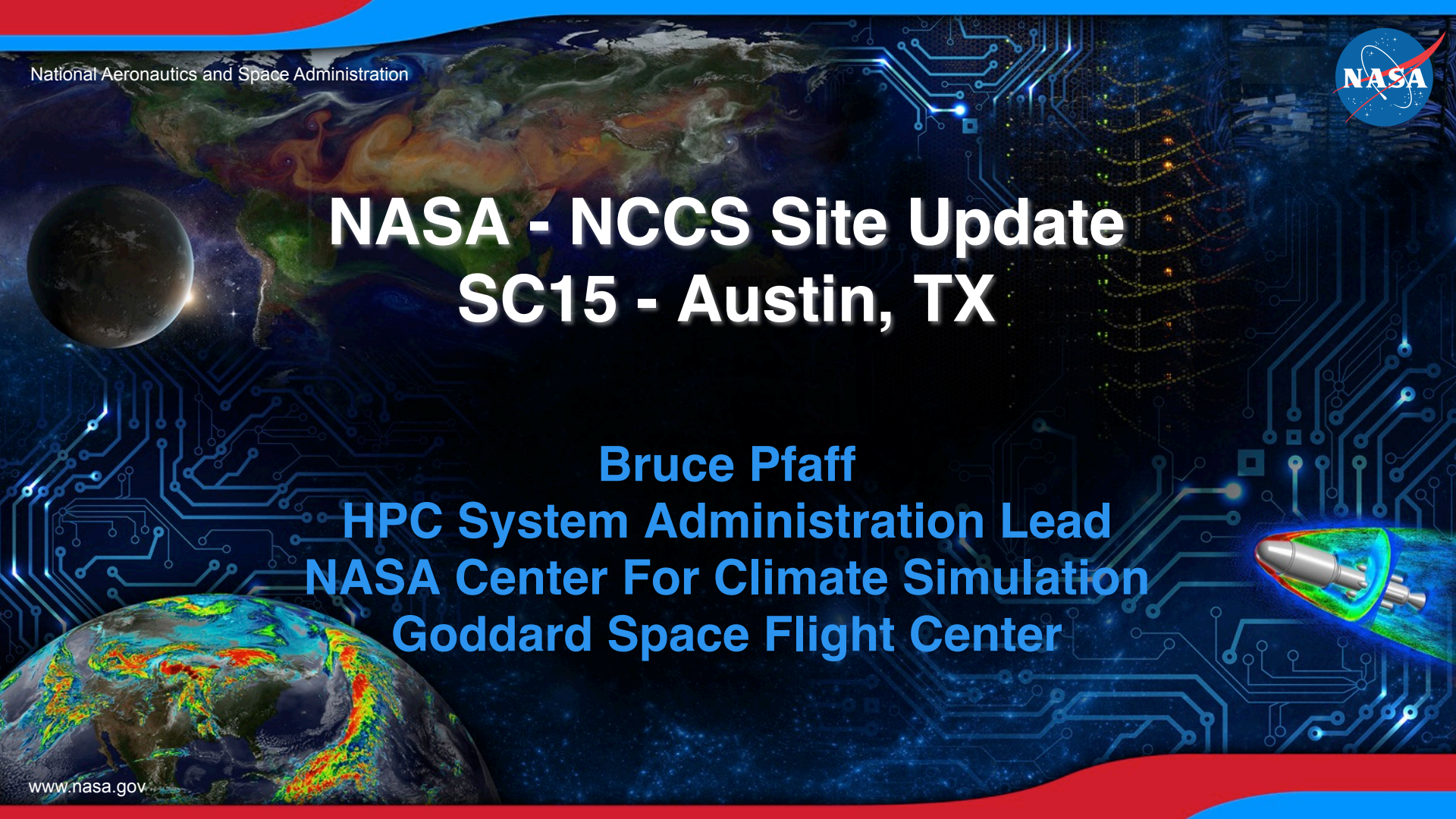
National Aeronautics and Space Administration



NASA - NCCS Site Update
SC15 - Austin, TX

Bruce Pfaff
HPC System Administration Lead
NASA Center For Climate Simulation
Goddard Space Flight Center

National Aeronautics and Space Administration

NCCS Site Update
(Two years later – how's it going?)

Bruce Pfaff
HPC System Administration Lead
NASA Center For Climate Simulation
Goddard Space Flight Center

Agenda

Who we are

(we're not quite as crazy as Ryan… but we're close)

Our system and configuration

Our resource manager transition

Responses to the transition

Configuration comparison

What's next?

NASA Center for Climate Simulation (NCCS)

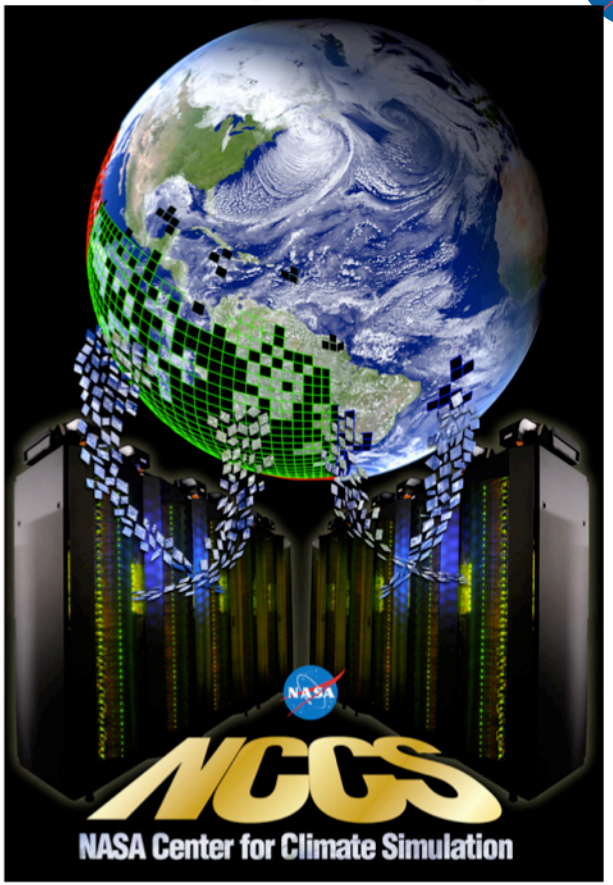
Funded by the Science Mission Directorate

- Located at the Goddard Space Flight Center (GSFC)
- Code 606.2 – Within the CISTO Organization

Provides an integrated high-end computing environment designed to support the specialized requirements of Climate and Weather modeling.

- State-of-the-art high-performance computing, data storage, and networking technologies
- Advanced analysis and visualization environments
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services

http://www.nccs.nasa.gov



Our Configuration

The "Discover" Cluster

- A heterogeneous cluster of x86 Intel based nodes
  - » 7 different generations of Intel processors through the years
  - » Currently 3.5Pf, was ~1.2Pf 6 months ago
- Three separate IB fabrics
- 30Pb+ of direct attached storage (fibre channel SAN)
- Dedicated login, gateway and I/O nodes
- ~3200 compute nodes

Running Slurm v14.03.10 (plus local mods to PBS wrappers)

- Converted from PBS Pro 12 in Oct/Nov 2013
- Plan to move to 15.08 this fall (and skip 14.11)

Recent System Upgrade (SCU10/11/12)

- Decommissioned all Westmere nodes (26000+ cores)
- Added 2200+ Haswell nodes (64,000+ cores)

Recent Upgrades (Nov 2014 - May 2015)

SCU10 – 1080 Haswell nodes, 30240 cores, 1.2Pf peak (11/12/14)

- Retire SCU7 first – 1200 Westmere nodes, 14,400 cores

SCU11 – 612 Haswell nodes, 17136 cores, 713Tf peak (12/15/14)

- Retire SCU3 & SCU 4 first – 516 Westmere nodes, 6,192 cores

SCU12 – 612 Haswell nodes, 17136 cores, 713Tf peak (5/26/15)

- Retire SCU1 & SCU2 first – 516 Westmere nodes, 6,192 cores

Total System is ~80,000 Xeon cores, 3.5Pf peak (SandyBridge & Haswell & Phi's)

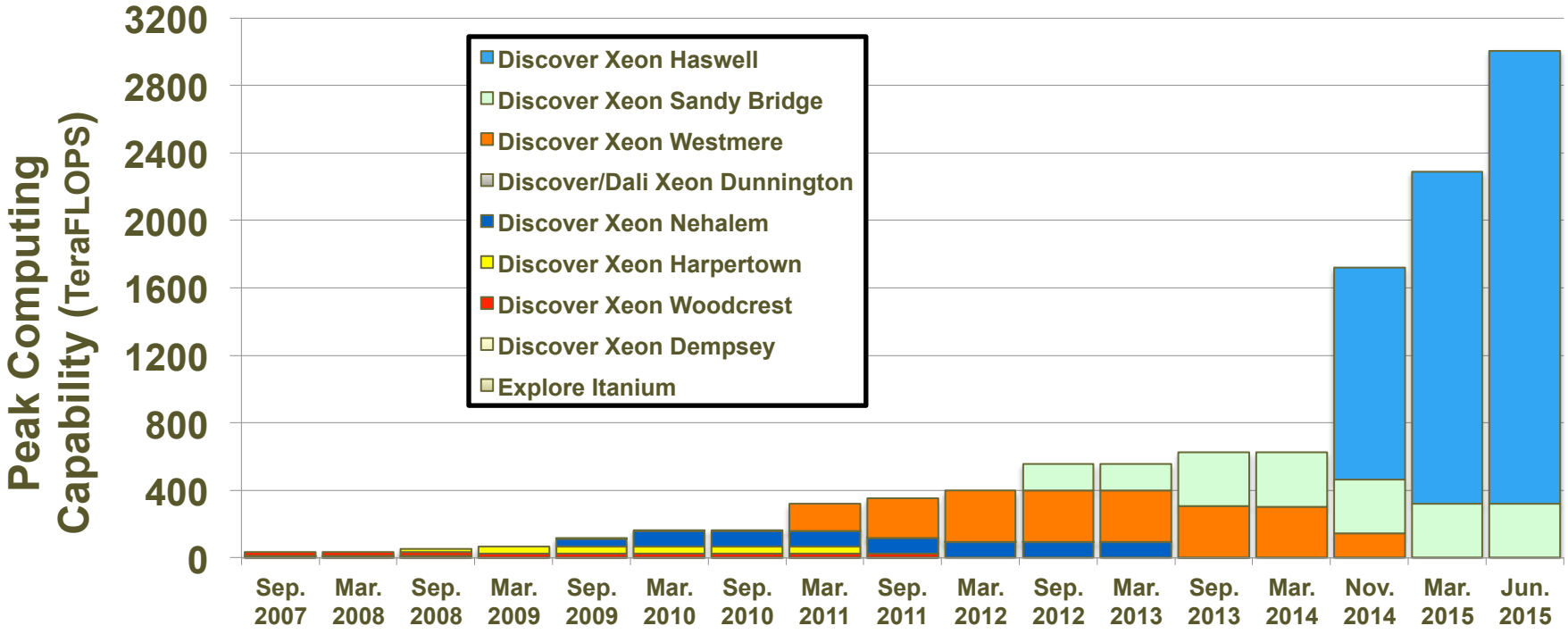
Added 20Pb of raw disk storage

- Retired 2 disk systems to be able to connect the new disks to the SAN
- 4 more older disk systems have been decommissioned

Converted all GPFS Metadata to SSD

- Spent a year testing and evaluating products

Discover Peak Computing Capacity



Discover: A Growing, Changing System

Year	Unit	Vendor	Intel Processor
2006 –	Base Unit	Linux Networx	Dempsey
2007 –	SCU1 & SCU2	Linux Networx	Woodcrest
2008 –	SCU3 & SCU4	IBM	Harpertown
2009 –	SCU5 & SCU6	IBM	Nehalem
2010 –	SCU7	Dell	Westmere
2011 –	SCU1,2,3,4+	IBM	Westmere

- Decommissioned all the Dempsy, Woodcrest and Harpertown systems

2012 – SCU8	IBM	SandyBridge + Intel Phi's
2013 – SCU9	IBM	SandyBridge

- Decommissioned SCU5 & SCU6 (all the Nehalem systems)

2014 – SCU10	SGI	Haswell

- Decommissioned SCU7 (Westmere systems)

2015 – SCU11 & SCU12	SGI	Haswell

- Decommissioned SCU1-4 (last of the Westmere systems)

Currently Active Hardware

Our Resource Managers Through The Years

1980's – Cyber 205

- BatchPro with local mods

1990's – Many Cray Systems (UNICOS)

- NQS/NQE
- Wrote our own scheduler on top of NQS

Early 2000's – HP/Compac DEC Alpha Cluster (Tru64)

- Platform LSF

Early 2000's – Small IBM SP Cluster (AIX)

- Load Leveler

2006-2013 – Discover Cluster (SUSE Linux)

- PBS

2013 – Discover Cluster (SUSE Linux)

- Slurm

Resource Manager Transition PBS -> Slurm

Evaluated multiple resource managers during summer of 2013

- Decision to transition to Slurm was based on technical and cost considerations (best value)

Started working on the transition in August

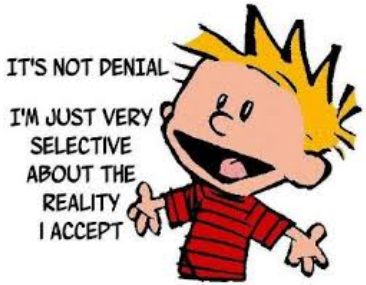
- Limited user testing by mid to late September
- Planned for a month of slow transition during October
- Government shutdown changed our plans

Converted 10 days after shutdown ended

- Mimicked the existing PBS commands and configuration
- Minor issues quickly resolved in the first few weeks
- Some users did not even know we transitioned
- Major customer (GMAO) said the transition went very smooth and had no interruptions in their operational processing

Slide from SC13

The Five Stages of ~~Grief~~ Transition



Denial

- I don't want to change
- I won't change

Anger

- Our previous vendor was angry
  - » Can't really blame them, it's a natural response

Negotiation

- It will look exactly like it currently looks, right?
- I won't have to change anything, right? Please…

Depression

- Why are you making us do this?
- You're taking my dedicated resources away? Why???

Acceptance

- Users: This isn't so bad, some of it is actually pretty cool.
- Admins: This is flexible and feels natural, we love this!

How Our Users Responded

Four major responses

I'm going to continue to use PBS syntax, I don't want to change anything

- Many users have a mental block and don't want to consider changing things

I'll only change what I have to in order to make things work

- When a problem arises, and we tell them the Slurm solution, they'll change only what they need to change to solve the problem

I'm converting everything to native Slurm

- Many of our power users have jumped right in, changing everything
- Some are asking, why did you pick this configuration setting?

We changed?? When??

- Some users still don't know we've changed

How Our Admins Responded

Cautiously Optimistic – Admins generally don't like change

- We've changed resource managers before, we can do it again

Concerned about speed of the transition

- We didn't have much time for the transition
- A government shutdown didn't help
- Ultimately converted in 10 days (kind of)

Extremely happy with Slurm

- Intuitive commands and interfaces
- Customizable output formats
- Understands RegEx's
- Flexible, rapid, configuration changes

Eeyore became Tigger

Previous Config: ~20 Queues

We mimicked ALL these queues and their limits and turned them into partitions during our transition

6 – general purpose

4 – supporting operational workloads

5 – specific user groups, some dedicated hardware

3 – tied to specific hardware (GPUs, Phi's, etc)

2 or 3 – admin use for testing and/or maint activities

WARNING: DON'T try this at home (v2.6 issues – it's fine now)

- sinfo won't like you (depending on your node names)
- 1+ minute response time with this many partitions vs. 5 seconds now
- sinfo performance improvements have been added to newer releases

Current Config: ~~8~~ 6 Partitions (and shrinking)

2 - general purpose

- Compute, datamove

2 - supporting operational workloads

- 1 still to be retired

1 - specific user group, dedicated hardware

- GMAO high-resolution Nature Runs

1 - tied to specific hardware (Native mode Phi's)

0 - admin use for testing and/or maint activities

- These tasks are generally accomplished via reservations now

HOWEVER…

Current Config: ~~26~~ 33 QoSs (and growing)

- ~~9~~ 11 – general purpose, different limits and pre-emption options
- ~~4~~ 4 – supporting operational workloads
- ~~9~~ 14 – specific user groups, priority and limit
- ~~2~~ 3 – used to facilitate transition
- ~~2~~ 1 – temporary (short-term requirements – conference papers, etc)

Of those 32 QoSs

- ~~9~~ 6 – Priority-only
- ~~5~~ 9 – Priority with constraints
- ~~3~~ 2 – Priority with exceptions
- ~~9~~ 16 – Priority with exceptions & constraints

9 are basically inactive (past or future use)

Why are QoSs better than Partitions?

We had over 700 nodes in dedicated partitions

- 9000 cores out of 40,000 – 22.5% (12-core Westmere nodes)

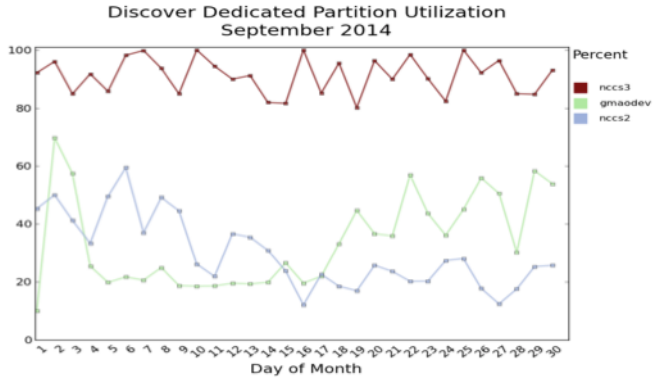
Only 140 nodes remain dedicated, supporting operational workload

- 3920 cores out of 64,512 = 6% (28-core Haswell nodes)
- 0 cores out of 15,360= 0% (16-core Sandybridge nodes)
- Plan to convert this work to reservations and/or QoS
  - » Operational challenges, so it's taking longer

Our dedicated resources were poorly utilized

- One partition regularly 80-90% utilized (3000 cores)
- Other 2 partitions regularly <30% utilized (6000 cores)

Rapid response to changing requirements



What's Next?

Upgrade from 14.03.10 to 15.08.x

Retire remaining partitions from the old regime – only 1 left

- Goal: Only 2 or 3 partitions total (OK, maybe 4 or 5)

Convert remaining operational workload to QoS and reservations

- Need to work closely with the users to manage this
- Some of this has already hapened

Provide preemption capabilities to general user community

- Did some prelim work on this with targets group of users
- Got distracted by large hardware and OS upgrade

Integrate a job submit plugin to allow some additional functions

- Enforce limits
- Reject invalid requests

Enhance epilog to provide additional reporting and job analysis

What's the Take-away From Our Transition?

The transition went well, a few challenges, but overall success

Superior performance characteristics of the RM

System startup takes seconds, not minutes or hours

Fewer dedicated resources -> greater system throughput

Higher utilization from defragmentation of resources

Rapid response to changing requirements and short term needs

Flexible policy engine, easy to customize and adjust

High Availability allows changes w/o scheduling downtimes

Plugin architecture allows for unlimited local enhancements

Great support from SchedMD (and from the Slurm community)

National Aeronautics and Space Administration

NCCS Site Update
(Two years later – how's it going?)

Bruce Pfaff
HPC System Administration Lead
NASA Center For Climate Simulation
Goddard Space Flight Center