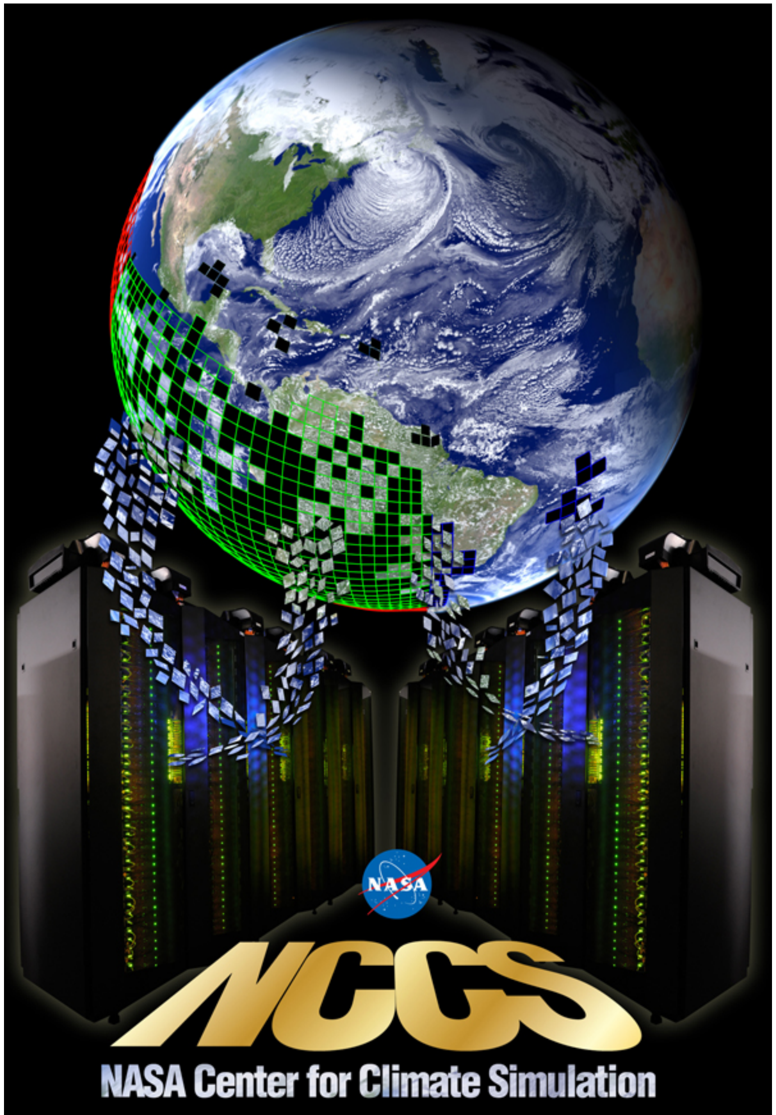


User Forum

NASA Center for Climate Simulation

High Performance Science

June 26, 2018

Agenda

- CISTO and NCCS Changes
- Discover Linux Cluster
- ADAPT Virtualization Environment
- Storage/Analytics Architecture and Data Management Plans

CISTO and NCCS Changes

Dan Duffy,
Chief, Computational and Information Sciences and Technology Office (CISTO)
HPC Lead and NCCS Lead Architect

June 2018: Staff Additions

Welcome to New Members of the NCCS and CISTO Team:

Kerman Bime/IT Coalition, System Administration

Jim Carlisi/GDIT, System Administration

Luli Laulu/Inuteq, Office Administrative Assistant

Elizabeth Nerdig/GDIT, Facilities Engineer

Jason Robbins/ GDIT, System Administration

George Roros, Kiosk Technology Consultant

Darryl Smallwood/ IT Coalition, Data Services

Colton Weinman, Graphics Design Consultant

Gabe Borroni, Disasters GIS Analyst

Garrett Layne, Disasters GIS Analyst

Buchi Oraegbu, AIST Managed Cloud Environment

Dan Sherman, AIST Managed Cloud Environment

Summer 2018: Interns

Welcome to New Members of the NCCS and CISTO Team:

Jordan Caraballo-Vega

Thomas Favata

August Morin

Paulo Paz

Carly Robbins

Matt Stroud

Donovan Murphy (working with Craig)

Chris Culver (working with Craig)

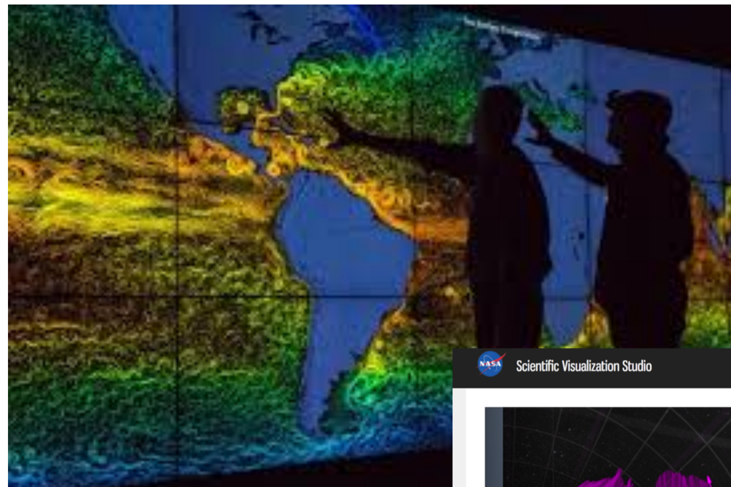
NASA Science Mission Directorate High End Computing (HEC) SMD19 Allocation Requests

- SMD19 period: October 1, 2018 - September 30, 2019.
- Deadline for Oct. 1 allocation requests is August 1, 2018.
- Principal Investigators can submit eBooks requests now!
- Webinar on the revised SMD HEC allocation process: 4 pm ET today!
  - Slides and webinar recording will be available later
- If you have questions about the new eBooks process, or an existing allocation or request, email support@nccs.nasa.gov to talk to Nancy Carney.

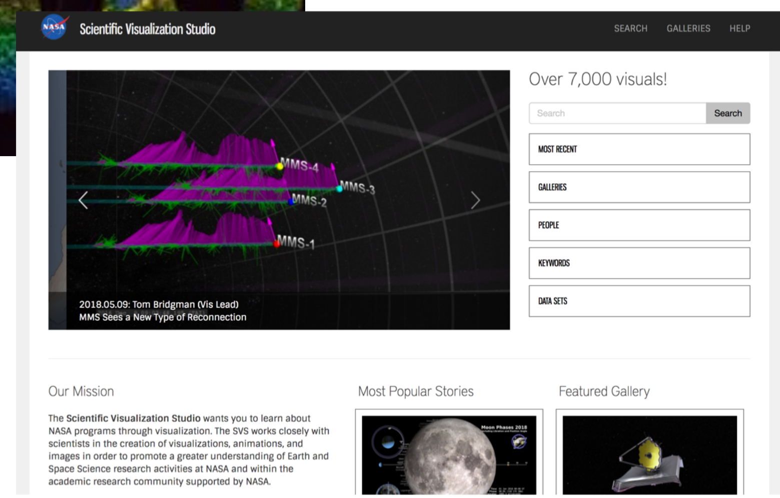
Scientific Visualization Studio

Scientific Visualization Studio

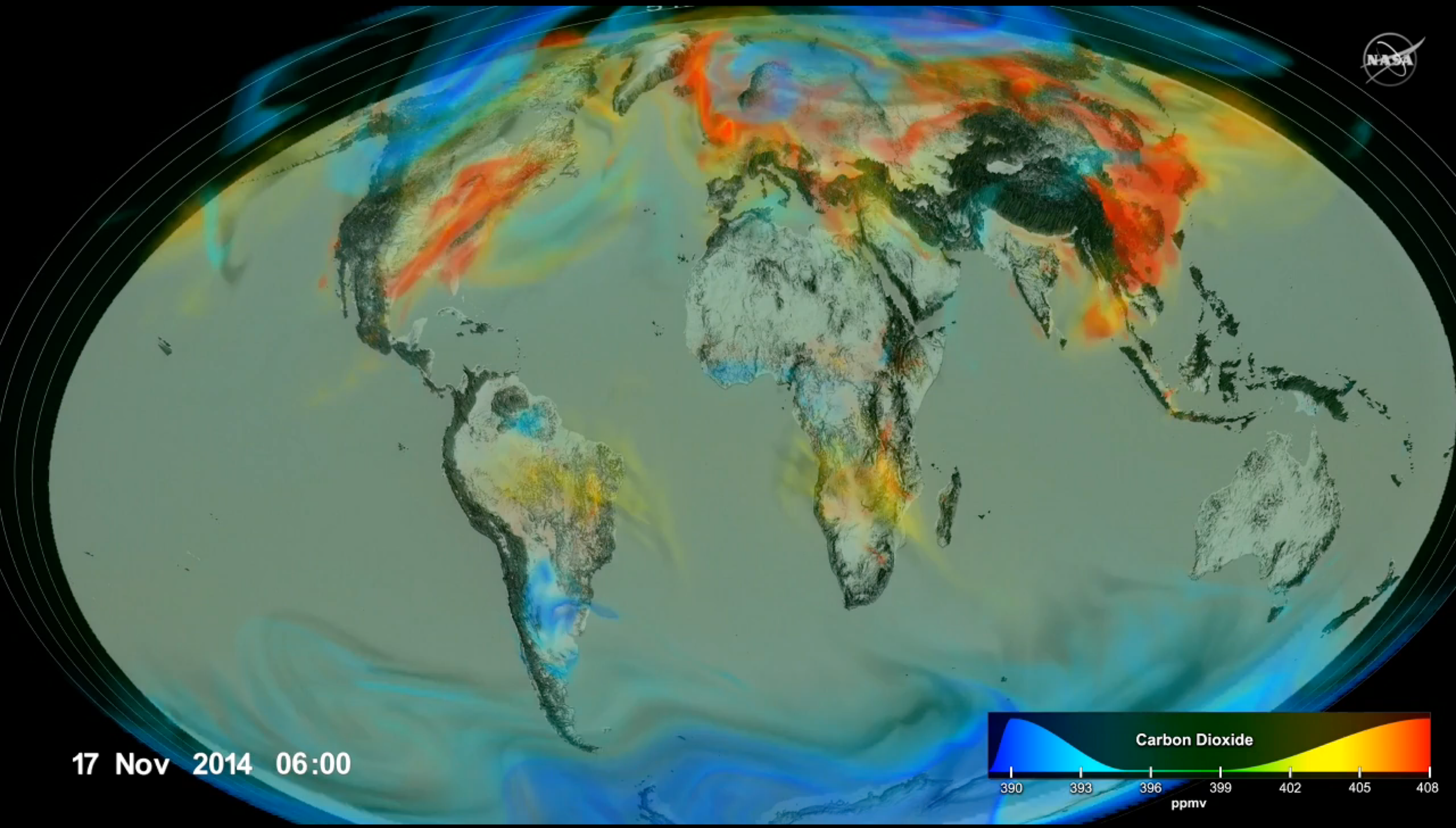
Visualizers working closely with scientists to create new ways of viewing massive amounts of data.



All content available at https://svs.gsfc.nasa.gov



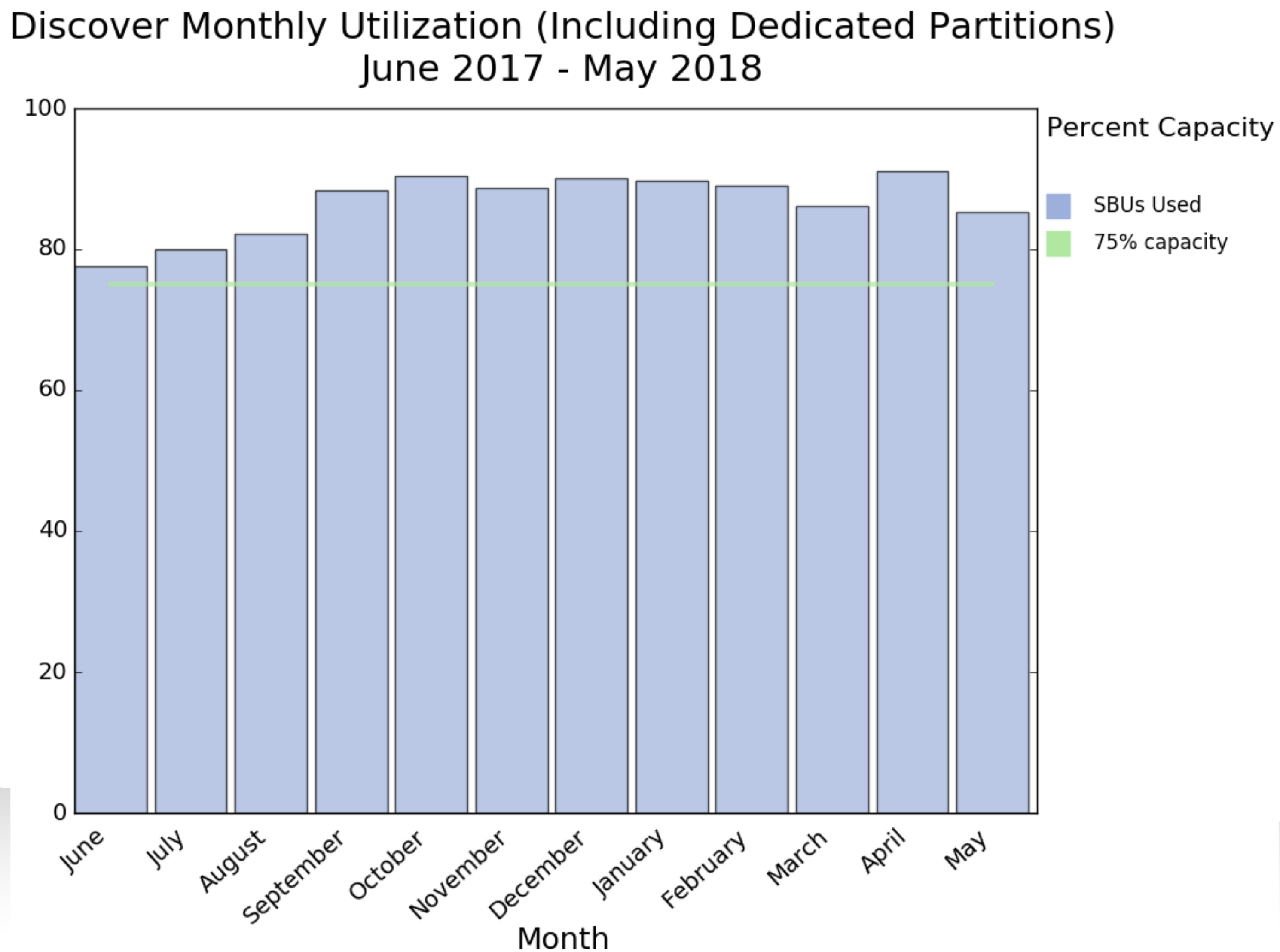
Contact Horace Mitchell for more information:
horace.g.mitchell@nasa.gov


NASA
17 Nov 2014 06:00
Carbon Dioxide
390
393
396
399
402
405
408
ppmv

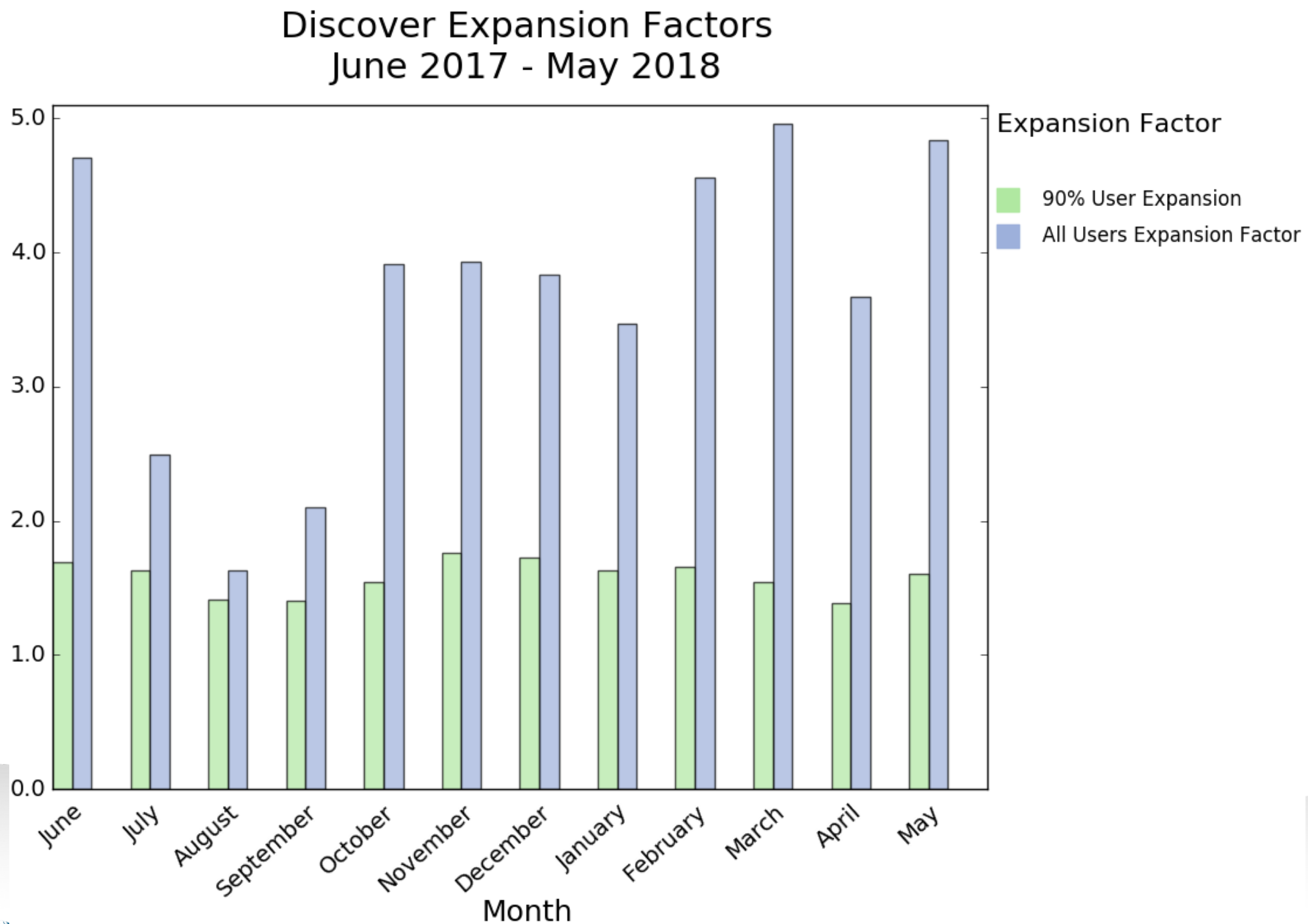
Discover Linux Cluster

Dan Duffy

Discover 12-Month Utilization Percentage Trend



Discover Expansion Factors – 12-Month Trend



SCU14

Discover FY17 Compute Upgrade Summary

- Scalable Unit 14 (SCU14)
  - First single NCCS system with >1.5 PF
  - First OPA system in NASA
- Edge Solutions and SuperMicro
- 520 Compute Nodes
  - Dual-socket with 20-core Intel Skylake 2.4 GHz processors
  - 20,800 total cores
  - 1,560 TF peak computing
  - 192 GB of RAM per node
- 24 Service Nodes
- 20 I/O Nodes
- Intel OmniPath (OPA) Interconnect
  - 100 Gbps
  - 2-to-1 blocking
  - Interconnect designed to easily scale to 2x the number of compute nodes

Come visit during the Science Jamboree: July 25th

Discover Scalable Unit Evolution

2014	SCU1 258 Nodes 3,096 Cores 2.8 GHz, 24 GB Westmere DDR IB 34.7 TF	SCU2 258 Nodes 3,096 Cores 2.8 GHz. 24 GB Westmere DDR IB 34.7 TF	SCU3 258 Nodes 3,096 Cores 2.8 GHz, 24 GB Westmere DDR IB 34.7 TF	SCU4 258 Nodes 3,096 Cores 2.8 GHz, 24 GB Westmere DDR IB 34.7 TF	SCU7 1,200 Nodes 14,400 Cores 2.8 GHz, 24 GB Westmere QDR IB 161.3 TF	SCU8 480 Nodes 7,680 Cores 2.6 GHz, 32 GB SandyBridge QDR IB Xeon Phi 606 TF	SCU9 480 Nodes 7,680 Cores 2.6 GHz, 64 GB SandyBridge FDR IB 160 TF
2015	SCU11 612 Nodes 17,136 Cores 2.6 GHz, 128 GB Haswell FDR IB 683 TF	SCU12 612 Nodes 17,136 Cores 2.6 GHz, 128 GB Haswell FDR IB 683 TF			SCU10 1,080 Nodes 30,240 Cores 2.6 GHz, 128 GB Haswell FDR IB 1,229 TF	SCU8 480 Nodes 7,680 Cores 2.6 GHz, 32 GB SandyBridge QDR IB Xeon Phi 606 TF	SCU9 480 Nodes 7,680 Cores 2.6 GHz, 64 GB SandyBridge FDR IB 160 TF
2016	SCU11 612 Nodes 17,136 Cores 2.6 GHz, 128 GB Haswell FDR IB 683 TF	SCU12 612 Nodes 17,136 Cores 2.6 GHz, 128 GB Haswell FDR IB 683 TF			SCU10 1,080 Nodes 30,240 Cores 2.6 GHz, 128 GB Haswell FDR IB 1,229 TF	SCU13 648 Nodes 18,144 Cores 2.6 GHz, 128 GB Haswell FDR IB 723 TF	SCU9 480 Nodes 7,680 Cores 2.6 GHz, 64 GB SandyBridge FDR IB 160 TF
2017	SCU11 612 Nodes 17,136 Cores 2.6 GHz, 128 GB Haswell FDR IB 683 TF	SCU12 612 Nodes 17,136 Cores 2.6 GHz, 128 GB Haswell FDR IB 683 TF		SCU10 1,080 Nodes 30,240 Cores 2.6 GHz, 128 GB Haswell FDR IB 1,229 TF	SCU13 648 Nodes 18,144 Cores 2.6 GHz, 128 GB Haswell FDR IB 723 TF	SCU14 520 Nodes 20,800 Cores 2.4 GHz, 192 GB Skylake OPA 1,560 TF	SCU9 240 Nodes 3,840 Cores 2.6 GHz, 64 GB SandyBridge FDR IB 80 TF

Current NCCS HPC Platform Summary

SCU9
280 Nodes
4,480 Cores
2.6 GHz
64 GB
Sandy Bridge
FDR IB
93.2 TF

SCU10
1,080 Nodes
30,240 Cores
2.6 GHz
128 GB
Haswell
FDR IB
1,258.0 TF

SCU11
612 Nodes
17,136 Cores
2.6 GHz
128 GB
Haswell
FDR IB
712.9 TF

Discover

- Intel Xeon nodes
  - 3,752 nodes
  - 107,936 cores
  - Peak 5,129 TFLOPS general purpose
- 42 PB disk

SCU12
612 Nodes
17,136 Cores
2.6 GHz
128 GB
Haswell
FDR IB
712.9 TF

SCU13
648 Nodes
18,144 Cores
2.6 GHz
128 GB
Haswell
FDR IB
754.8 TF

SCU14
520 Nodes
20,800 Cores
2.4 GHz
192 GB
Skylake
OPA 100
1,597.4 TF

Centralized Storage (evolved from DASS)

- 11 HP Apollo Intel Xeon nodes
- 15 PB disk

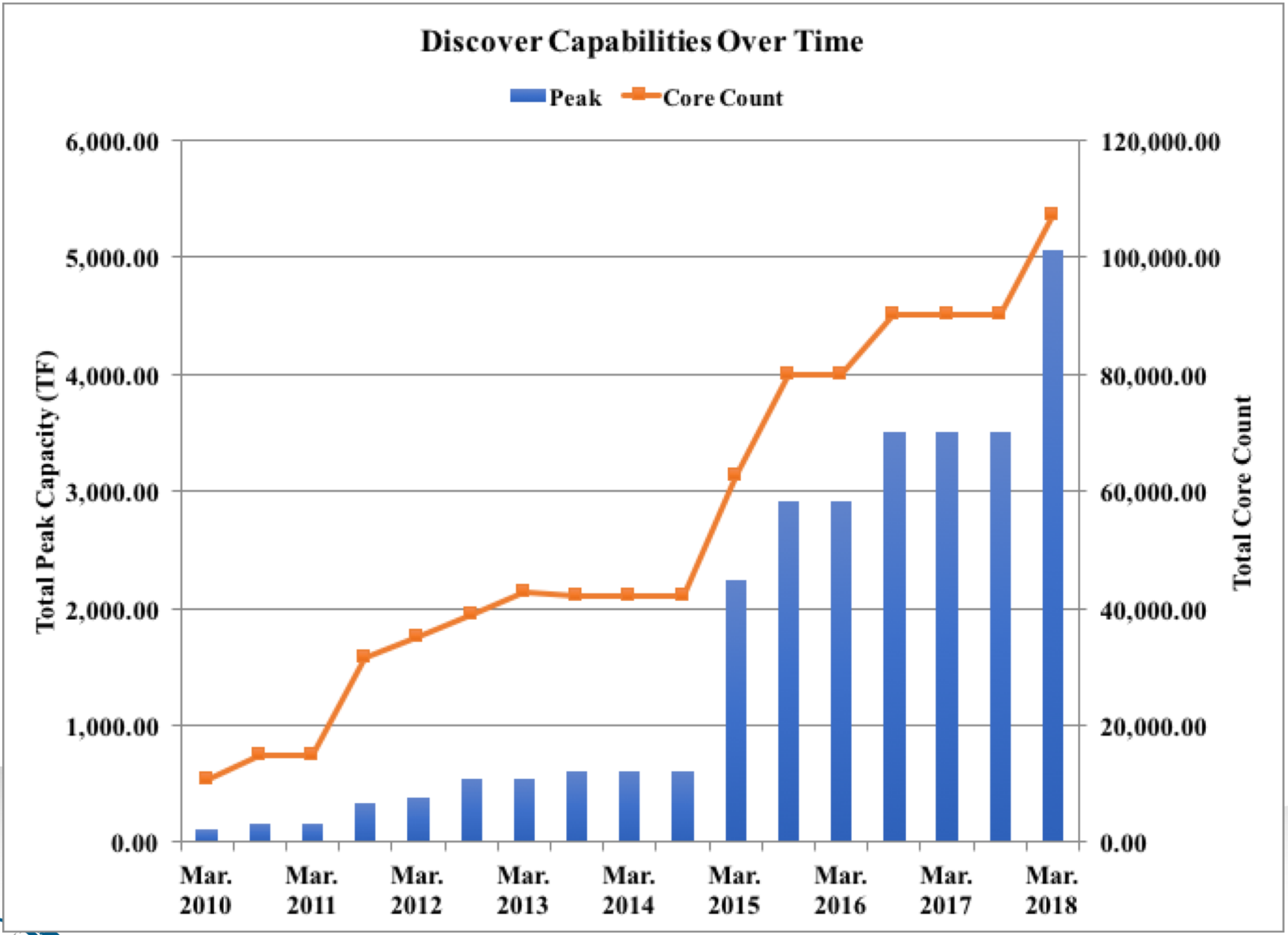
ADAPT

- Intel Xeon nodes
  - 554 nodes
  - 12,148 cores
  - Peak 232.2 TF
- 8 PB disk

Mass Storage System

- ~90 PB robotic tape library
- Data Management Facility (DMF) space management
- 4.4 PB disk

Discover Total Capacity Over Time



Discover Upgrades: Omnipath/OPA, SLES 12 and Slurm17

- All three are being refined and tested in the more isolated SCU14 environment first, with measured deployment to the rest of Discover later.
- A number of changes, especially for Omnipath/OPA and SLES 12.
- Look for upcoming Brown Bags and Documentation!

Discover Upgrades for FY19 – Plans

- Compute
  - SCU15 (specific details TBD)
  - Compute nodes only, to be integrated with SCU14 to approximately double the total number of cores and peak computing capability
- Storage
  - High speed disk
  - Minimum of 15 PBs RAW (prior to RAID)

Advanced Data Analytics Platform (ADAPT)

Dan Duffy

Advanced Data Analytics Platform (ADAPT)
High-Performance Science Cloud

Capability and Description	Configuration
 **Persistent Data Services** Virtual machines or containers deployed for web services, examples include ESGF, GDS, THREDDS, FTP, etc.	128 GB of RAM, 10 GbE, and FDR IB
 **DataBase** High available database nodes with solid state disk.	128 GB of RAM, 3.2 TB of SSD, 10 GbE, and FDR IB
 **Remote Visualization** Enable server side graphical processing and rendering of data.	128 GB of RAM, 10 GbE, FDR IB, and GPUs
 **High Performance Compute and Machine Learning** More than ~10,000 cores coupled via high speed networks for elastic or itinerant computing requirements.	~300 nodes with between 24 and 256 GB of RAM; Small set of nodes with 6 TB of SSD; 16 Nvidia Tesla V100s
 **High-Speed/High-Capacity Storage** Petabytes of storage accessible to all the above capabilities over the high speed Infiniband network.	Storage nodes configured with ~10 PB's of usable capacity
 **High Performance Networks** Internal networks enable high speed access to storage, while external networks provide high performance data movement.	External: 10 and 40 GbE Internal: 10 GbE and Infiniband

ADAPT, Current and Coming Soon

- Access available now for NCCS users, email support@nccs.nasa.gov
- Coming Soon:
  - Additional compute nodes (modular container)
  - New NVIDIA V100 GPU systems (more detail on next slides), K40 GPU nodes available now
  - Convert InfiniBand network to Ethernet
    - Better utilization of container-based hypervisors
  - Fold ADAPT 1.0, where feasible, into OpenStack (ADAPT 2.0) control
    - Future user portal for self-provisioning
  - Introduce Cloud Bursting
    - Leverage commercial clouds to augment processing

Three New Systems in ADAPT for Machine Learning/Deep Learning (ML/DL)

	2 @ 4 GPU Systems	8 GPU System
Total Cores	24	24
Speed of Core	2.3 GHz	2.3 GHz
RAM	512 GB	512 GB
GPUs	4 by V100	8 by V100
Network Interface	2 by 50 Gbps	2 by 50 Gpbs
Local Storage	2 by 800 GB SSD 2 by 3.2 TB NVMe	2 by 800 GB SSD 2 by 3.2 TB NVMe
Operating System	CentOS	CentOS
Software	Python, Caffe, TensorFlow, Custom	Python, Caffe, TensorFlow, Custom

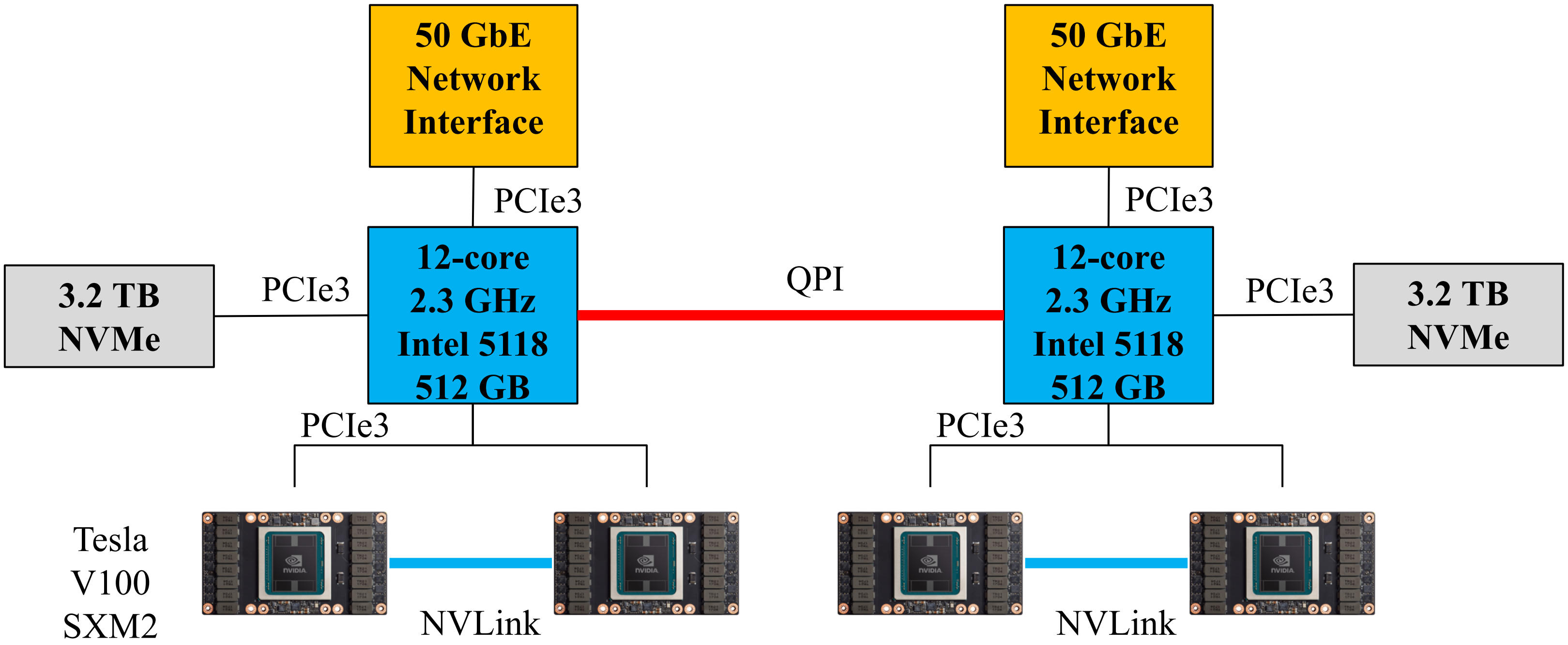
Accessible by NASA credentialed users; must have an NCCS account. Email support@nccs.nasa.gov to inquire about getting access.

NVidia Tesla V100 Specifications

Specifications	Performance
Double Precision Performance (64-bit)	7.8 teraFLOPS
Single Precision Performance (32-bit)	15.7 teraFLOPS
Tensor Performance (16-bit)	125 teraFLOPS
High Bandwidth Memory (HBM)	16 GB
HBM Throughput	900 GB/sec
Interconnect Bandwidth (NVLink)	300 GB/sec

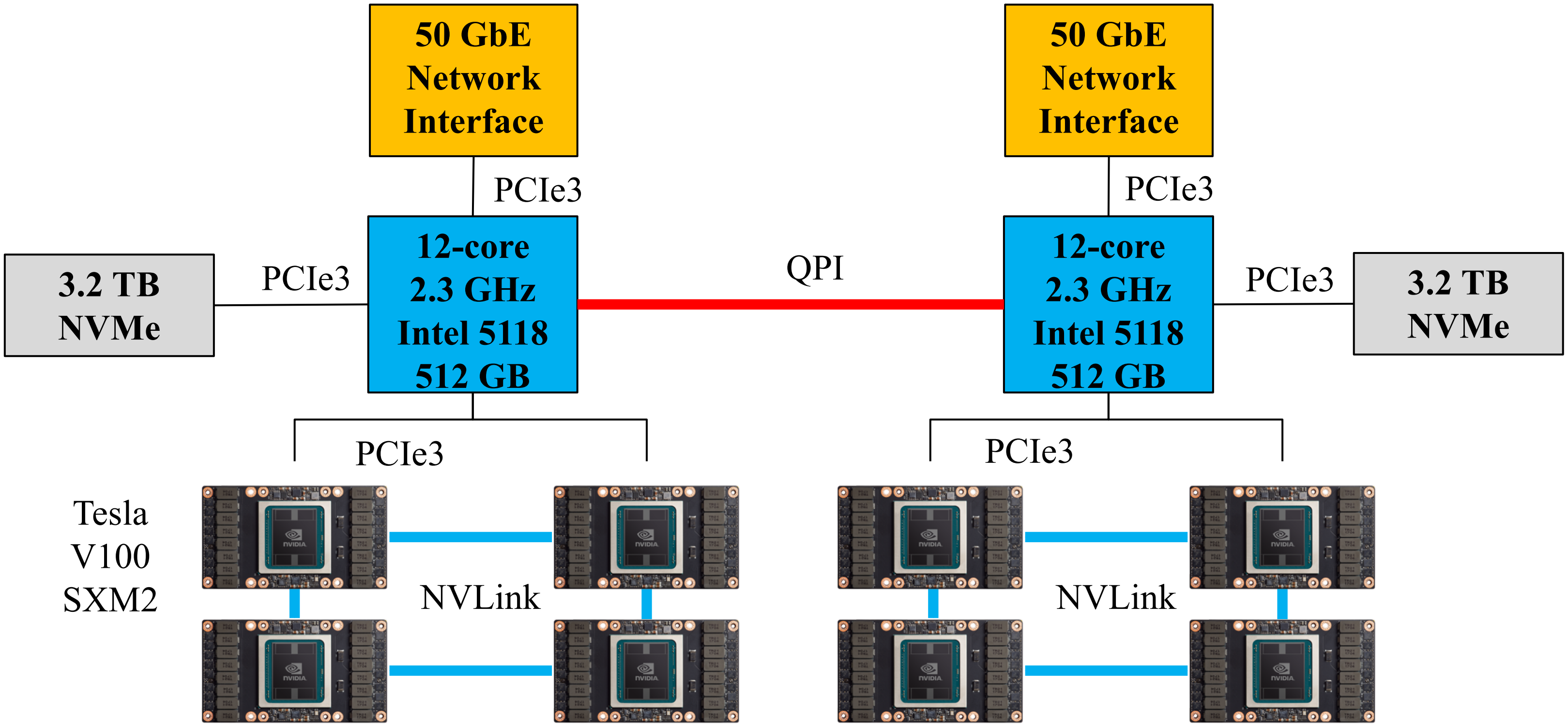
https://www.nvidia.com/en-us/data-center/tesla-v100/
https://images.nvidia.com/content/technologies/volta/pdf/437317-Volta-V100-DS-NV-US-WEB.pdf

Two Systems with 4 GPUs



Specifications	Total Performance
Double Precision Performance (64-bit)	31.2 teraFLOPS
Single Precision Performance (32-bit)	62.8 teraFLOPS
Tensor Performance (16-bit)	500 teraFLOPS

One Systems with 8 GPUs



Specifications	Total Performance
Double Precision Performance (64-bit)	62.4 teraFLOPS
Single Precision Performance (32-bit)	125.6 teraFLOPS
Tensor Performance (16-bit)	1,000 teraFLOPS

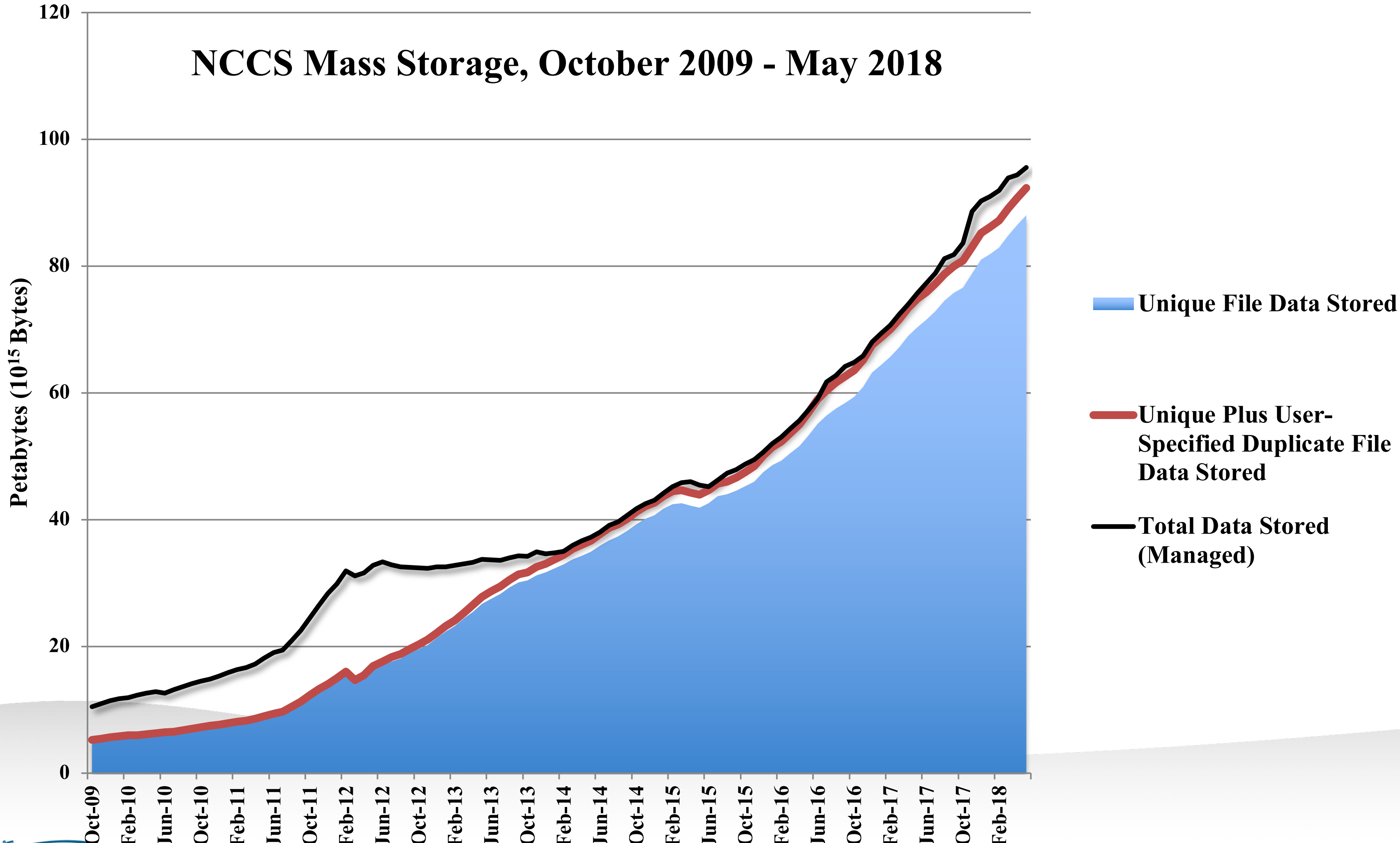
Storage/Analytics Architecture Evolution

and

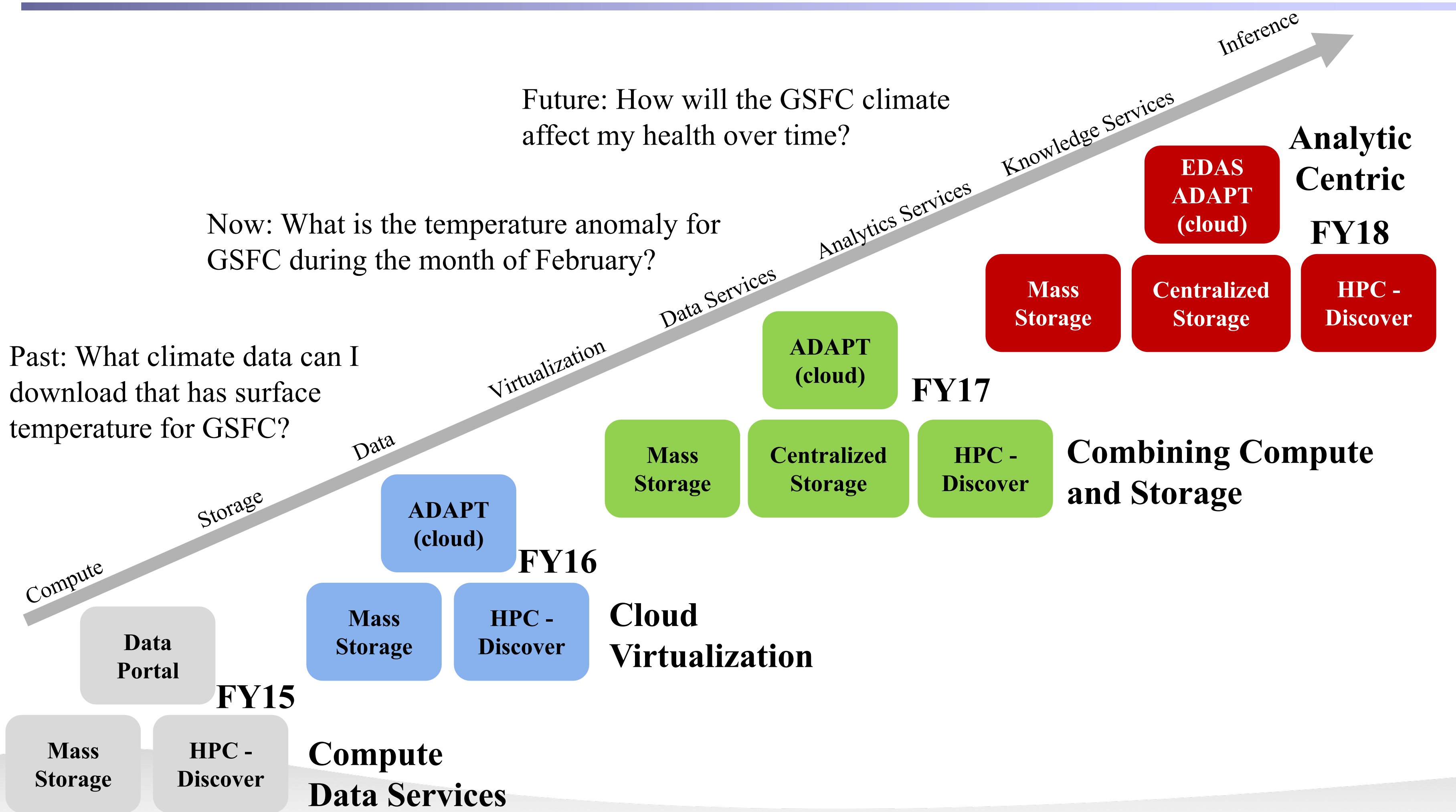
Data Management Plans

Dan Duffy and Laura Carriere

NCCS Mass Storage
October 2009 – May 2018



Evolution of Major NCCS Systems



FY19-FY20 Centralized Storage Concept

Current capacity: 9 PB
Planned capacity: 25 PB

Data Services

ADAPT

Analytics and Inference Based Services

ADAPT

High Performance Science Cloud

Centralized Storage
Data Lake Concept
Tiered Storage

Tier 1 – Spinning Disk and Compute

Tier 2 – Object Store (future)

Discover

High Performance Computing

Mass Storage
Legacy System
Becomes Read Only

What Goes into the Centralized Storage?

- Not an archive, not a scratch space
- Long term curated data sets, e.g.
  - Final data product, either public or private, optionally with an official or second copy at another location, e.g. a NASA archive, or on different media
  - Suitable for use as input by other projects
- Require a Data Management Plan – coming up next
- Write once/read many data sets
- Sharable through services
- Examples:
  - Model input data
  - Reanalysis and forecast data
  - Model Intercomparison and IPCC runs
  - Research runs, Nature Runs, high resolution simulations
  - Digital Globe
  - Other relevant observation data sets

Draft Data Management Plans

- Manage input data, intermediate files, in-house software, final products
- Plan for disposition of each data type
- Tied to allocation process in the future
  - Will submit along with allocation requests
  - No automatic centralized storage allocation
- Advantages: easier to locate, share, and delete
- NCCS will provide help developing plans

Data Management Tools – in Progress

- Track storage usage of running jobs
- Track trends in disk usage, both in MSS and online
- Identify duplication
- Provide usage within group quotas
- Find and delete within MSS –
  (data older than x days)
- Improved public usage statistics

Downsizing Your Mass Storage Data

- Know what you have in Mass Storage (see Tools, previous slide)
- Review & remove unneeded data, e.g.
  - Experiments no longer needed
  - Data already archived elsewhere
- Thanks for some recent large deletions:
  - Lena Marshak (1.6 PB)
  - dao_it (GMAO) (0.5 PB)
- Prize – Pizza for Petabytes (Deleted)!
- Email support@nccs.nasa.gov if you'd like to be included

Be a Part of the Conversation

- Make sure we understand your requirements
- Help influence the process
- We'll meet with your group or one-on-one
- Contact support@nccs.nasa.gov, Laura.Carriere@nasa.gov, Ellen.Salmon@nasa.gov

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

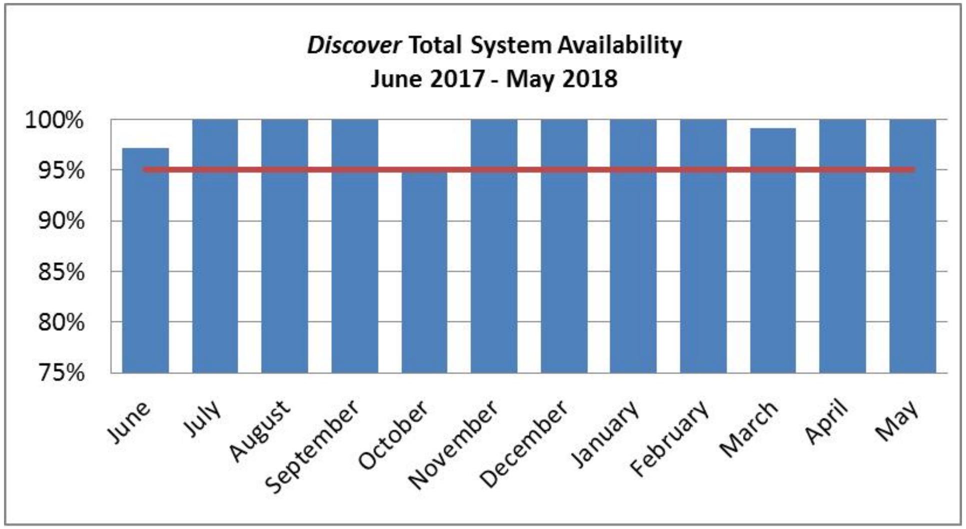
https://www.nccs.nasa.gov

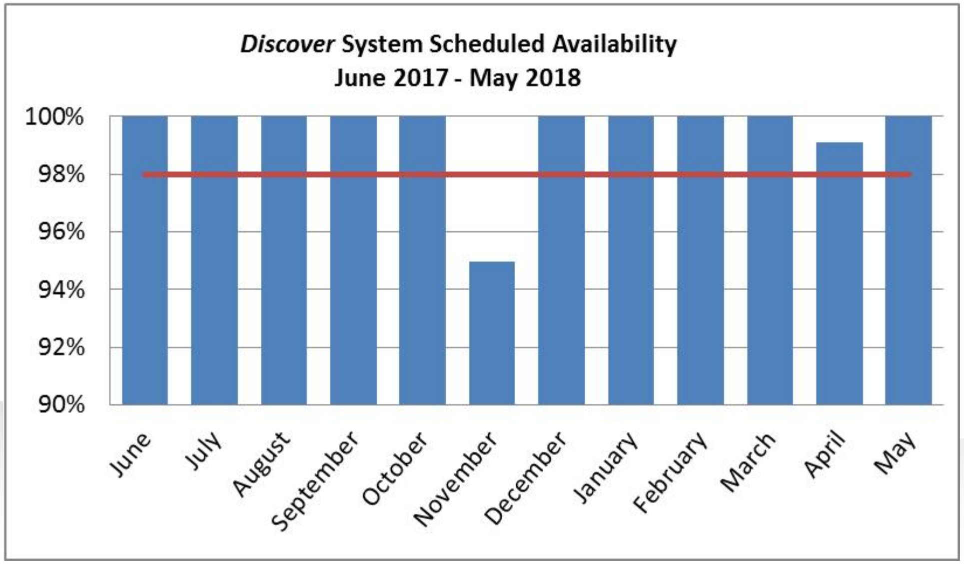
http://twitter.com/NASA_NCCS

Thank you

SUPPLEMENTAL SLIDES

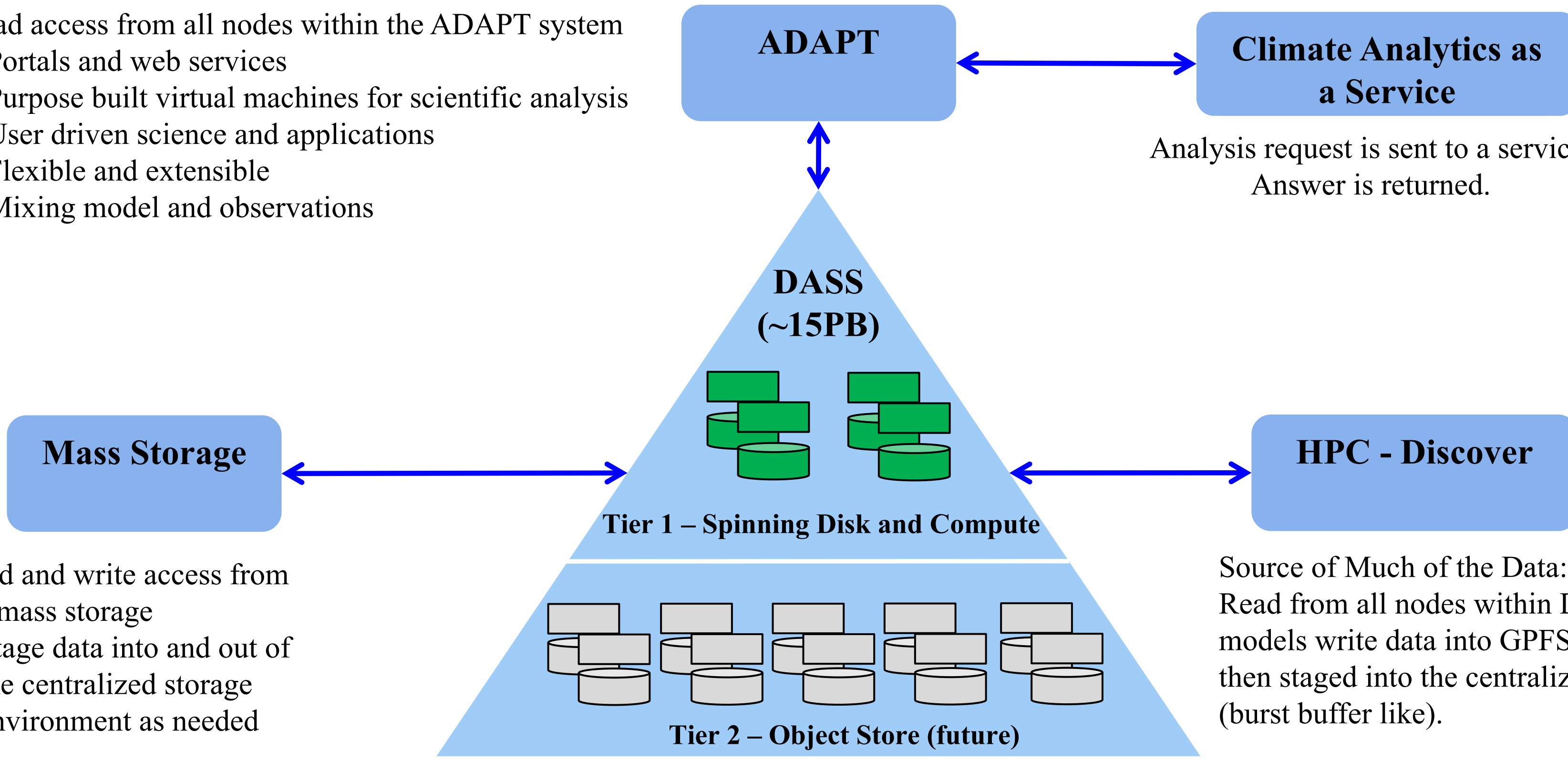
Discover System Availability





Data Analytics Storage System (DASS) Concept

Read access from all nodes within the ADAPT system
- Portals and web services
- Purpose built virtual machines for scientific analysis
- User driven science and applications
- Flexible and extensible
- Mixing model and observations



Analysis request is sent to a service. Answer is returned.

Read and write access from the mass storage
- Stage data into and out of the centralized storage environment as needed

Source of Much of the Data: Write and Read from all nodes within Discover – models write data into GPFS which is then staged into the centralized storage (burst buffer like).

Note that all the services will still have local file systems to enable local optimized writes and reads as needed within their respective security domains.

Discover Scratch Disk Evolution

Calendar	Description	Decommission	Total Usable Capacity (TB)
2012	Combination of DDN disks	None	3,960
Fall 2012	NetApp1: 1,800 by 3 TB Disk Drives; 5,400 TB RAW (prior to RAID protection)	None	9,360
Fall 2013	NetApp2: 1,800 by 4 TB Disk Drives; 7,200 TB RAW (prior to RAID protection)	None	16,560
Early 2015	DDN10: 1,680 by 6 TB Disk Drives, 10,080 TB RAW (prior to RAID protection)	DDNs 3, 4, 5	~26,000
Mid 2015	DDN11: 1,680 by 6 TB Disk Drives, 10,080 TB RAW (prior to RAID protection)	DDNs 7, 8, 9	~33,000
Mid 2016	DDN12: 1,680 by 6 TB Disk Drives, 10,080 TB RAW (prior to RAID protection)	None	~40,000
Early 2017	13+ PB RAW (prior to RAID protection)	TBD	~50,000

- Usable capacity differs from raw capacity for two reasons. First, the NCCS uses RAID6 (double parity) to protect against drive failures. This incurs a 20% overhead for the disk capacity. Second, the file system formatting is estimated to also need about 5% of the overall disk capacity. The total reduction from the RAW capacity to usable space is about 25%.

Data Management Plans & Centralized Storage Motivation

- Increasing costs due to uncontrolled growth in Mass Storage
  - 94 PB and growing by 1.5 PB/month
- Duplication of data due to individual storage practices
- Slow access to data in MSS, e.g. GMAO input data

Working towards the goal of a sustainable storage environment that embodies analytics

ADAPT Use Cases

- Arctic Boreal Vulnerability Experiment (ABoVE)
- CALET (CALorimetric Electron Telescope)
- High Mountain Asia Terrain (HiMAT)
- Asteroid Hunters – Near Earth Objects
- Biomass in South Sahara
- NCCS Data Services
- Laser Communications Relay Demonstration (LCRD) Project - FPGA simulations

Distributed Science Cloud Grid

CPUs Total:	**12689**
Hosts up:	**923**
Hosts down:	**0**

Current Load Avg (15, 5, 1m):
7%, 6%, 6%
Avg Utilization (last hour):
7%
Localtime:
2018-01-31 10:57