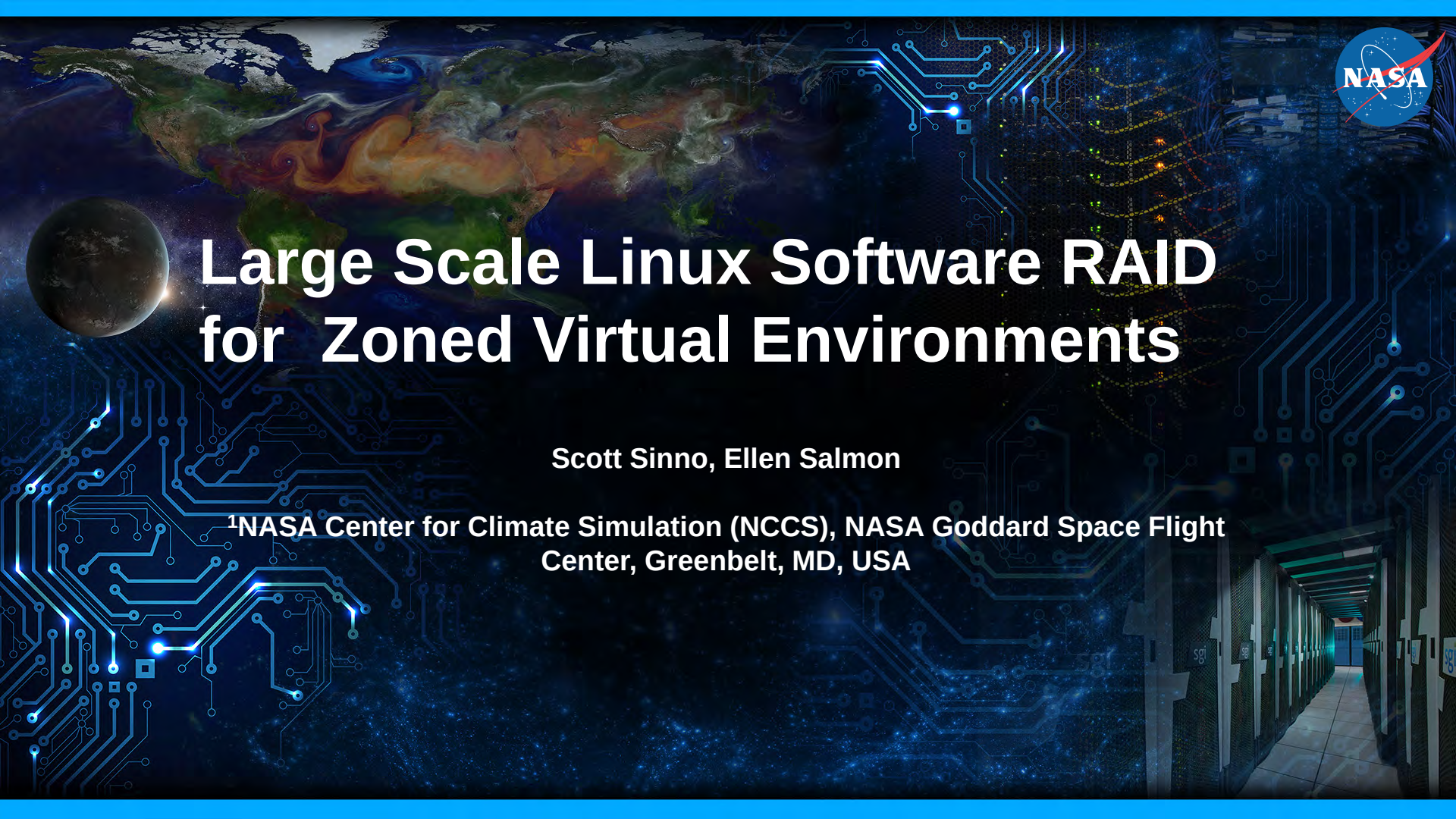


Large Scale Linux Software RAID for Zoned Virtual Environments

Scott Sinno, Ellen Salmon

[1]NASA Center for Climate Simulation (NCCS), NASA Goddard Space Flight Center, Greenbelt, MD, USA

NASA Center for Climate Simulation (NCCS)

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

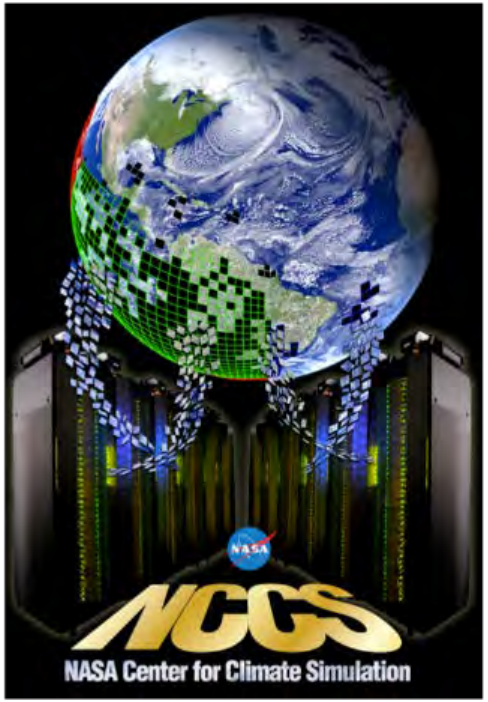
- •High-performance computing, cloud computing, data storage, and networking technologies
- •High-speed access to petabytes of Earth Science data
- •Collaborative data sharing, publication, and analysis services

Primary Customers (NASA Science)

- •NASA funded science projects can get access to these resources
- •Global Modeling and Assimilation Office (GMAO)
- •Land Information Systems (LIS)
- •Goddard Institute for Space Studies (GISS)
- •Variety of other Research and Development (R&D) and Engineering
  - »ABoVE, HiMAT, CALET, WFIRST

High-Performance Science

- •http://www.nccs.nasa.gov
- •Funded by the High End Computing (HEC) program under SMD
  - »Dr. Tsengdar Lee, Program Manager
- •Code 606.2 at NASA Goddard Space Flight Center in Greenbelt, MD.



Challenges

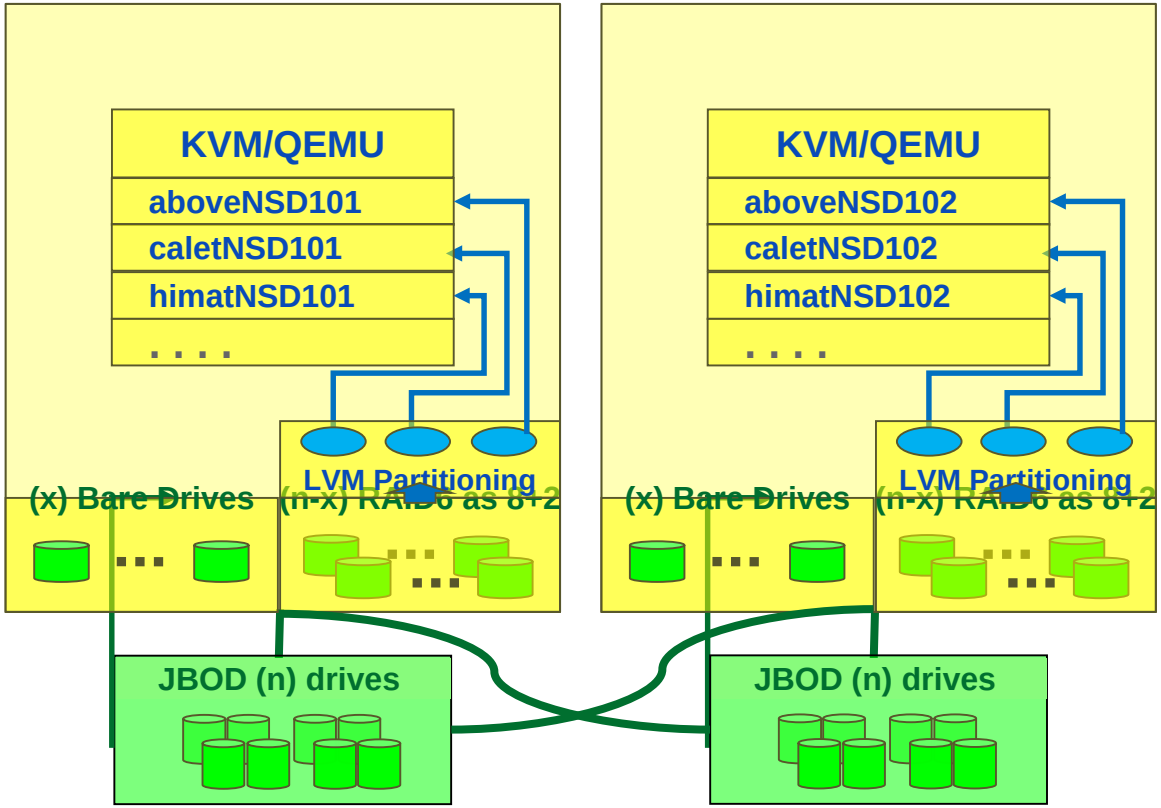
- Security : Zoned Architecture
  - No Hypervisor may have IP network connectivity to a virtual machine.
  - Segregation of services & nodes by risk class/category.
  - No writable shared storage between VM zones.
  - Hypervisors exist in their own Zone.

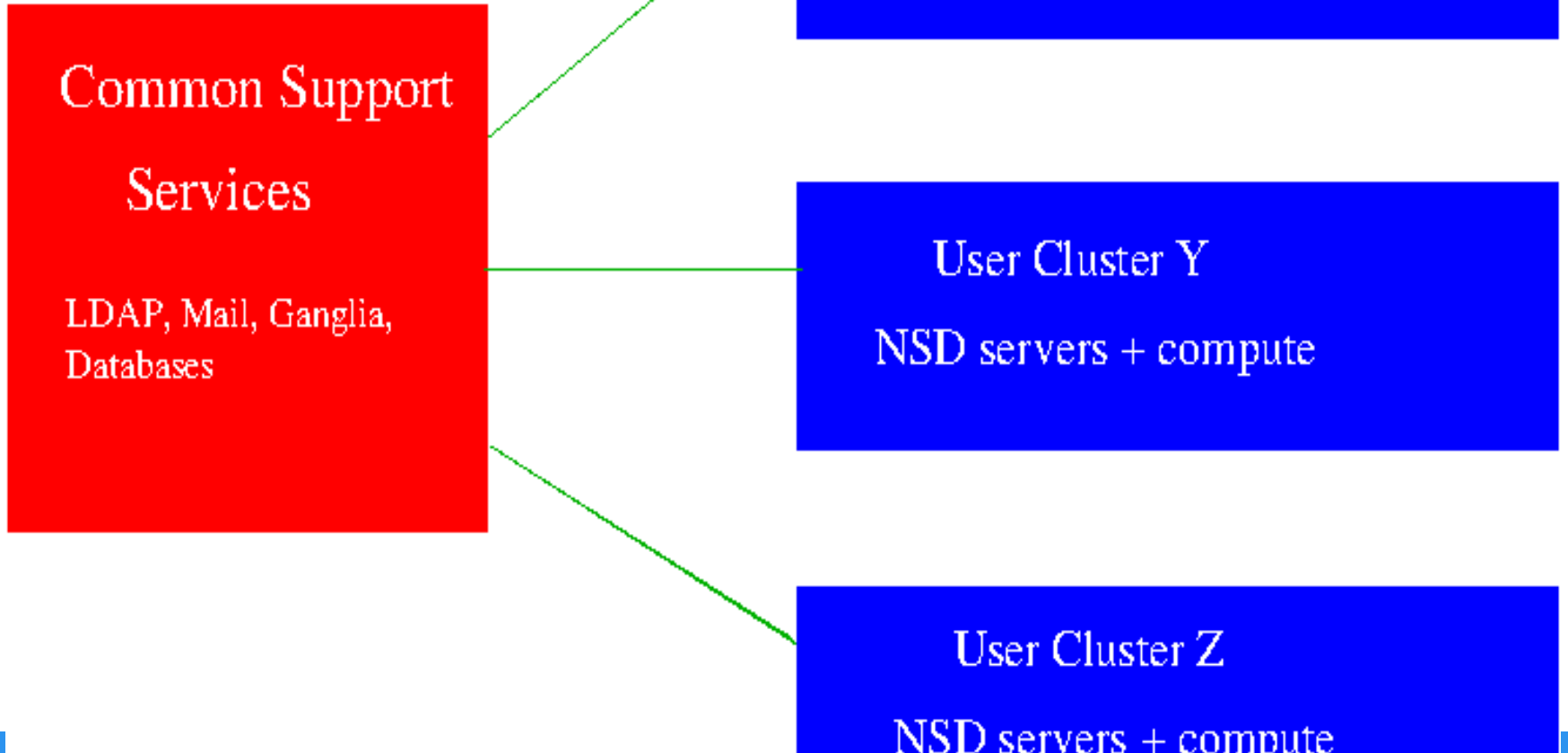
- Financial
  - Hand-me-down compute hardware.
  - JBOD only storage-budget.
  - Nearly exclusive leveraging of opensource software requisite.

Software Stack

- Hypervisors : CentOS6
- Custom KVM/QEMU build with tailored management scripts (virall, rvirsh).
- Puppet & PuppetDB
- Custom LTS Kernel (currently 4.4.49)
- Linux MDADM/LVM
- GPFS
- Mellanox OFED

Storage Serving Unit




Common Support Services
LDAP, Mail, Ganglia, Databases
User Cluster X
NSD servers + compute
User Cluster Y
NSD servers + compute
User Cluster Z
NSD servers + compute

Cluster Mgmt

- Custom script & database driven.
  - PuppetDB actively updates inventory & resource allocations on each invocation of puppet.
  - Database is aware of free & allocated resources on each Hypervisor.
  - Database is aware of all active & inactive-but-defined guests.
  - Provides automatic per-cluster SLURM configuration.
  - Provides alternative means of job execution via 'pupsh', a front-end wrapper to pdsh
  - 'rvirsh' and 'virall' provide means of managing bulk KVM/QEMU VM's
  - Zone & GPFS Cluster isolation prevents wide-scale deadlocks in an unpredictable job environment with low-memory nodes.
  - Guests may be marked as 'inactive' in DB, which permits leaving definitions in place but prevents guests from starting.

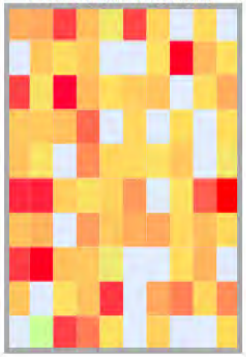
```
[ssinno@broker01 ~]pupsh
Usage : pupsh [-o "pdsh options"] "SQLquery" ["command"]

        pupsh is a front-end to pdsh.  You may pass any valid pdsh options
along via the '-o' flag.

        All SQL columns & associated values are derived from facter.
Supported attributes are :

hostname domain fqdn operatingsystem operatingsystemrelease lsbdistcodename
kernel kernelversion kernelrelease architecture manufacturer virtual is_virtual
ht_enabled hardwaremodel hardwareisa processorcount physicalprocessorcount
processormodel processortype memorysize_mb cluster gl_cluster vzparent
kvmparent kvmguests vzguests is_forward_facing has_ib uncommitted_memory
uncommitted_cores

If no command is provided, a list of hostnames that match the query
will be returned instead.

Some helpful examples:

#Execute on all nodes matching hostname 'ssinno'
pupsh "hostname ~ 'ssinno' " "uname -a"

#Execute no more than 4 nodes concurrently matching hostname 'ssinno'
pupsh -o "-f 4" "hostname ~ 'ssinno' " "uname -a"

#Execute on hosts matching 'ssinno' with at least 16GB memory
pupsh "hostname ~ 'ssinno' and memorysize_mb >= 16384" "uname -a"

#Execute on hosts matching 'ssinno' naming output by hostname
pupsh "hostname ~ 'ssinno'" "myjob >| /att/nobackup/myuserid/%h.out "
[ssinno@broker01 ~]
[ssinno@broker01 ~]
[ssinno@broker01 ~]
```

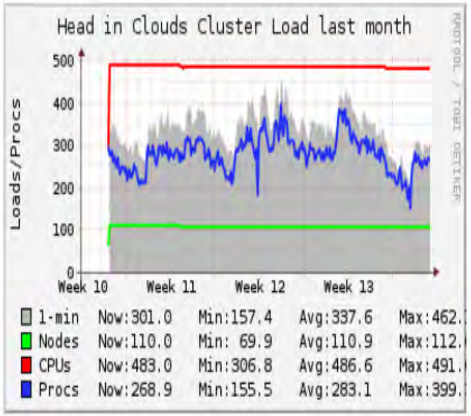
Overview of Head in Clouds @ 2017-04-05 10:40

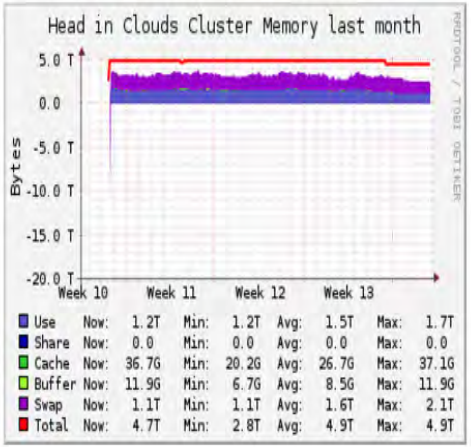
CPUs Total: **483**
Hosts up: **110**
Hosts down: **1**

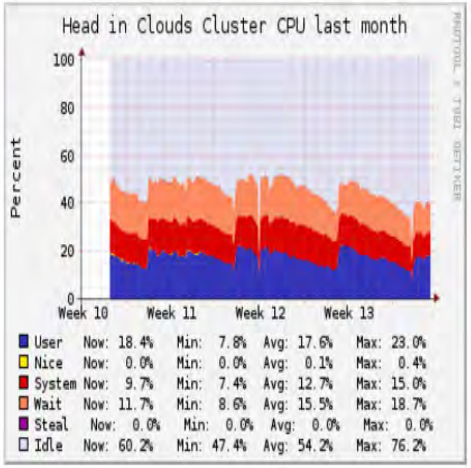
Current Load Avg (15, 5, 1m):
55%, 56%, 57%
Avg Utilization (last month):
69%

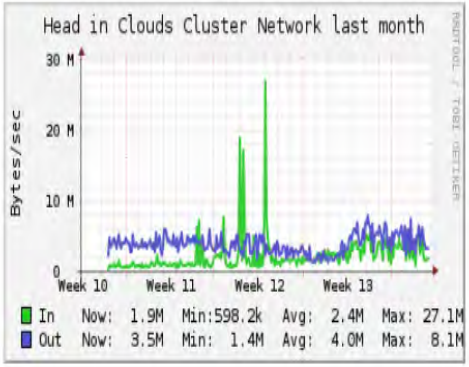
Server Load Distribution











Resources & Services

- Hardware
  - 15PB Raw disk, mix of 4TB/6TB/8TB
  - ~ 12PB usable disk (RAID/Replication)
  - 364 Hypervisors, currently hosting 674 guests (2017/05/16)

- Static Services (Data Service Zone)
  - ArcGIS
  - Earth System Grid
  - FTP/HTTP/Tomcat

Questions?