ARCITECTA + SPECTRA

RMACC

Does HPC Really Need Data Management?

ARCITECTA | SPECTRA

RMACC

Introductions

Normally we give this talk as a panel, much prefer a discussion that us talking!

The speakers

Different views of HPC

What do you consider Data Management

What is the value of your data

How well do you know your data?

What about data Governance and Access

Hybrid Workflows in HPC

ARCITECTA + SPECTRA

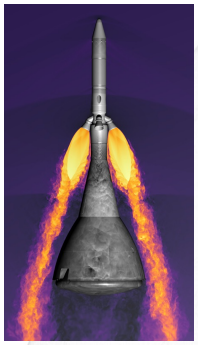
RMACC

When thinking about HPC…

ARCITECTA + SPECTRA

RMACC

Sci-Vis



Schedulers (SLURM, PBS..)
Clusters (Beowulf, Hybrid), Cloud
IO channels, fabric, interconnect

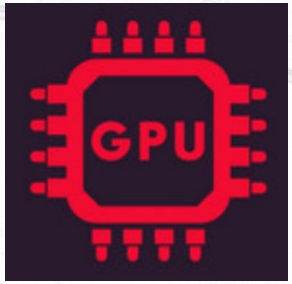
Scratch Storage

NAS: FLASH – HDD

HOT DATA

COLD DATA

Cloud/ Hybrid

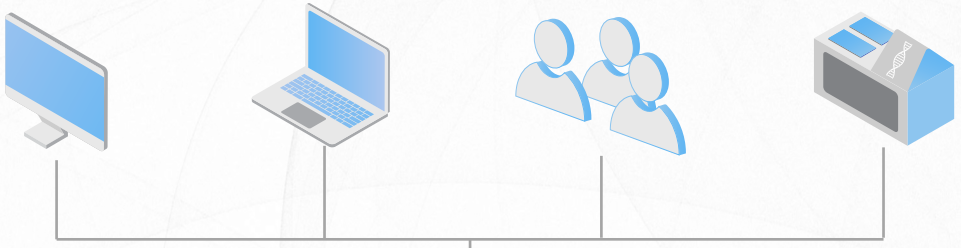


Images courtesy of NASA

ARCITECTA + SPECTRA

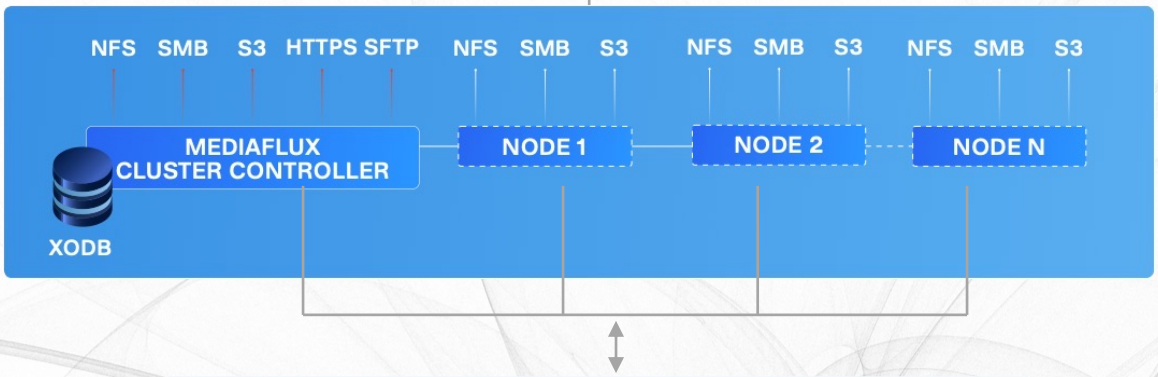
RMACC

Different Views of HPC



Metadata Enabled Data Discovery
Provides NFS and SMB access for
researchers and users

HPC CLUSTER AND
“SCRATCH” STORAGE



VIRTUALIZED STORAGE –
SINGLE GLOBAL NAMESPACE

ARCITECTA + SPECTRA

RMACC

What do you consider data management?

Effective data and storage management accelerates outcomes, while preserving data for future use and examination. Finding the right data and having it in the right location drives HPC workload performance – keeping scratch space cleared and active data hot.

Does your data management plan…

- Address immediate computational needs?
- Meet long-range archival requirements for access, management and preservation?
- Include a data lifecycle plan?
- Consider deletion?
- Utilize FAIR data principles?

ARCITECTA | SPECTRA

RMACC

What is the value of your data?

Is there such a thing as ROD (Return on Data)?

What is the cost of replacing your data? Is there a data classification system in place? Is all data treated the same, or are there different policies based on different projects or data types?

Identifying a storage cost problem…

- Is the cost too high to keep it in an archive?
- How do you discover unknown/abandoned data?
- Can AI/ML help with abandoned data?
- Who makes decisions on that data when the owner leaves?

ARCITECTA | SPECTRA

RMACC

How well do you know your data?

If you need to recompute something, do you know where to go to find that data? What algorithms were used to calculate? Was the code saved? What were the inputs? Were there environmental variables?

Identifying a storage cost problem…

- Whose data it is?
- File count?
- Maximum file size?
- Longest path name?
- What data can be deleted?
- If data is valid/uncorrupted?

ARCITECTA | SPECTRA

RMACC

What about data governance and access?

How do you ensure your data is accessible in 10 years time?

What is fast access? Time to data?

What is the role of keyword tagging and user defined metadata?

Where do Data Rights Management and protection regulation like the GDPR fit in?

Identifying a storage cost problem…

- Is your data protected from internal bad actors?
- What measures are in place to prevent cybercrime?
- Do you know what to do during and after a ransomware attack?

ARCITECTA | SPECTRA

RMACC

Hybrid workflows in HPC

On-prem data with cloud compute or cloud storage with on-prem compute?

Pipeline workflow in cloud?

Data sharing across WAN

What does the future of
HPC data management look like?

ARCITECTA + SPECTRA

RMACC

Q&A