HPC CLUSTER AND DATA ARCHIVING CENTER AT OATS:

ARCHITECTURE AND PROVISIONING

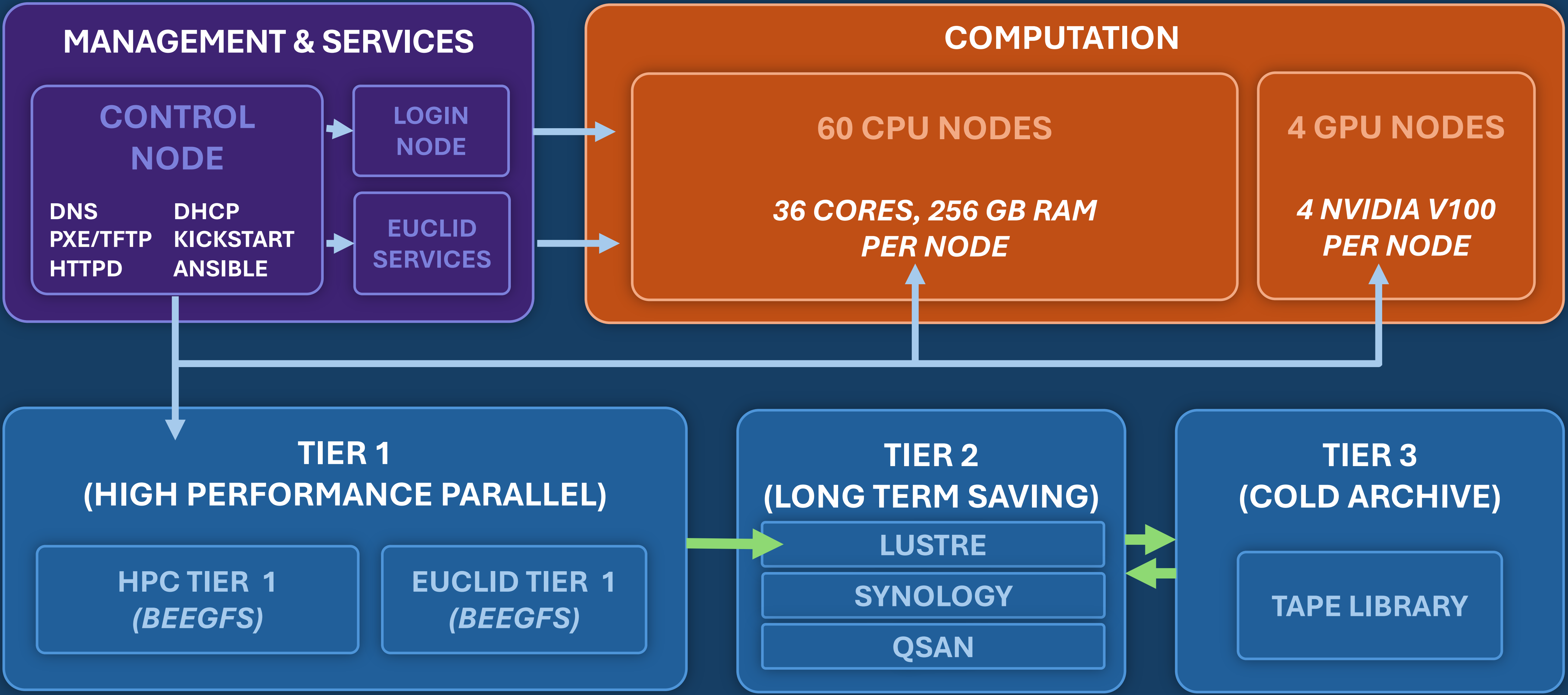
Presented by

Gianmarco Maggio

On behalf of the OATS Computing and Data Center Team:

Massimo Costantini, Marco Frailis, Federico Gasparo, Cristina Knapic, Massimo Sponza, Fabio Stocco, Giuliano Taffoni, Daniele Tavagnacco, Cristiano Urban, Claudio Vuerli

HPC CLUSTER ARCHITECTURE WITH MULTI-TIER STORAGE SYSTEM



MANAGEMENT & SERVICES

CONTROL NODE

DNS
DHCP
PXE/TFTP
KICKSTART
HTTPD
ANSIBLE

LOGIN NODE

EUCLID SERVICES

COMPUTATION

60 CPU NODES

36 CORES, 256 GB RAM PER NODE

4 GPU NODES

4 NVIDIA V100 PER NODE

TIER 1 (HIGH PERFORMANCE PARALLEL)

HPC TIER 1 *(BEEGFS)*

EUCLID TIER 1 *(BEEGFS)*

TIER 2 (LONG TERM SAVING)

LUSTRE

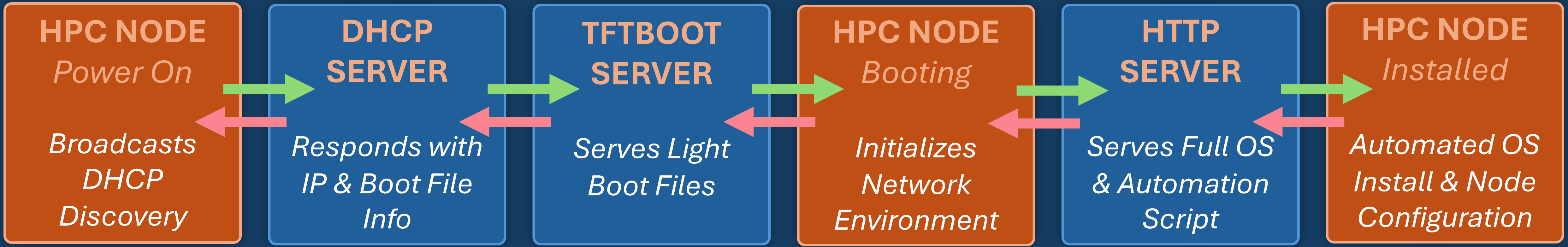
SYNOLOGY

QSAN

TIER 3 (COLD ARCHIVE)

TAPE LIBRARY

HPC AUTOMATED NETWORK INSTALLATION


HPC NODE
Power On
Broadcasts DHCP Discovery
DHCP SERVER
Responds with IP & Boot File Info
TFTBOOT SERVER
Serves Light Boot Files
HPC NODE
Booting
Initializes Network Environment
HTTP SERVER
Serves Full OS & Automation Script
HPC NODE
Installed
Automated OS Install & Node Configuration

CONTROL NODE(S) SERVICES

DHCP DNS PXE/TFTP HTTPD KICKSTART ANSIBLE

	KICKSTART **(RedHat)**	**PRESEED** **(Debian)**
CONFIGURATION STRUCTURE	Structured sections	Flat Key-Value: Linear List of Answers
READABILITY & USABILITY	High	Moderate to Low

HPC AUTOMATION & ORCHESTRATION

PROVISIONING & CONFIGURATION

- Idempotency: same configuration across all nodes
- Installation of drivers, libraries (MPI, CUDA)

LIFECYCLE MANAGEMENT & UPDATES

- Simultaneous security patches
- Rolling updates management without job interruptions

SCALABILITY & REPRODUCIBILITY

- Easy addition of new nodes
- Infrastructure as Code (IaC) definition

ANSIBLE	SALTSTACK
AGENTLESS (SSH)	AGENT-BASED (Minions)
PUSH	PULL (default) / PUSH
YAML / JINJA	YAML / JINJA / PYTHON
SSH	zeroMQ
Initial setup and <100 nodes: Ansible save ~20-30% time	With 1000+ nodes: Saltstack save ~50-80% execution time

Key advantages:

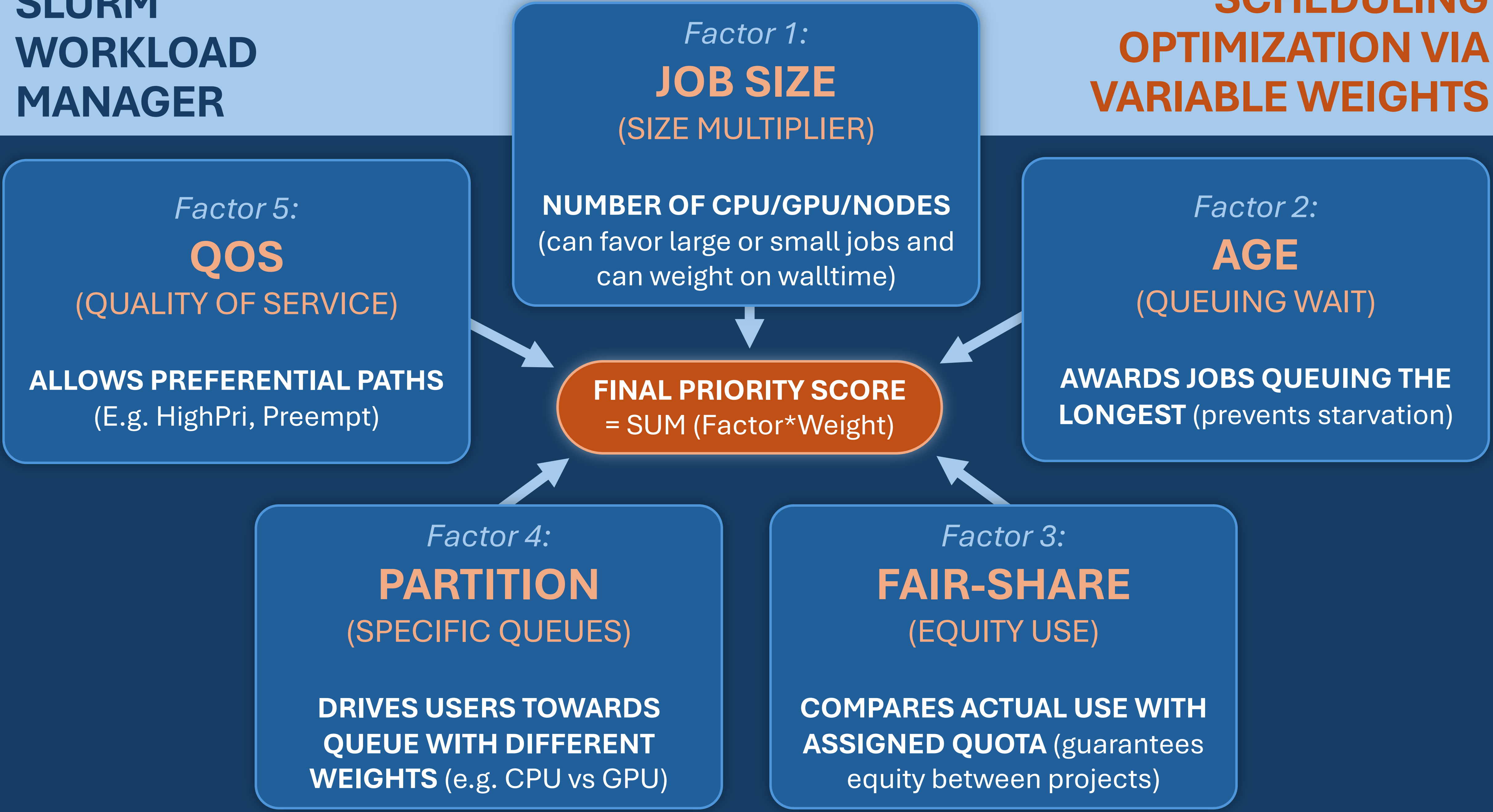
- *Reduction of human errors*
- *Time savings for SysAdmins*
- *Consistent environments for Researchers*
- *Rapid resource deployment*

HPC SOFTWARE DEPLOYMENT

SPACK **(modules)**		**APPTAINER** **(containers)**
Compile from source on target hardware	*SOFTWARE CREATION METHOD*	Encapsulated pre-built environment
Maximum	*PERFORMANCE LEVEL*	Excellent
Low portability	*PORTABILITY & REPRODUCIBILITY*	High portability
Complex	*EASY OF USE & MANAGEMENT*	Simple
High	*STORAGE EFFICIENCY*	Moderate (bind mounts overhead)
HPC requiring locally optimizated binaries Standardized user software libraries	*OPTIMAL USE CASES*	Highly reproducible scientific publications and paper data Workflow conflicting with cluster OS

SLURM WORKLOAD MANAGER

SCHEDULING OPTIMIZATION VIA VARIABLE WEIGHTS



Factor 1:

JOB SIZE

(SIZE MULTIPLIER)

NUMBER OF CPU/GPU/NODES (can favor large or small jobs and can weight on walltime)

Factor 2:

AGE

(QUEUING WAIT)

AWARDS JOBS QUEUING THE LONGEST (prevents starvation)

Factor 3:

FAIR-SHARE

(EQUITY USE)

COMPARES ACTUAL USE WITH ASSIGNED QUOTA (guarantees equity between projects)

Factor 4:

PARTITION

(SPECIFIC QUEUES)

DRIVES USERS TOWARDS QUEUE WITH DIFFERENT WEIGHTS (e.g. CPU vs GPU)

Factor 5:

QOS

(QUALITY OF SERVICE)

ALLOWS PREFERENTIAL PATHS (E.g. HighPri, Preempt)

FINAL PRIORITY SCORE
= SUM (Factor*Weight)

TIER 1 BEEGFS STORAGES

BEEGFS	
ARCHITECTURAL PHYLOSOPHY	USER FRIENDLY & AGILITY
COST MODEL	LOW (open-source)
INSTALLATION DIFFICULTY	EASY
MAINTENANCE & MANAGEMENT	SIMPLE
PERFORMANCE FOCUS	SMALL/MEDIUM CLUSTERS & MIXED WORKLOADS

Max Seq Write: 7.5 GB/s
Max Seq Read: 10.0 GB/s

HPC/PLEIADI BEEGFS STORAGE

4 BeeGFS STORAGE SERVERS

EACH SERVER: 2x12 DISKS, RAID6

~600 TB TOTAL USABLE SPACE

Metadata Buddy Mirroring

DATA PATH: Omnipath 100 Gb/s

EUCLID BEEGFS STORAGE

2 BeeGFS STORAGE SERVERS

EACH SERVER: 3x10 DISKS, RAID6

~720 TB TOTAL USABLE SPACE

METADATA ON RAID10 SSDs

Metadata Buddy Mirroring

Max Seq Write: 4.5 GB/s
Max Seq Read: 6.5 GB/s

IA2 TIER 2 FACILITIES

LUSTRE

(PARALLEL FILESYSTEM)

USED AS TIER 2 STORAGE

2 JBOD 90x HDD (1.8 PB total)
2 SERVERS (controlling
½ JBOD 1 + ½ JBOD 2)
1 METADATA STORAGE
(SSDs, RAID 10, 4 TB total)

- Large Storage Focus
- Metadata Redundancy
- Distributed Namespace
- Software Defined (ZFS) Data Integrity

SYNOLOGY

(NAS)

USED FOR OWNCLOUD
AND
VMWARE VM BACKUPS

270 TB OF SPACE ON
HDDs

Easy Of Use.
Mixed Data Hosting.
Versatile Services.

QSAN

(UNIFIED STORAGE)

TRANSITION STORAGE TO
LONG TERM PRESERVATION
FOR ARI-L/ALMA AND PRISMA
PROJECTS

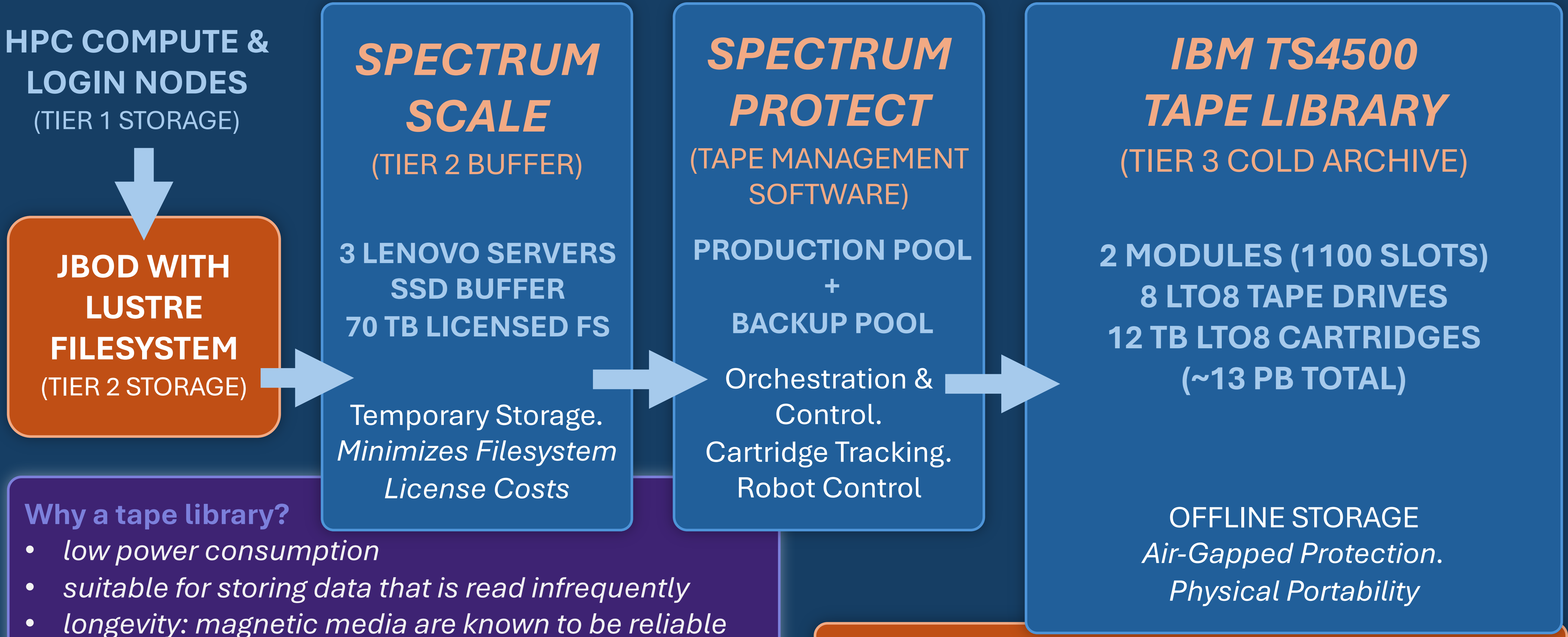
400 TB OF SPACE ON
HDDs

High Availability.
Data Integrity for Transition.
Secure Bridging.

NETWORK: 10 Gb/s Ethernet
DATA PATH: Fibre Channel, Serial Attached SCSI (SAS)

CROSS BACKUP: Synology + Lustre => QSAN and QSAN => Synology

IA2 TIER 3 COLD STORAGE ARCHITECTURE



HPC COMPUTE & LOGIN NODES
(TIER 1 STORAGE)

JBOD WITH LUSTRE FILESYSTEM
(TIER 2 STORAGE)

SPECTRUM SCALE
(TIER 2 BUFFER)

3 LENOVO SERVERS SSD BUFFER
70 TB LICENSED FS

Temporary Storage.
Minimizes Filesystem License Costs

SPECTRUM PROTECT
(TAPE MANAGEMENT SOFTWARE)

PRODUCTION POOL
+
BACKUP POOL

Orchestration & Control.
Cartridge Tracking.
Robot Control

IBM TS4500 TAPE LIBRARY
(TIER 3 COLD ARCHIVE)

2 MODULES (1100 SLOTS)
8 LTO8 TAPE DRIVES
12 TB LTO8 CARTRIDGES
(~13 PB TOTAL)

OFFLINE STORAGE
Air-Gapped Protection.
Physical Portability

Why a tape library?

- *low power consumption*
- *suitable for storing data that is read infrequently*
- *longevity: magnetic media are known to be reliable for long-term data preservation (10+ years)*

ACTUALLY USED: ~1.8 PB (including redundancy)

IA2 SERVICES

VMWARE VSPHERE VIRTUALIZATION INFRASTRUCTURE

3 LENOVO SERVERS
(WITH 32 CPUS, 1 TB RAM)
STORAGE FOR VMS
(150 TB ALL FLASH)
100+ VMS IN PRODUCTION

Migration to Proxmox VE during the next years

USC-C/INAF Services hosted on IA2 infrastructure:

- **ownCloud:** ~ 2.6k users
- **Easy Redmine:** ~ 350 users
- **GitLab:** ~ 1k users / ~ 2k projects
- **Indico:** ~ 6k users
- **DOI** service: https://doi.ict.inaf.it/
- **INAF Open Access Repository**
- **Website hosting** (e.g. USC-C website and several other projects)

DATABASE machines with active-passive replica

Physical machines used for applications where performance is critical:
2 DELL servers with 2 CPUs (2 x 32 core), 256 GB RAM, 3.5 TB all flash

IA2 ISTITUTIONAL DUTIES

LIVE ARCHIVES (OBSERVATORIES, SIMULATIONS, CATALOGUES)

- Data ingestion from astronomical instruments VM DB LUSTRE
- Web portals to allow data retrieval VM DB LUSTRE
- Preservation of older than one year data in cold storage LUSTRE

STORAGE SERVICES

- Online storage LUSTRE
- Long-term preservation TAPE
- Cloud storage: ownCloud SYNOLOGY

SUPPORT SERVICES

- Collaborative tools (Indico, GitLab, ownCloud, Easy Redmine,...)
- Web Hosting

HPC CLUSTER AND DATA ARCHIVING CENTER AT OATS:

ARCHITECTURE AND PROVISIONING

Presented by

Gianmarco Maggio

On behalf of the OATS Computing and Data Center Team:

Massimo Costantini, Marco Frailis, Federico Gasparo, Cristina Knapic, Massimo Sponza, Fabio Stocco, Giuliano Taffoni, Daniele Tavagnacco, Cristiano Urban, Claudio Vuerli