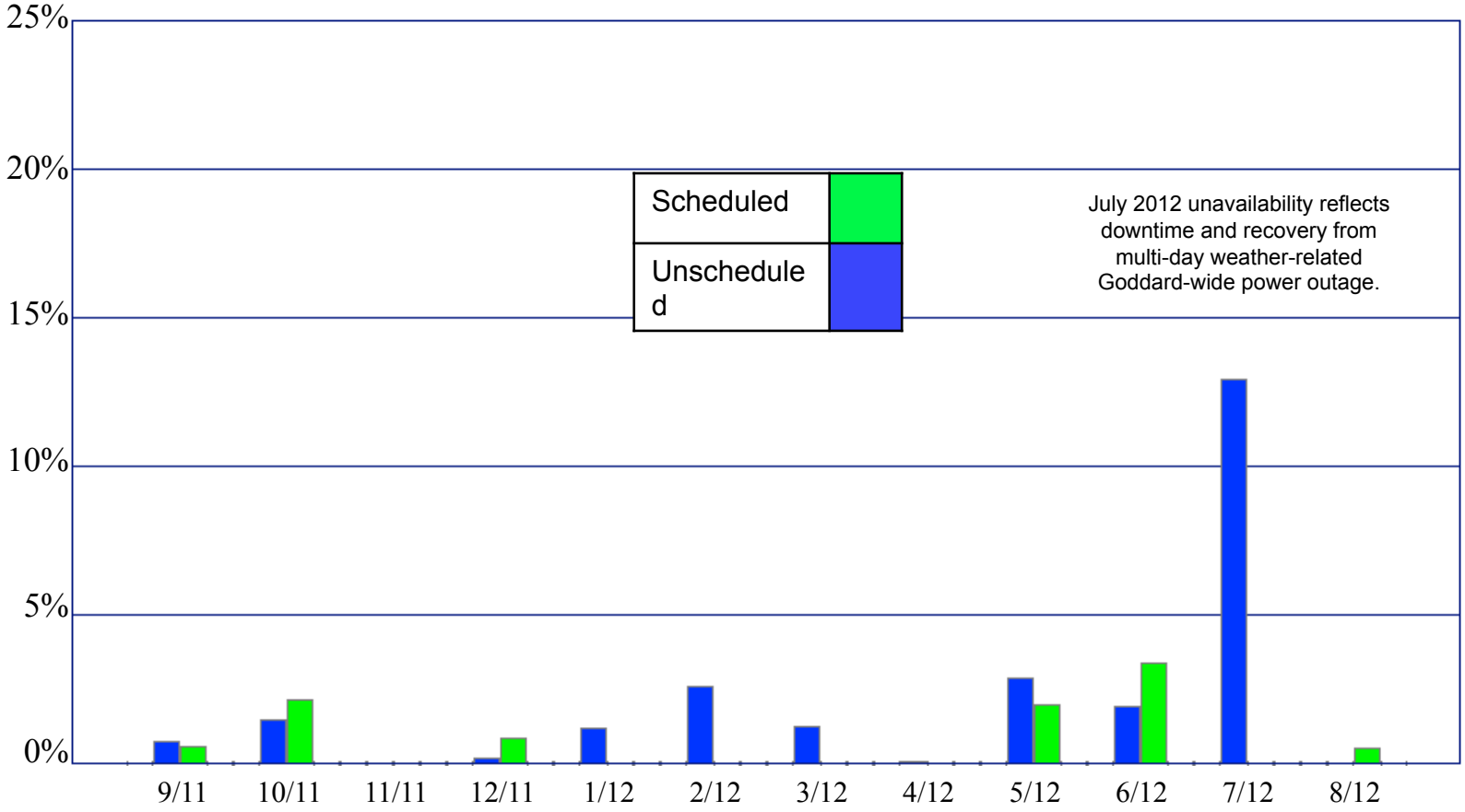


Discover Linux Cluster Expansion Factor

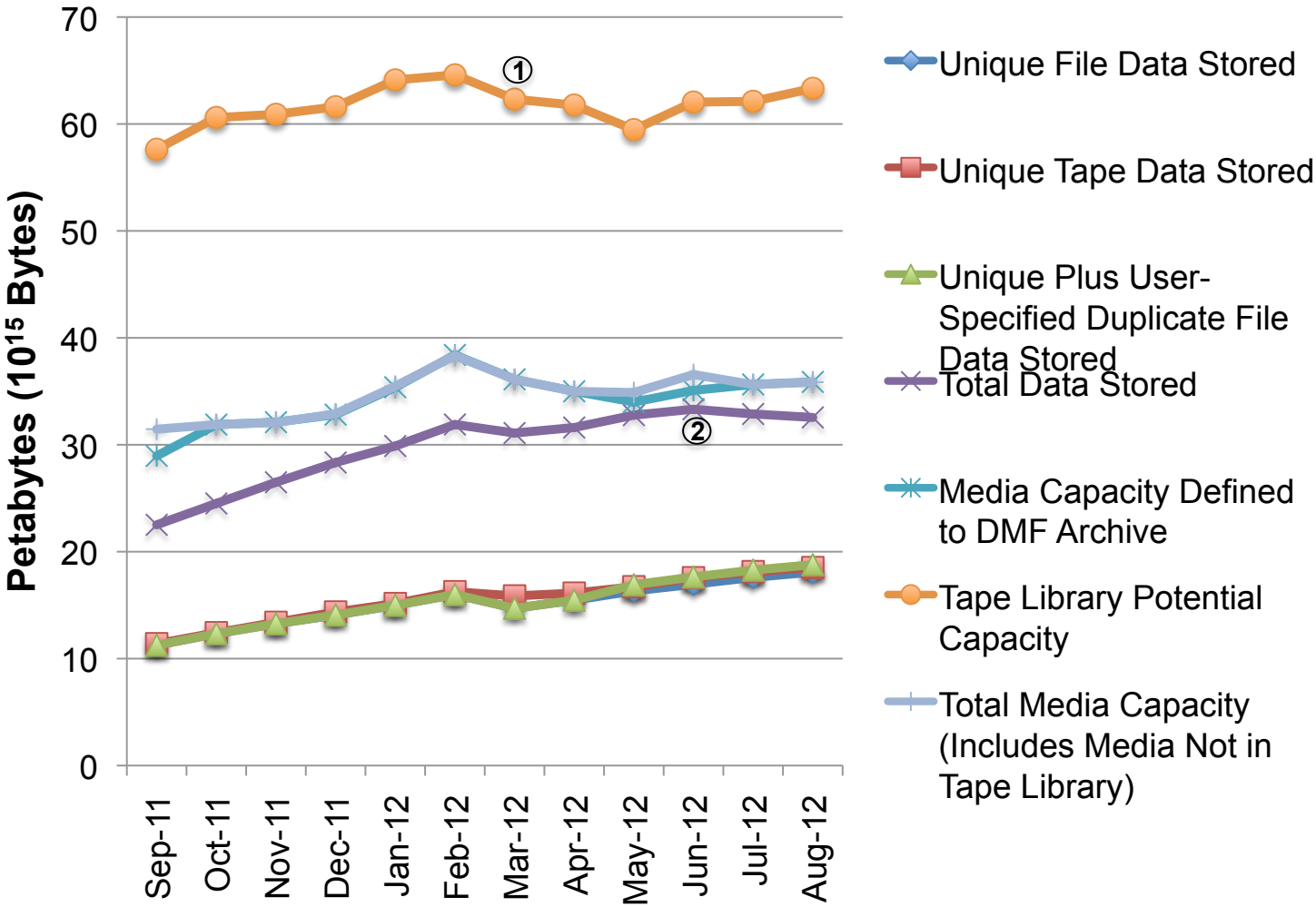
Expansion Factor = (Queue Wait + Runtime) / Runtime



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.