The SAGE2 project has received funding from the European Union's Horizon2020 Research & Innovation Programme under grant agreement 800999

SEAGATE

CORTX/Motr in Sage2

March 2021

Seagate Systems EU R&D

Ganesan.Umanesan@seagate.com (Sr staff software Eng)
Andriy.Tkachuk@seagate.com (Staff Software Eng)
Sai.Narasimhamurthy@seagate.com (Eng Director)

One Storage System to rule them all!

<u>Extreme Computing</u>

Changing I/O Needs

HDDs cannot Keep Up

Sage2

<u>Big Data Analysis</u>

Avoid Data Movements

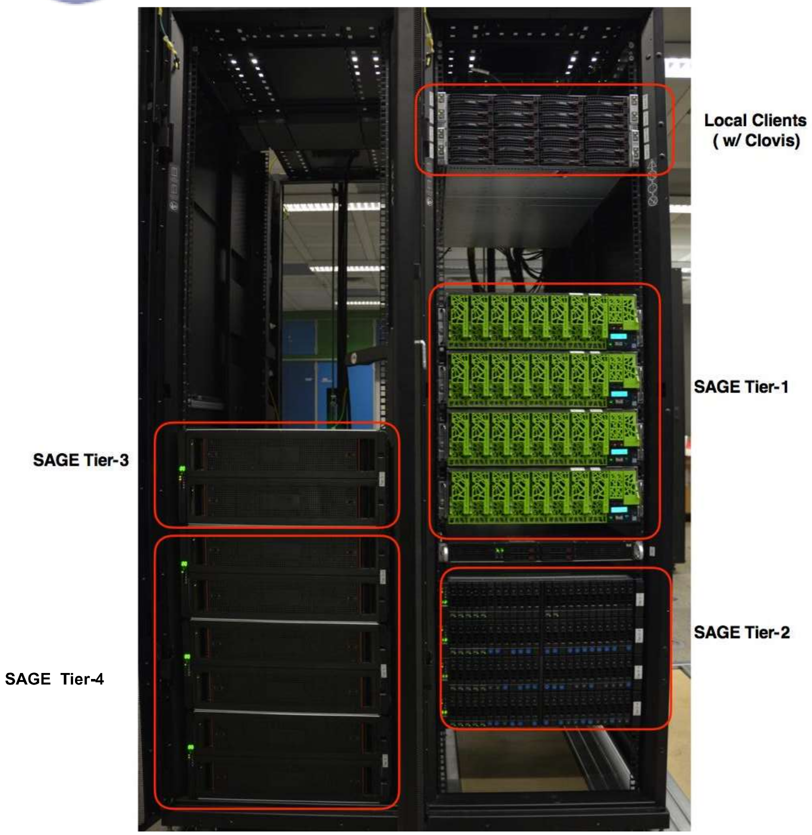
Manage and Process extremely large data sets

<u>AI/DL</u>

Large Memory Requirements

Storage and I/O Reqs significantly different

SAGE Project Recap [2015 - 2018]



- ✓ Storage system based CORTX Motr
- ✓ Co-designed with "BDEC" Use Cases (**B**ig **D**ata **E**xtreme **C**ompute)
- ✓ Assembled @ Seagate, UK
- ✓ Deployed @ Juelich Supercomputing, Germany
- ✓ Porting of Stack Components done
- ✓ Porting of BDEC applications done

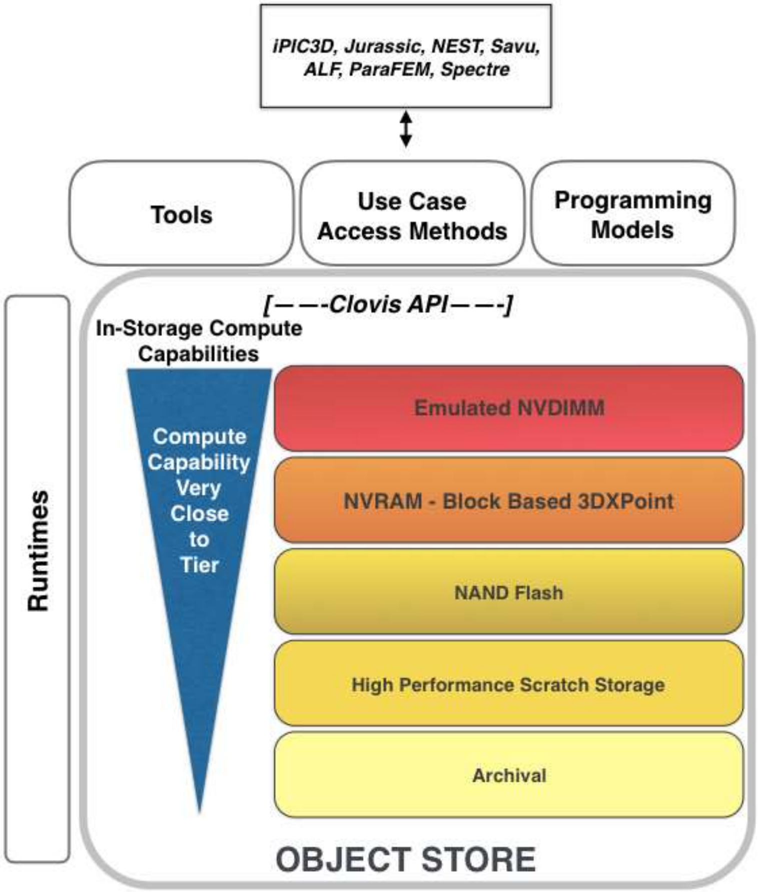
Key Takeaways from SAGE

Motr Basic Services

- Layouts
- Containers
- Porting on different media tiers
- Function shipping (PoC)
- Clovis (Motr API) usage

Runtimes

- Cache Management
- Virtual Memory Hierarchy (Both using USM)


iPIC3D, Jurassic, NEST, Savu, ALF, ParaFEM, Spectre
Tools
Use Case Access Methods
Programming Models
[——-Clovis API——-]
In-Storage Compute Capabilities
Compute Capability Very Close to Tier
Runtimes
Emulated NVDIMM
NVRAM - Block Based 3DXPoint
NAND Flash
High Performance Scratch Storage
Archival
OBJECT STORE

Use Case Access

- PNFS
- Apache Flink

Programming Models

- Exploring Avoiding MPI-IO

Tools

- Allinea Performance Tools
- HSM

Sage2 - Continuing to build on the vision

SEAGATE
UK

Bull
atos technologies
France

CCFE
CULHAM CENTRE FOR FUSION ENERGY
UK

Kitware
France

DE LA RECHERCHE À L'INDUSTRIE
cea
France

Sage2

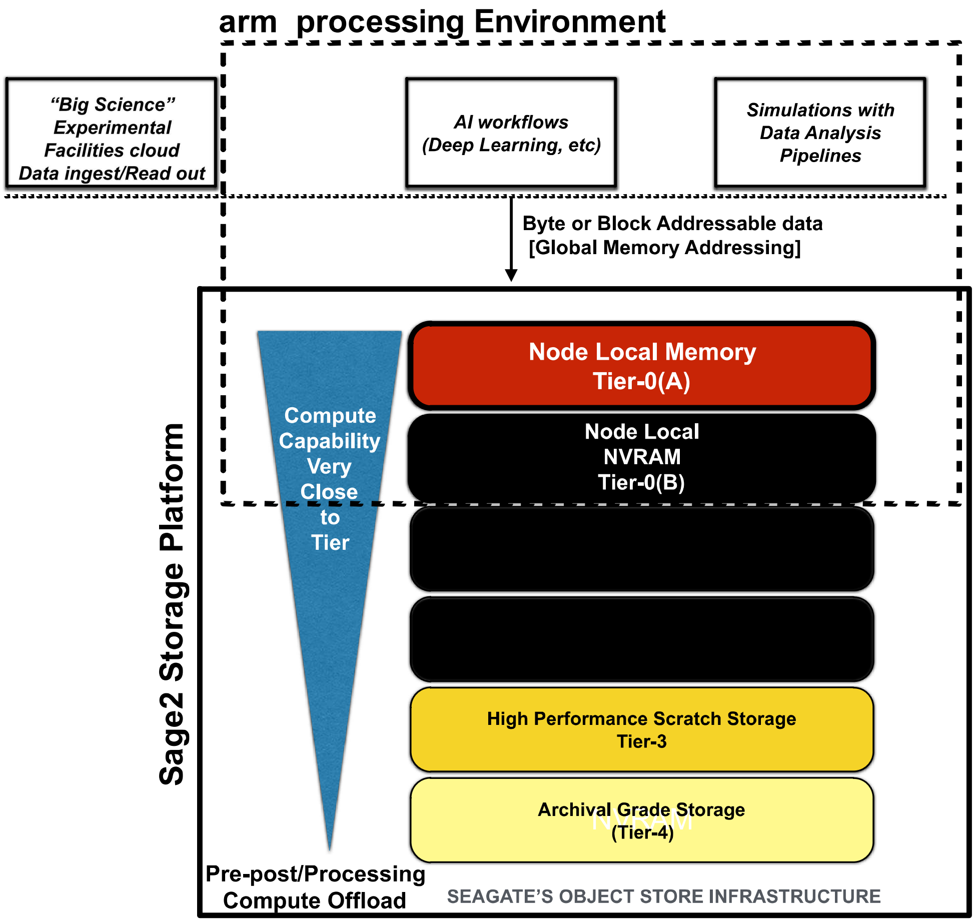
KTH
VETENSKAP OCH KONST
Sweden

JÜLICH
Forschungszentrum
Germany

epcc
UK

arm
UK

Sage2 Innovation


arm processing Environment
"Big Science" Experimental Facilities cloud Data ingest/Read out
AI workflows (Deep Learning, etc)
Simulations with Data Analysis Pipelines
Byte or Block Addressable data [Global Memory Addressing]
Sage2 Storage Platform
Compute Capability Very Close to Tier
Node Local Memory Tier-0(A)
Node Local NVRAM Tier-0(B)
High Performance Scratch Storage Tier-3
Archival Grade Storage (Tier-4)
Pre-post/Processing Compute Offload
SEAGATE'S OBJECT STORE INFRASTRUCTURE

Vision:

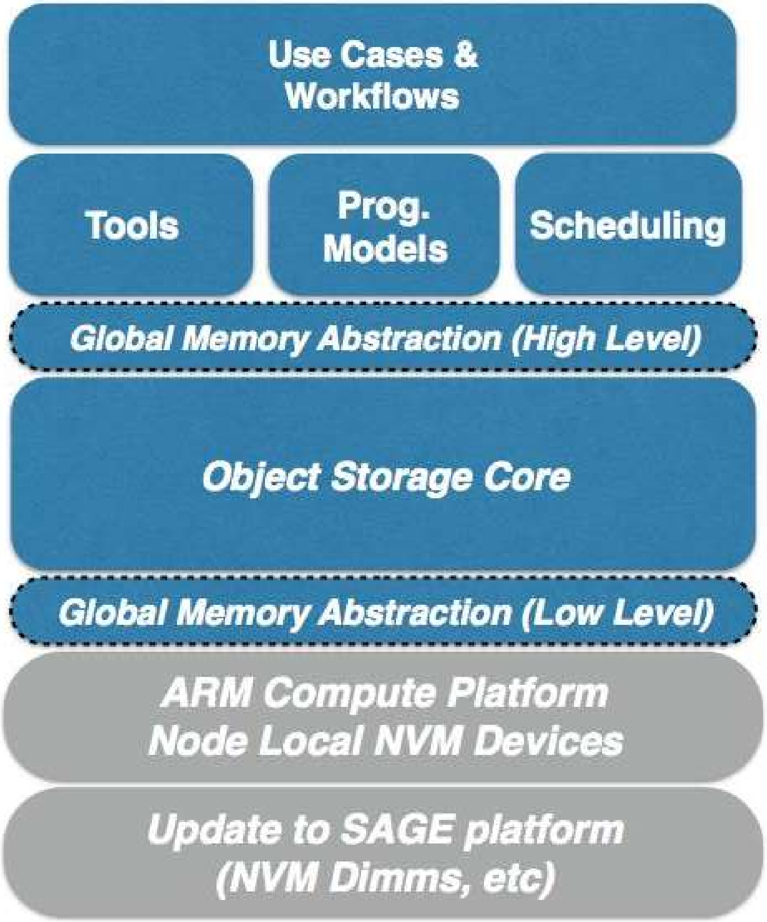
Extending storage systems into Compute nodes & blurring the lines between memory & storage

<u>Four</u> primary Innovations

1. **Compute node local Memories** part of storage stack
2. **Byte Addressable extensions** into Persistent storage (Global Memory Abstraction)
3. **Co-design** with new workflows: Mainly Data analytics pipelines w/ **AI/Deep learning**
4. **Co-design** with **ARM based environments** – moving towards European HPC Ecosystem Goals.

AI/DL use cases expected to be memory intensive & will exploit node local memory which will need to be extended

Sage2 - Key Stack Components


Use Cases & Workflows
Tools
Prog. Models
Scheduling
Global Memory Abstraction (High Level)
Object Storage Core
Global Memory Abstraction (Low Level)
ARM Compute Platform Node Local NVM Devices
Update to SAGE platform (NVM Dimms, etc)

Tools/ Prog. Models/Schedulers

- dCache, High Speed Object Transfer, I/O Containers, TensorFlow, Slurm for Motr, Object access Prog. Mod, Simple Access Interface

GMA

- High Level – API for mapping Objects in Memory
- Low Level – Incorporating NVDIMMs

Object Storage Core

- Motr for GMA
- Motr extreme scale comps. - QoS, DTM, Function Shipping
- Motr for Sage2 (Incl. ARM port)

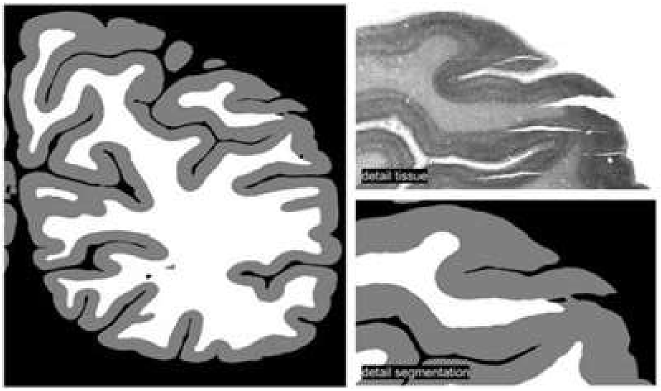
ARM

- ARM support for NVDIMMs

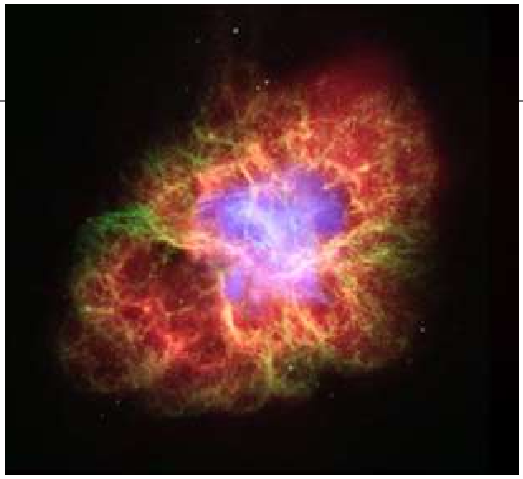
Sage2 Use Cases

AI Based Data Analysis
[1]Cervical Cancer Diagnosis

AI Based Data Analysis
[2] Multi-label Classification of Large Videos

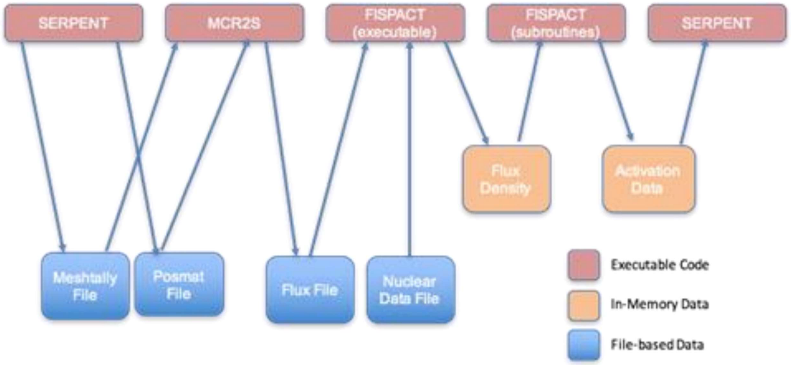


[3] Brain Image Data Analysis



[4] Radio Astronomy Data Analysis

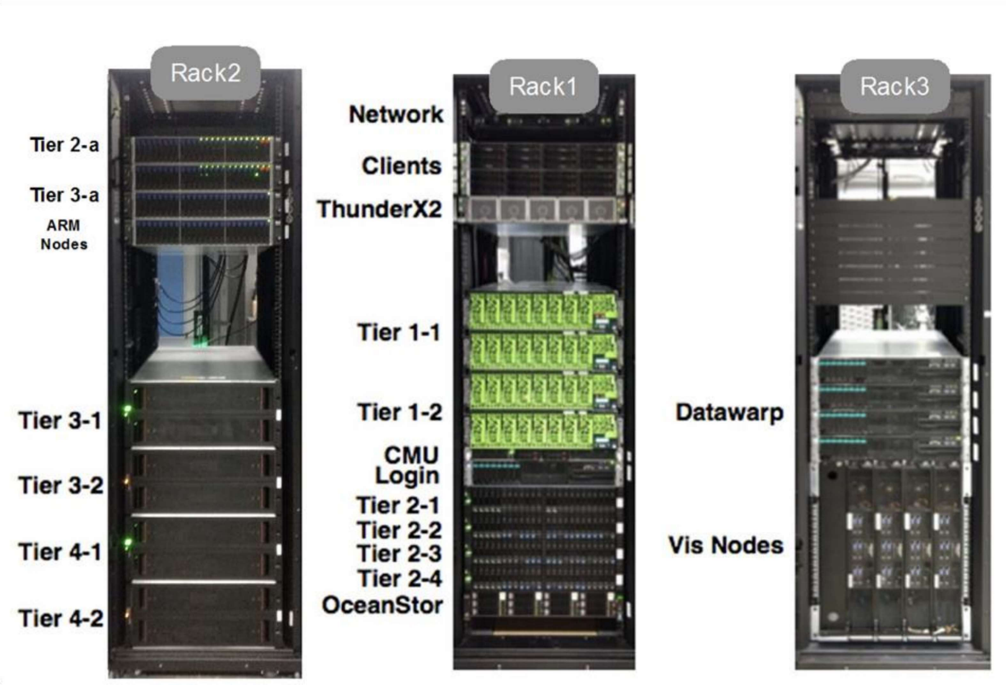
Machine Learning
[6]Tensorflow for machine learning monitoring data



[5] Multi-Physics Multi-stage workflows (Nuclear Fusion)

[7] Classic HPC Applications

Sage2 Update



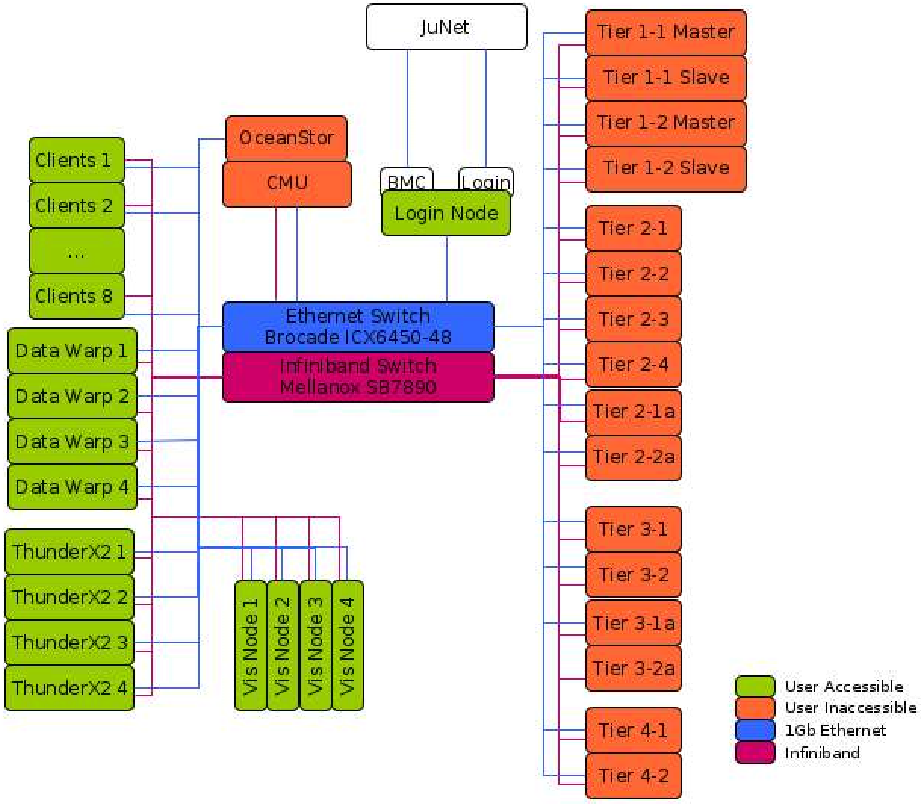
□ Prototype updated with latest Motr+Hare
□ Focus on Application Porting
□ Completion of Prototype Implementations
□ Detailed Performance analysis of CORTX on SAGE – Coming up

Sage2 – Ongoing POCs/ Implementations (In Motr, & on top of Motr API)

□ QoS (HSM & Performance Throttling) with Motr
□ CORTX Arm Porting with Motr
□ TensorFlow on Motr API
□ dCache on Motr API
□ 3DXPoint NVDIMM Interoperability
□ Deployed AI applications on Motr
□ Slurm CORTX Burst Buffer Plugin on Motr API
□ Global Memory Abstraction APIs & Motr Driver on Motr API
□ Function Shipping in Motr
□ Simple Access Interface on Motr API
□ Distributed Transactions in Objects (Motr)
□ Clovis Apps Framework on Motr API
□ Go binding on Motr MPI

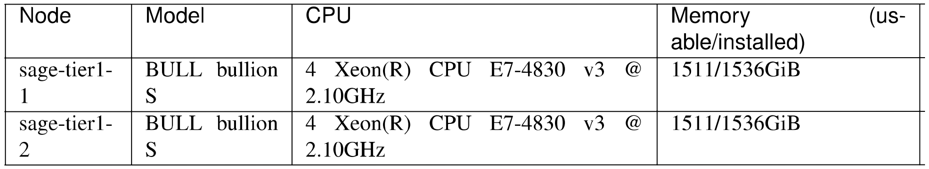
Open-Source Code (Incl. Documentation) (Q3, Q4 2021)

More on SAGE prototype

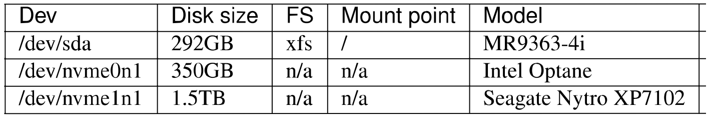


	Rack 3	Rack 1	Rack 2
42		Mellanox SB7890 Infiniband Swtich	
41		Mellanox SX6036 Infiniband Swtich	
40		Brocade ICX6430-24 Ethernet Switch	
39		Brocade ICX6450-48 Ethernet Switch	
38		**Clients**	**Tier2-1a**
37		Supermicro 2U 4-Server	ARM server
36	**Visualisation Nodes**	**Clients**	**Tier2-2a**
35	visnode-01	Supermicro 2U 4-Server	ARM server
34	visnode-02	**ThunderX2 Nodes**	**Tier3-1a**
33	visnode-03	client-tx2-[1-4]	ARM server
32	visnode-04		**Tier3-2a**
31			ARM server
30			
29			
28			
27		**Tier1-1 Master**	
26		Bull Bullion Server	
25			
24		**Tier1-1 Slave**	
23		Bull Bullion Server	
22	**Data Warp Nodes**		
21	datawarp-01	**Tier1-2 Master**	
20	datawarp-02	Bull Bullion Server	**Tier3-1**
19	datawarp-03		Seagate 5U84 Enclosure
18	datawarp-04	**Tier1-2 Slave**	
17		Bull Bullion Server	
16			
15		**CMU** Bull R421-E4 Server	**Tier3-2**
14		**Login** Cray S2600WTTR Server	Seagate 5U84 Enclosure
13			
12		**Tier2-1**	
11		Seagate 2U24 Enclosure	
10		**Tier2-2**	**Tier4-1**
9		Seagate 2U24 Enclosure	Seagate 5U84 Enclosure
8		**Tier2-3**	
7		Seagate 2U24 Enclosure	
6		**Tier2-4**	
5		Seagate 2U24 Enclosure	**Tier4-2**
4		**Scratch Storage** OceanStor	Seagate 5U84 Enclosure
3			
2			
1			

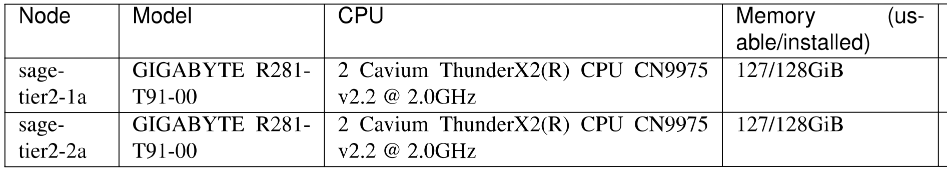
SAGE – Tiers 1 and 2



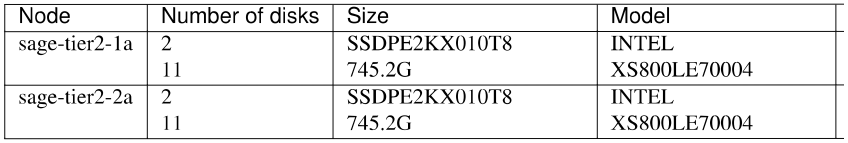
Node	Model	CPU	Memory (usable/installed)
sage-tier1-1	BULL bullion S	4 Xeon(R) CPU E7-4830 v3 @ 2.10GHz	1511/1536GiB
sage-tier1-2	BULL bullion S	4 Xeon(R) CPU E7-4830 v3 @ 2.10GHz	1511/1536GiB



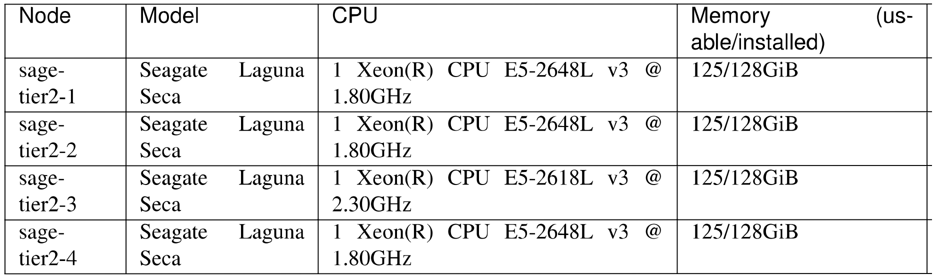
Dev	Disk size	FS	Mount point	Model
/dev/sda	292GB	xfs	/	MR9363-4i
/dev/nvme0n1	350GB	n/a	n/a	Intel Optane
/dev/nvme1n1	1.5TB	n/a	n/a	Seagate Nytro XP7102



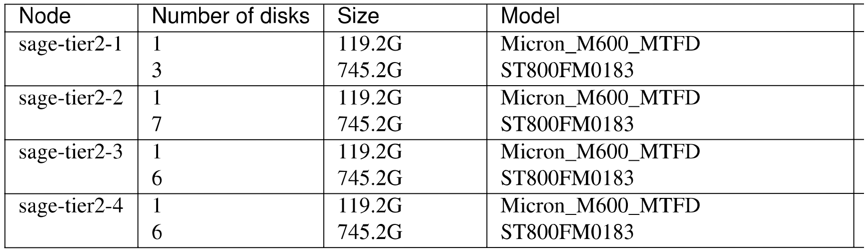
Node	Model	CPU	Memory (usable/installed)
sage-tier2-1a	GIGABYTE R281-T91-00	2 Cavium ThunderX2(R) CPU CN9975 v2.2 @ 2.0GHz	127/128GiB
sage-tier2-2a	GIGABYTE R281-T91-00	2 Cavium ThunderX2(R) CPU CN9975 v2.2 @ 2.0GHz	127/128GiB



Node	Number of disks	Size	Model
sage-tier2-1a	2 11	SSDPE2KX010T8 745.2G	INTEL XS800LE70004
sage-tier2-2a	2 11	SSDPE2KX010T8 745.2G	INTEL XS800LE70004

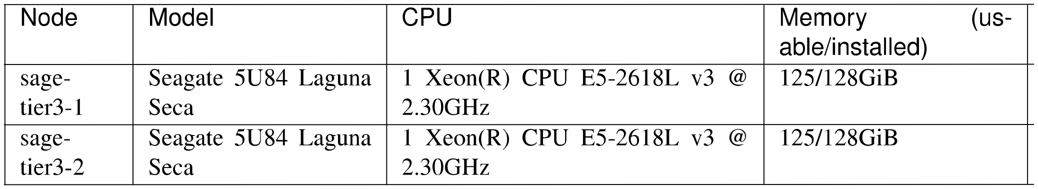


Node	Model	CPU	Memory (usable/installed)
sage-tier2-1	Seagate Laguna Seca	1 Xeon(R) CPU E5-2648L v3 @ 1.80GHz	125/128GiB
sage-tier2-2	Seagate Laguna Seca	1 Xeon(R) CPU E5-2648L v3 @ 1.80GHz	125/128GiB
sage-tier2-3	Seagate Laguna Seca	1 Xeon(R) CPU E5-2618L v3 @ 2.30GHz	125/128GiB
sage-tier2-4	Seagate Laguna Seca	1 Xeon(R) CPU E5-2648L v3 @ 1.80GHz	125/128GiB

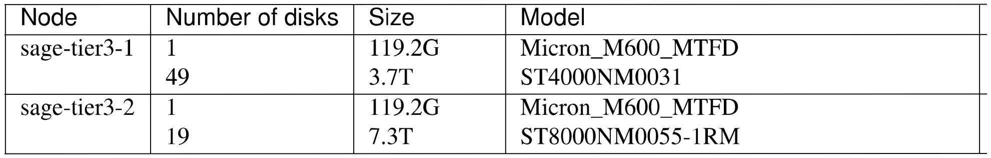


Node	Number of disks	Size	Model
sage-tier2-1	1 3	119.2G 745.2G	Micron_M600_MTFD ST800FM0183
sage-tier2-2	1 7	119.2G 745.2G	Micron_M600_MTFD ST800FM0183
sage-tier2-3	1 6	119.2G 745.2G	Micron_M600_MTFD ST800FM0183
sage-tier2-4	1 6	119.2G 745.2G	Micron_M600_MTFD ST800FM0183

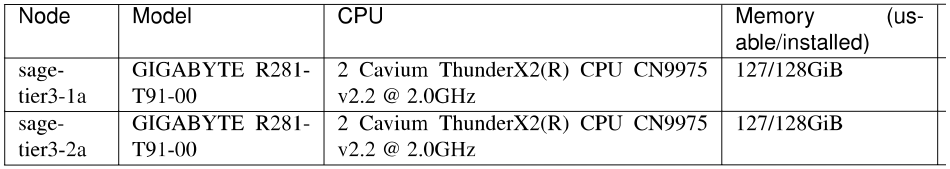
SAGE – Tiers 3 and 4



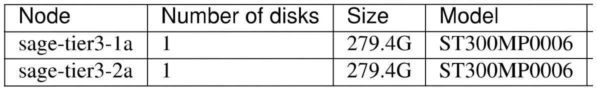
Node	Model	CPU	Memory (usable/installed)
sage-tier3-1	Seagate 5U84 Laguna Seca	1 Xeon(R) CPU E5-2618L v3 @ 2.30GHz	125/128GiB
sage-tier3-2	Seagate 5U84 Laguna Seca	1 Xeon(R) CPU E5-2618L v3 @ 2.30GHz	125/128GiB



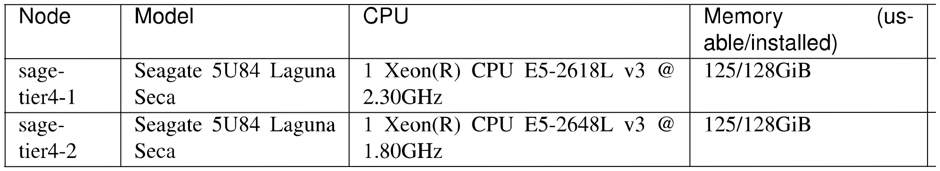
Node	Number of disks	Size	Model
sage-tier3-1	1 49	119.2G 3.7T	Micron_M600_MTFD ST4000NM0031
sage-tier3-2	1 19	119.2G 7.3T	Micron_M600_MTFD ST8000NM0055-1RM



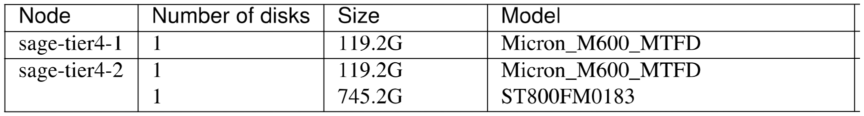
Node	Model	CPU	Memory (usable/installed)
sage-tier3-1a	GIGABYTE R281-T91-00	2 Cavium ThunderX2(R) CPU CN9975 v2.2 @ 2.0GHz	127/128GiB
sage-tier3-2a	GIGABYTE R281-T91-00	2 Cavium ThunderX2(R) CPU CN9975 v2.2 @ 2.0GHz	127/128GiB



Node	Number of disks	Size	Model
sage-tier3-1a	1	279.4G	ST300MP0006
sage-tier3-2a	1	279.4G	ST300MP0006

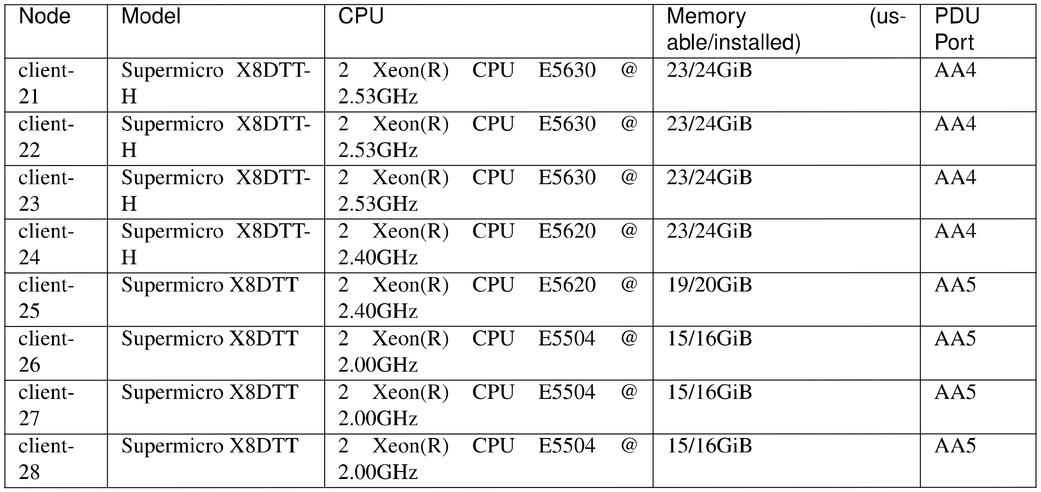


Node	Model	CPU	Memory (usable/installed)
sage-tier4-1	Seagate 5U84 Laguna Seca	1 Xeon(R) CPU E5-2618L v3 @ 2.30GHz	125/128GiB
sage-tier4-2	Seagate 5U84 Laguna Seca	1 Xeon(R) CPU E5-2648L v3 @ 1.80GHz	125/128GiB

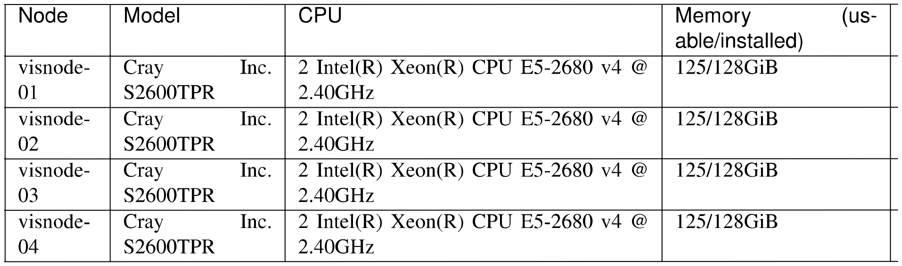


Node	Number of disks	Size	Model
sage-tier4-1	1	119.2G	Micron_M600_MTFD
sage-tier4-2	1 1	119.2G 745.2G	Micron_M600_MTFD ST800FM0183

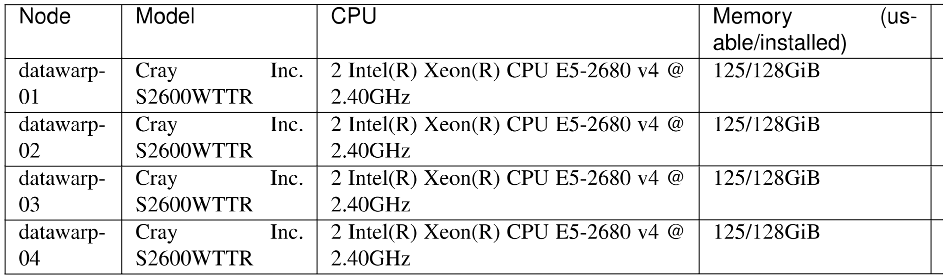
SAGE – The 16 Clients



Node	Model	CPU	Memory (usable/installed)	PDU Port
client-21	Supermicro X8DTT-H	2 Xeon(R) CPU E5630 @ 2.53GHz	23/24GiB	AA4
client-22	Supermicro X8DTT-H	2 Xeon(R) CPU E5630 @ 2.53GHz	23/24GiB	AA4
client-23	Supermicro X8DTT-H	2 Xeon(R) CPU E5630 @ 2.53GHz	23/24GiB	AA4
client-24	Supermicro X8DTT-H	2 Xeon(R) CPU E5620 @ 2.40GHz	23/24GiB	AA4
client-25	Supermicro X8DTT	2 Xeon(R) CPU E5620 @ 2.40GHz	19/20GiB	AA5
client-26	Supermicro X8DTT	2 Xeon(R) CPU E5504 @ 2.00GHz	15/16GiB	AA5
client-27	Supermicro X8DTT	2 Xeon(R) CPU E5504 @ 2.00GHz	15/16GiB	AA5
client-28	Supermicro X8DTT	2 Xeon(R) CPU E5504 @ 2.00GHz	15/16GiB	AA5

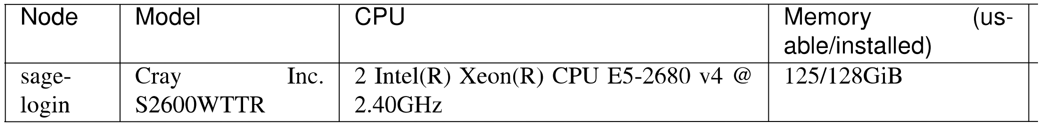


Node	Model	CPU	Memory (usable/installed)
visnode-01	Cray Inc. S2600TPR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB
visnode-02	Cray Inc. S2600TPR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB
visnode-03	Cray Inc. S2600TPR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB
visnode-04	Cray Inc. S2600TPR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB

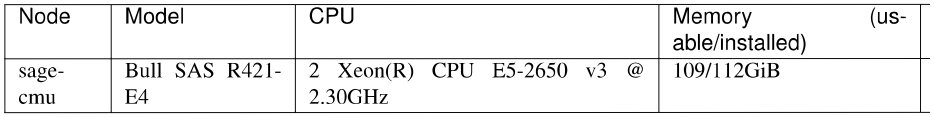


Node	Model	CPU	Memory (usable/installed)
datawarp-01	Cray Inc. S2600WTTR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB
datawarp-02	Cray Inc. S2600WTTR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB
datawarp-03	Cray Inc. S2600WTTR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB
datawarp-04	Cray Inc. S2600WTTR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB

SAGE – Login Node and CMU/ Software



Node	Model	CPU	Memory (usable/installed)
sage-login	Cray Inc. S2600WTTR	2 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz	125/128GiB



Node	Model	CPU	Memory (usable/installed)
sage-cmu	Bull SAS R421-E4	2 Xeon(R) CPU E5-2650 v3 @ 2.30GHz	109/112GiB

server nodes
CentOS Linux release 7.9.2009 (Core)
cortx-motr-1.0.0-1_git89f7737_3.10.0_1127.19.1.el7.x86_64
cortx-hare-1.0.0-1_git28f3372.el7.x86_64
kmod-lustre-client-2.12.4.2_171_g9356888-1.el7.x86_64

compute nodes
CentOS Linux release 7.8.2003 (Core)
cortx-motr-1.0.0-1_git89f7737_3.10.0_1127.19.1.el7.x86_64
cortx-hare-1.0.0-1_git28f3372.el7.x86_64
kmod-lustre-client-2.12.4.2_171_g9356888-1.el7.x86_64

Usage of the SAGE System with Clovis Apps (Demo)

- **`c0ct`**
  `Read motr object to a file`
- **`c0cp`**
  `Write motr object from a file`
- **`c0rm`**
  `Remove motr object`

- All three applications run natively on Motr clients.
- They use the Motr client interface (Clovis) to connect directly to servers for performing object I/O.
- All IO and other operations performed on native/raw motr objects.
- Do not handle composite objects yet.
- Not at all S3 and other high-level objects.

Git Repo:
https://gitlab.version.fz-juelich.de/sage2/clovis-sample-apps
(Ongoing work to consolidate repository)

HSM Demo

HSM Summary

```
m0hsm> help
Usage: m0hsm <action> <fid> [...]
  actions:
    create <fid> <tier>
    show <fid>
    dump <fid>
    write <fid> <offset> <len> <seed>
    write_file <fid> <path>
    read <fid> <offset> <len>
    copy <fid> <offset> <len> <src_tier> <tgt_tier> [options: mv,keep_prev,w2dest]
    move <fid> <offset> <len> <src_tier> <tgt_tier> [options: keep_prev,w2dest]
    stage <fid> <offset> <len> <tgt_tier> [options: mv,w2dest]
    archive <fid> <offset> <len> <tgt_tier> [options: mv,keep_prev,w2dest]
    release <fid> <offset> <len> <tier> [options: keep_latest]
    multi_release <fid> <offset> <len> <max_tier> [options: keep_latest]
    set_write_tier <fid> <tier>


  <fid> parameter format is [hi:]lo. (hi == 0 if not specified.)
  The numbers are read in decimal, hexadecimal (when prefixed with `0x')
  or octal (when prefixed with `0') formats.
m0hsm>
```

Note "first cut" performance for tiers as follows:

Tier1 – 2.6 GB/s (4 NVME devs)
Tier2 – 1.9 GB/s (4 SSD devs)
Tier3 – 0.6 GB/s (4 HDD devs)

(Note: the pool width of 4 devices was used in Tier2 and Tier3 (as in Tier1) to make the perf measurements comparable.

Git Repo

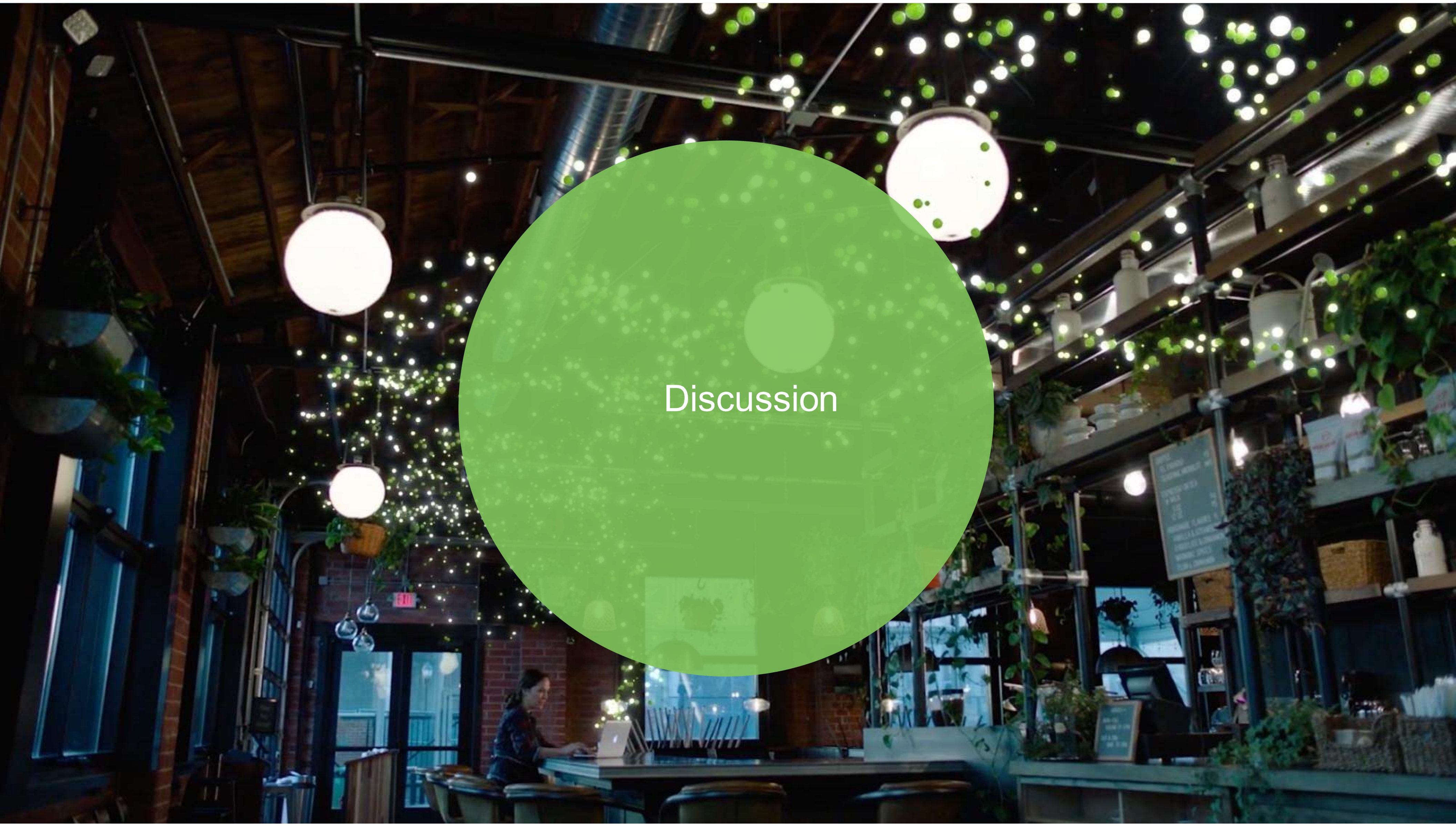
https://github.com/Seagate/cortx-motr
https://github.com/Seagate/cortx-motr/tree/main/hsm

Additional Notes (Code & software management)

- Performance tests currently being run by mcp utility (written in Go) (We are getting multiple GB/s across tiers – more detailed performance characterizations TBD)

- Code that will be available (Many will be integrated/linked from CORTX github)

  - MIO in Maestro (Seagate) - currently in Maestro gitlab repos
    - https://github.com/Seagate/cortx-mio
  - TensorFlow
  - DCache
  - Slurm Interface
  - Clovis Driver for GMA
  - Simple Access Interface
  - ESDM Middleware work in EsiWACE2 (Seagate) - currently in DKRZ gitlab repos



Discussion