IBM DCS Storage – the Ideal Building Block for Flexible Spectrum Scale

Kumaran Rajaram, Staff Engineer
IBM, Storage Benchmarking Team

Dexter Pham, Consulting Engineer
DCS Technical Specialists

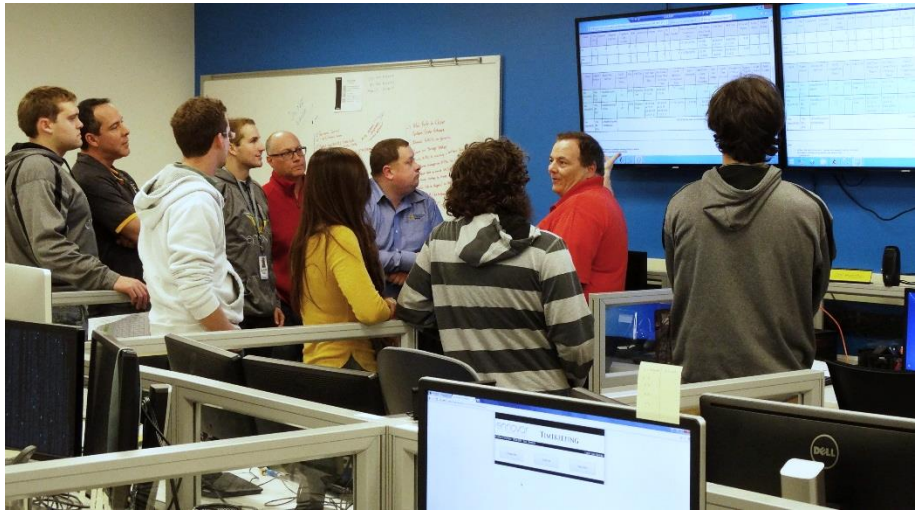
Matt Forney, Research and Technology Director
Ennovar, Institute of Emerging Technologies and Market Solutions Wichita State University

Ennovar Assets, Benchmarks, Best Practices, and Assistance

Ennovar Spectrum Scale Solutions Center

Institute of Emerging Technologies and Marketing Solutions at Wichita State University

- Hands-on student applied learning with Spectrum Scale
- Spectrum Scale and Spectrum Archive (SME) subject matter experts
- Consulting—including benchmarks, best practices and technical marketing services.
- Remote Customer Demo Online Access
  - Customer defined use case and application environments.
  - Online Spectrum Scale Solutions Lab Portal provides VPN customer access on Ennovar's high-performance FSS configurations.



WICHITA STATE UNIVERSITY

Team and Collaboration

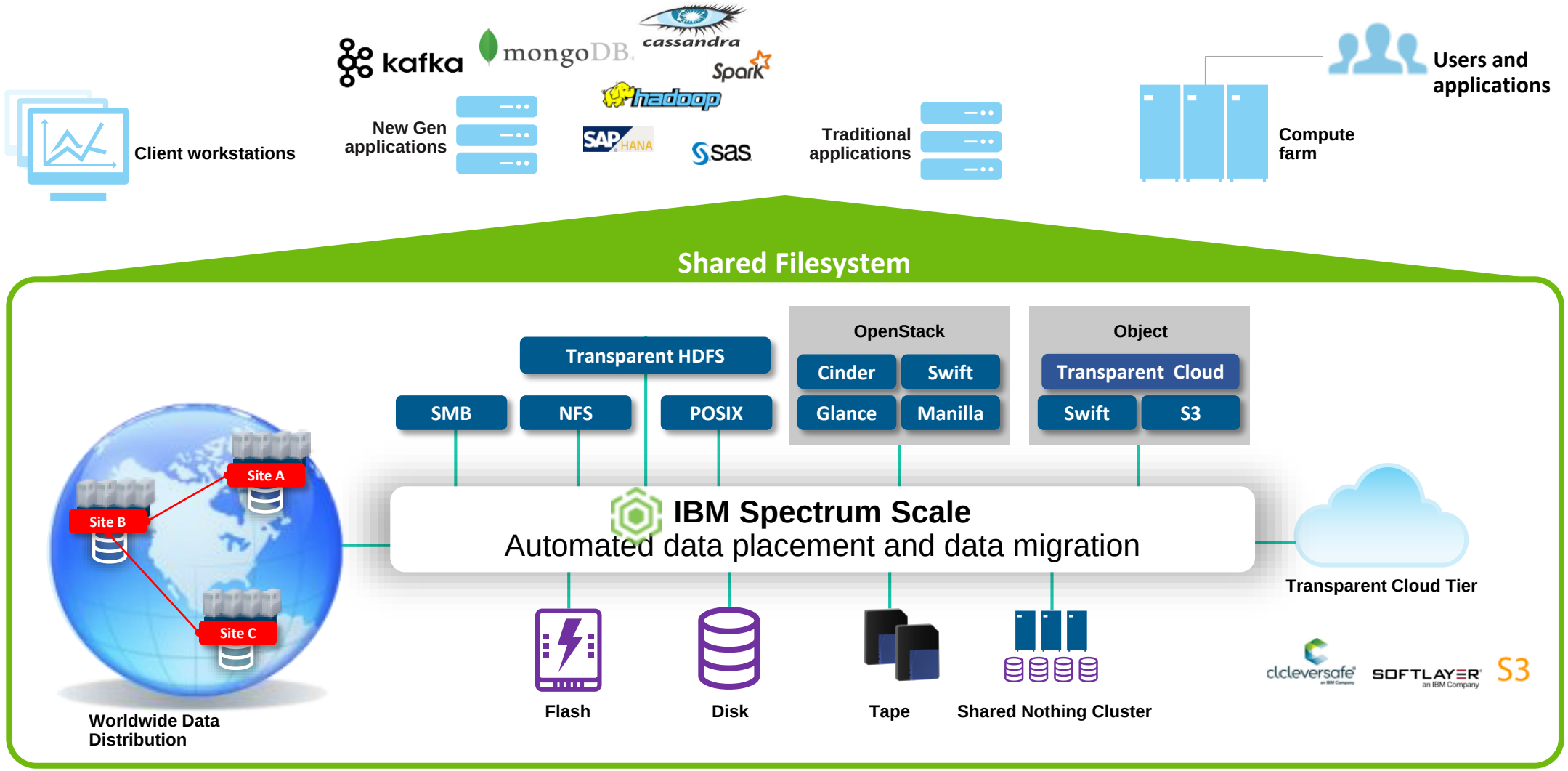
Olga Yiparaki, Chief Engineer, Storage Performance
Vernon Miller, Spectrum Scale performance
Jay Vaddi, Spectrum Scale Performance
Kumaran "Kums" Rajaram, Spectrum Scale Performance

Matt Forney, Research and Technology Director
Alan Snyder, Technical Marketing Director
Dexter Pham, Consulting Engineer
Tom Rose, Technical Manager
Joel Hatcher, Technical Lead

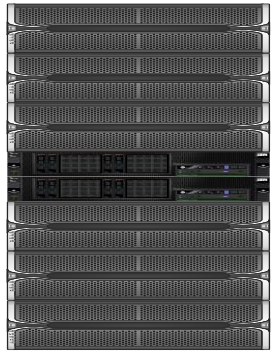
Flexible Spectrum Scale Solution with DCS3860 storage

- Focused on IBM Spectrum Scale with DCS3860 Gen 2 storage systems
  - Develop storage reference architectures (building blocks)
    - High-performance (*today's update*)
    - Balanced-workload (*future work*)
    - High-capacity (*future work*)
  - Develop and run benchmarks for each architectural model using
    - IBM DCS3860 Gen 2 storage systems
    - Intel-based NSD servers
  - Develop Implementation Guide
    - Define modular building blocks for I/O and capacity expansion
    - Hardware selection criteria and tuning guidelines
    - Detailed system build instructions
    - Benchmark scripts and test results on reference models

Spectrum Scale: Unleash new storage economics on a global scale

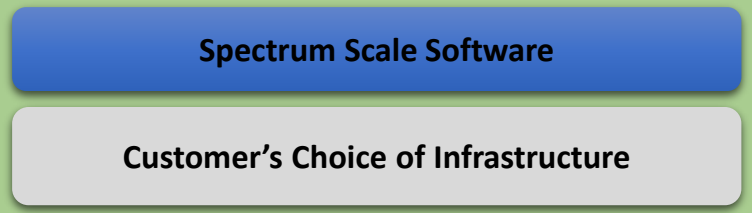


Three Options for Licensing Spectrum Scale



IBM's Elastic Storage Server (ESS)

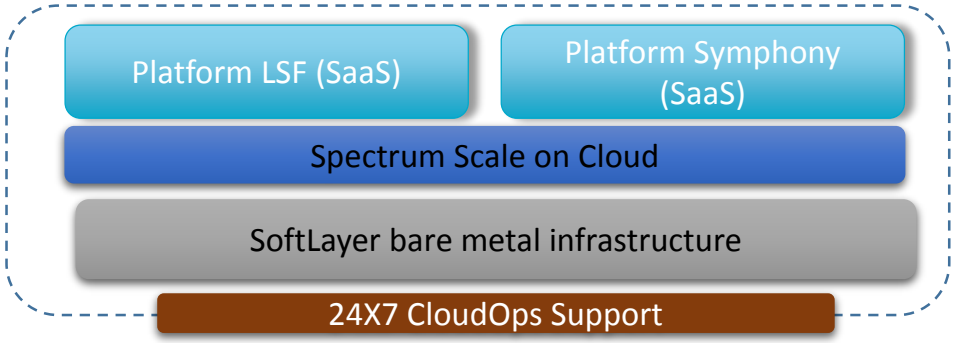
An Integrated IBM Offering



Flexible Spectrum Scale (FSS)

IBM Spectrum Scale Software on flexible hardware

Software + x86/Power + storage



Cloud Service

Ready to use, Spectrum Scale on the Cloud

IBM Spectrum Scale on Cloud

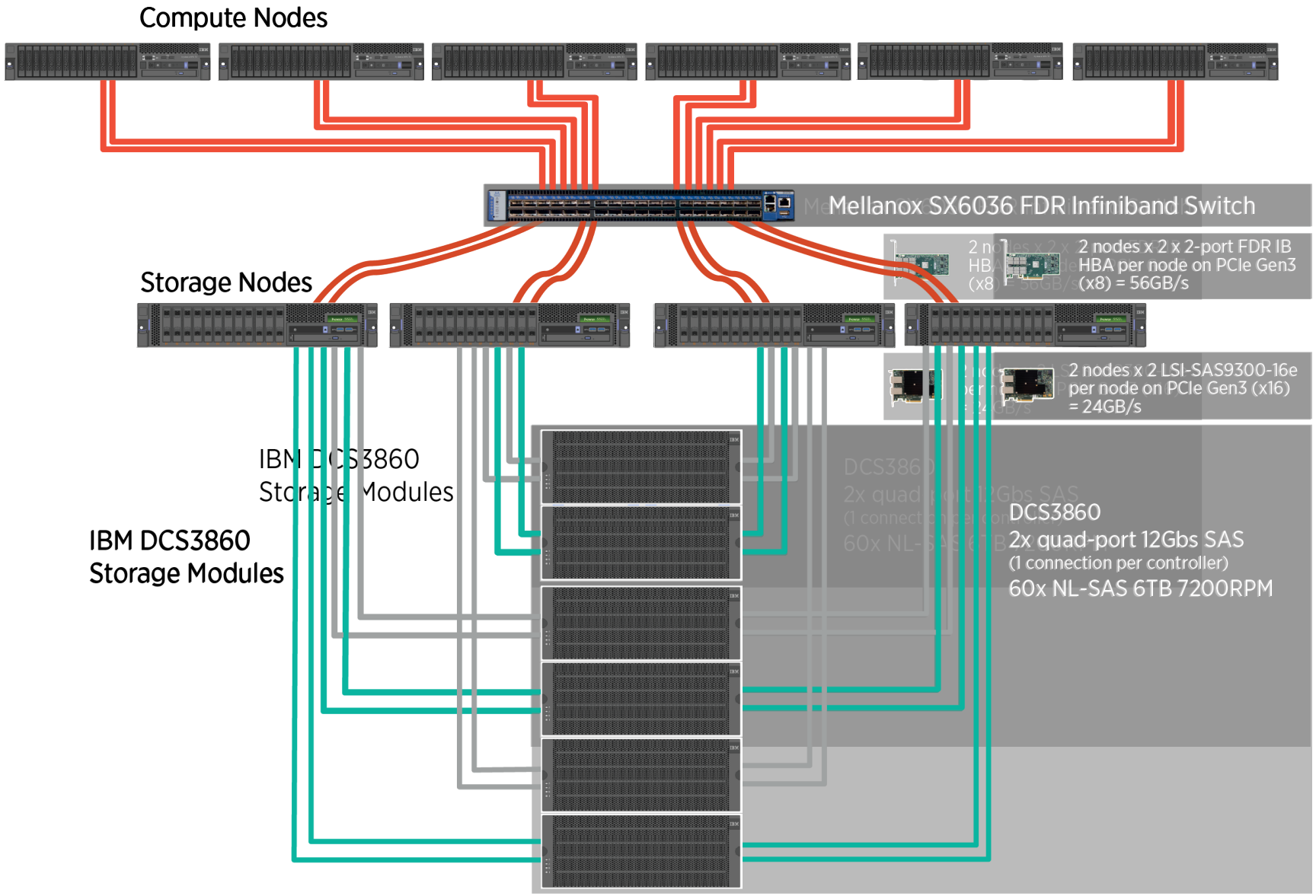
On IBM SoftLayer Cloud

Flexible Spectrum Scale Storage Building Block based on DCS3860

	High-Performance Building Block
Controllers/ NSD Servers	Up to 4 single-proc Intel E5-1650 v 3 6-core “Haswell"
Storage Subsystem	1 -- 6 systems, DCS3860-G2 model 1 Controller Enclosure per system 60 6TB NL-SAS per system, 6 RAID6 arrays of 10 drives each (8 + P + Q) 12Gb SAS connections (max. 360 disks, 1.7 PB usable capacity)

- Rack space - 30U (free space to add SSD enclosures)
- Storage enclosure – 4U 60 x 3.5” drive slots, 6 x 10 drive RAID6 = approx. 288TB usable capacity
- 1U Mellanox top-of-rack IB switch + 1U 1Gb Ethernet management switch
- DCS3860-G2 allows up to 5 expansion enclosures

Building Blocks – Intel x86 NSD Servers – 1 to 6 storage systems



Performance Updates

Storage	1 x DCS3860	6 x DCS3860	comments
NSD clients	10 total: 6 x Dual Socket Intel Xeon CPU E6540 @ 2.00GHz + 4 x Dual Socket Intel Xeon CPU E5-2650 v4 @ 2.20GHz	10 total: 6 x Dual Socket Intel Xeon CPU E6540 @ 2.00GHz + 4 x Dual Socket Intel Xeon CPU E5-2650 v4 @ 2.20GHz	
NSD servers	2 single-proc Intel E5-1650, v 3 6-core “Haswell"	4 single-proc Intel E5-1650, v 3 6-core “Haswell"	
Capacity	288 TB	1.7 PB	Net Capacity after RAID6
Drives	60 NL SAS x 6 TB each	6 x 60 NL SAS x 6 TB each	
Read GB/s	8.6 GB/s	43 GB/s	• Read BW can scale up to 52 GB/s but limited by IB network • Measured with both IOR and GPFSperf, same performance
Write GB/s	4.2 GB/s	25 GB/s	• Linear scaling, sustainable bandwidth • Measured with both IOR and GPFSperf, same performance
70/30 Read/Write GB/s	5.9 GB/s (sequential) 6.9 GB/s (random)	34 GB/s (sequential) 40 GB/s (random)	• Linear scaling, sustainable bandwidth

Performance measured with IOR and GPFSperf. Note that performance depends on client configuration and good Interconnect and can vary between environments. Performance is not guaranteed; rather it is a demonstration of the technical capabilities of this cluster under good conditions. See backup pages with settings and configuration details.

Summary

- Sustainable, high performance is achieved with flexible Spectrum Scale
- Performance scalability: linear scaling as storage expands to grow capacity, while maintaining a single file system
- Completed FSS with DCS3860 Implementation Guide

Next Steps

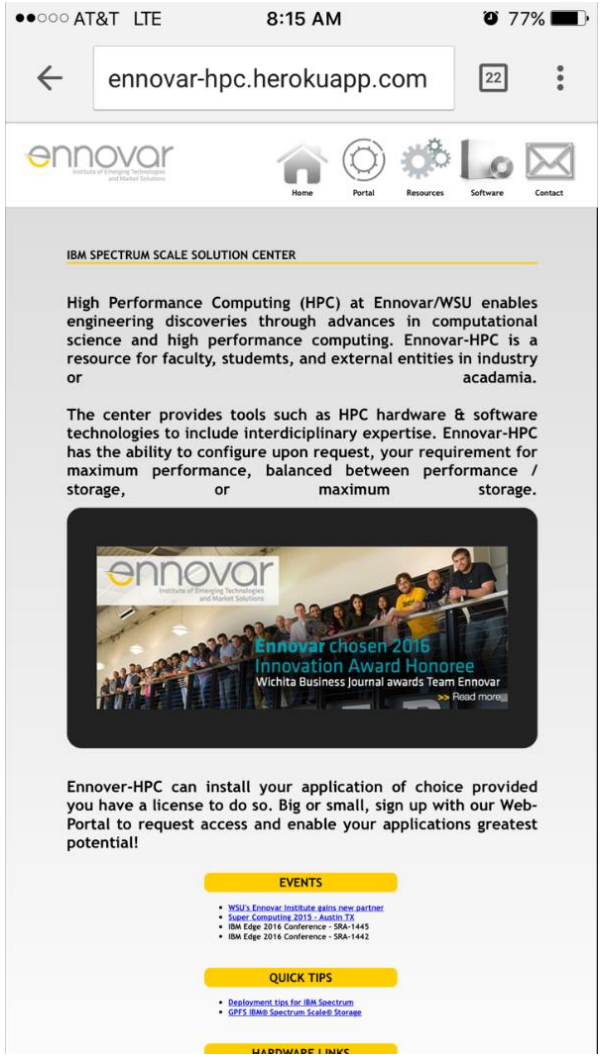
- Performance using different building blocks: `balanced workload' and `high-capacity' building blocks
- Performance with SPEC SFS and other benchmarks
- Spectrum Scale Solutions Center Lab Portal

Ennovar IBM Spectrum Scale Solutions Center Lab Portal

- Provides Spectrum Scale partners and end-users an opportunity for hands on experience performing upgrade, interop and benchmark testing in a lab environment
  - Eliminates risk to production environments
  - Spectrum Scale and SAN/NAS SMEs on site
    - Assist or perform testing
    - Provide or configure hardware changes and/or customizations
    - Provide Spectrum Scale and SAN/NAS expertise
  - Lab available 24x7 with personnel onsite between 8:00 and 5:00 US Central (available after hours/weekends via pre-arranged agreements)

- Go to https://ennovar-hpc.herokuapp.com/
  - Fill out the online registration form to request a portal account
  - Download the VPN client for accessing the lab
  - Ennovar will automatically receive an email once you register and will contact you via email or phone in order to provide you the necessary credentials for logging in to the lab VPN

- Contact the Director Matt Forney or SRA Lab Manager Tom Rose and a conference call will be arranged to discuss access and/or provide assistance
  - Email: Matt.Forney@Wichita.edu
  - Email: Thomas.Rose@Wichita.edu



Backup and References

Configuration with Spectrum Scale 4.2.0.3

mmlsconfig

```
maxMBpS 14336
pagepoolMaxPhysMemPct 75
nsdbufspace 70
workerThreads 1024
scatterBufferSize 256k
verbsRdma enable
verbsRdmaSend yes

[nsds]
nsdMultiQueueType 1
nsdMaxWorkerThreads 1k
nsdMinWorkerThreads 16
maxblocksize 16m
nsdMultiQueue 64
nsdThreadsPerQueue 12
nsdThreadsPerDisk 12
pagepool 48g
[clients] pagepool 8G
```

GPFS Version

```
=== mmdiag: version ===
Current GPFS build: "4.2.0.3 ".
Built on May  4 2016 at 09:29:46
```

mmlscluster

```
GPFS cluster information
========================
  GPFS cluster name:         ennovar1.sgi1_gpfs
  GPFS cluster id:           13168939376820416695
  GPFS UID domain:           ennovar1.sgi1_gpfs
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address      Admin node name  Designation
-------------------------------------------------------------------------
   1   sgi1_gpfs         192.168.212.20  sgi1_gpfs        quorum-manager
   2   sgi2_gpfs         192.168.212.21  sgi2_gpfs        quorum-manager
   3   sgi3_gpfs         192.168.212.22  sgi3_gpfs        quorum-manager
   4   sgi4_gpfs         192.168.212.23  sgi4_gpfs        quorum-manager
   5   compute1_gpfs     192.168.212.10  compute1_gpfs    quorum
   6   compute2_gpfs     192.168.212.11  compute2_gpfs
   7   compute3_gpfs     192.168.212.12  compute3_gpfs
   8   compute4_gpfs     192.168.212.13  compute4_gpfs
   9   compute5_gpfs     192.168.212.14  compute5_gpfs
  10   compute6_gpfs     192.168.212.15  compute6_gpfs
  11   lenovo1_gpfs      192.168.212.24  lenovo1_gpfs
  12   lenovo2_gpfs      192.168.212.25  lenovo2_gpfs
  13   lenovo3_gpfs      192.168.212.26  lenovo3_gpfs
  14   lenovo4_gpfs      192.168.212.27  lenovo4_gpfs
```

mmlsfs

```
flag                value                    description
------------------- ------------------------ -----------------------------------
 -f                 524288                   Minimum fragment size in bytes
 -i                 4096                     Inode size in bytes
 -I                 32768                    Indirect block size in bytes
 -m                 1                        Default number of metadata replicas
 -M                 2                        Maximum number of metadata replicas
 -r                 1                        Default number of data replicas
 -R                 2                        Maximum number of data replicas
 -j                 cluster                  Block allocation type
 -D                 nfs4                     File locking semantics in effect
 -k                 all                      ACL semantics in effect
 -n                 14                       Estimated number of nodes that will
                                                 mount file system
 -B                 16777216                 Block size
 -Q                 none                     Quotas accounting enabled
                    none                     Quotas enforced
                    none                     Default quotas enabled
 --perfileset-quota No                       Per-fileset quota enforcement
 --filesetdf        No                       Fileset df enabled?
 -V                 15.01 (4.2.0.0)          File system version
 -z                 No                       Is DMAPI enabled?
 -L                 16777216                 Logfile size
 -E                 Yes                      Exact mtime mount option
 -S                 No                       Suppress atime mount option
 -K                 whenpossible             Strict replica allocation option
 --fastea           Yes                      Fast external attributes enabled?
 --encryption       No                       Encryption enabled?
 --inode-limit      134217728                Maximum number of inodes
 --log-replicas     0                        Number of log replicas
 --is4KAligned      Yes                      is4KAligned?
 --rapid-repair     Yes                      rapidRepair enabled?
 --write-cache-threshold 0                   HAWC Threshold (max 65536)
```

mmlsnsd

```
File system   Disk name    NSD servers
---------------------------------------------
gpfs1         S1V00        sgi1_gpfs,sgi2_gpfs
gpfs1         S1V01        sgi2_gpfs,sgi1_gpfs
gpfs1         S1V02        sgi2_gpfs,sgi1_gpfs
gpfs1         S1V03        sgi1_gpfs,sgi2_gpfs
gpfs1         S1V04        sgi1_gpfs,sgi2_gpfs
gpfs1         S1V05        sgi2_gpfs,sgi1_gpfs
gpfs1         S2V00        sgi2_gpfs,sgi1_gpfs
gpfs1         S2V01        sgi1_gpfs,sgi2_gpfs
gpfs1         S2V02        sgi1_gpfs,sgi2_gpfs
gpfs1         S2V03        sgi2_gpfs,sgi1_gpfs
gpfs1         S2V04        sgi2_gpfs,sgi1_gpfs
gpfs1         S2V05        sgi1_gpfs,sgi2_gpfs
gpfs1         S3V00        sgi3_gpfs,sgi4_gpfs
gpfs1         S3V01        sgi4_gpfs,sgi3_gpfs
gpfs1         S3V02        sgi4_gpfs,sgi3_gpfs
gpfs1         S3V03        sgi3_gpfs,sgi4_gpfs
gpfs1         S3V04        sgi3_gpfs,sgi4_gpfs
gpfs1         S3V05        sgi4_gpfs,sgi3_gpfs
gpfs1         S4V00        sgi4_gpfs,sgi3_gpfs
gpfs1         S4V01        sgi3_gpfs,sgi4_gpfs
gpfs1         S4V02        sgi3_gpfs,sgi4_gpfs
gpfs1         S4V03        sgi4_gpfs,sgi3_gpfs
gpfs1         S4V04        sgi4_gpfs,sgi3_gpfs
gpfs1         S4V05        sgi3_gpfs,sgi4_gpfs
gpfs1         S5V00        sgi1_gpfs,sgi2_gpfs
gpfs1         S5V01        sgi2_gpfs,sgi1_gpfs
gpfs1         S5V02        sgi2_gpfs,sgi1_gpfs
gpfs1         S5V03        sgi1_gpfs,sgi2_gpfs
gpfs1         S5V04        sgi1_gpfs,sgi2_gpfs
gpfs1         S5V05        sgi2_gpfs,sgi1_gpfs
gpfs1         S6V00        sgi4_gpfs,sgi3_gpfs
gpfs1         S6V01        sgi3_gpfs,sgi4_gpfs
gpfs1         S6V02        sgi3_gpfs,sgi4_gpfs
gpfs1         S6V03        sgi4_gpfs,sgi3_gpfs
gpfs1         S6V04        sgi4_gpfs,sgi3_gpfs
gpfs1         S6V05        sgi3_gpfs,sgi4_gpfs
```

NSD clients & servers

	Spectrum Scale NSD Client Nodes	Spectrum Scale NSD Server Nodes
Operating System	CentOS v7.2	CentOS v 7.2
Processing Elements	6 x Dual Socket Intel Xeon CPU E6540 @ 2.00GHz	4 x Single Socket Intel Xeon CPU E5-1650 v3 @ 3.50GHz
	4 x Dual Socket Intel Xeon CPU E5-2650 v4 @ 2.20GHz	
RAM Size	128 GiB	64 GiB

DSC3860 Storage Configuration

- Controller Firmware Version: 08.20.21.00
- RAID Level: 6
- Segment Size: 512 KB
- Read Cache: Enabled
- Write Cache: Enabled
  - Write cache without batteries: Disabled
  - Write cache with mirroring: Enabled
- Flush write cache after (in seconds): 10.00
- Dynamic cache read prefetch: Disabled

Six Storage Systems

IOR Read Test

```
Run began: Tue Nov  8 23:06:59 2016
Command line used: /usr/local/bin/IOR -i 10 -d 5 -r -eg -E -F -k -t 1M -b 32g -o /gpfs1/gpfsperf/seq
Machine: Linux compute1

Summary:
        api                = POSIX
        test filename      = /gpfs1/gpfsperf/seq
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 500 (50 per node)
        repetitions        = 10
        xfersize           = 1 MiB
        blocksize          = 32 GiB
        aggregate filesize = 16000 GiB

Operation  Max (MiB)  Min (MiB)  Mean (MiB)   Std Dev  Max (OPs)  Min (OPs)  Mean (OPs)   Std Dev  Mean (s)
---------  ---------  ---------  ----------   -------  ---------  ---------  ----------   -------  --------
read        42246.53   41456.85    41804.58    228.13   42246.53   41456.85    41804.58    228.13 391.93047   EXCEL

Max Read:  42246.53 MiB/sec (44298.70 MB/sec)

Run finished: Wed Nov  9 00:13:06 2016
```

Six Storage Systems

IOR Write Test

```
Run began: Wed Nov  9 03:03:28 2016
Command line used: /usr/local/bin/IOR -i 10 -d 5 -w -eg -E -F -k -t 1M -b 32g -o /gpfs1/gpfsperf/seq
Machine: Linux compute1

Summary:
        api                = POSIX
        test filename      = /gpfs1/gpfsperf/seq
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 20 (2 per node)
        repetitions        = 10
        xfersize           = 1 MiB
        blocksize          = 32 GiB
        aggregate filesize = 640 GiB

Operation  Max (MiB)  Min (MiB)  Mean (MiB)   Std Dev  Max (OPs)  Min (OPs)  Mean (OPs)   Std Dev  Mean (s)
---------  ---------  ---------  ----------   -------  ---------  ---------  ----------   -------  --------
write       24780.20   23755.76    24231.24    306.19   24780.20   23755.76    24231.24    306.19  27.05041   EXCEL

Max Write: 24780.20 MiB/sec (25983.92 MB/sec)

Run finished: Wed Nov  9 03:08:49 2016
```

One Storage System

IOR Write Test

```
Run began: Wed Nov  9 22:31:42 2016

Command line used: /usr/local/bin/IOR -i 10 -d 5 -w -eg -E -F -k -t 1M -b 32g -o /gpfs1/gpfsperf/seq
Machine: Linux compute1

Summary:
        api                = POSIX
        test filename      = /gpfs1/gpfsperf/seq
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 10 (1 per node)
        repetitions        = 10
        xfersize           = 1 MiB
        blocksize          = 32 GiB
        aggregate filesize = 320 GiB

Operation  Max (MiB)  Min (MiB)  Mean (MiB)   Std Dev  Max (OPs)  Min (OPs)  Mean (OPs)   Std Dev  Mean (s)
---------  ---------  ---------  ----------   -------  ---------  ---------  ----------   -------  --------
write        4009.32    3931.79     3976.78     22.79    4009.32    3931.79     3976.78     22.79  82.40101   EXCEL

Max Write: 4009.32 MiB/sec (4204.08 MB/sec)

Run finished: Wed Nov  9 22:46:15 2016
```

One Storage System

IOR Read Test

```
Run began: Wed Nov  9 20:21:26 2016

Command line used: /usr/local/bin/IOR -i 10 -d 5 -r -eg -E -F -k -t 1M -b 32g -o /gpfs1/gpfsperf/seq
Machine: Linux compute1

Summary:
        api                = POSIX
        test filename      = /gpfs1/gpfsperf/seq
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 10 (1 per node)
        repetitions        = 10
        xfersize           = 1 MiB
        blocksize          = 32 GiB
        aggregate filesize = 320 GiB

Operation  Max (MiB)  Min (MiB)  Mean (MiB)   Std Dev  Max (OPs)  Min (OPs)  Mean (OPs)   Std Dev  Mean (s)
---------  ---------  ---------  ----------   -------  ---------  ---------  ----------   -------  --------
read         8259.74    8055.46     8194.47     57.18    8259.74    8055.46     8194.47     57.18  39.98991   EXCEL

Max Read:  8259.74 MiB/sec (8660.97 MB/sec)

Run finished: Wed Nov  9 20:28:56 2016
```