Data Management from IRIS to MUSE

**Randolf Klein, Neal Hurlburt, Bart De Pontieu,
Ryan Timmons**

LOCKHEED MARTIN

Data Analysis & Software in Heliophysics (DASH), ESAC, Oct. 2024

Introduction

Past and current missions use local data processing, storage, and distribution methods.

In our case, the Interface Region Imaging Spectrograph (IRIS) mission delivers 20GB per day of uncompressed Level 2 data relying on local data processing and storage at Stanford's Joint Science Operations Center (JSOC) and at the Lockheed Martin Solar and Astrophysics Laboratory (LMSAL).

In addition, a cloud provider is already used for data distribution to the community to guarantee high data download rates - Amazon Web Services (AWS) S3 and Cloudfront services.

High data volumes of new missions and the availability of cloud storage and processing allow new paradigms for data processing, storage, and distribution.

For the Multi-Slit Spectral Explorer (MUSE, having a more than 40x higher data rate than IRIS, we are evaluating whether to continue the traditional data management methods with existing hardware and software or switching - partially or wholly - to cloud storage and processing.

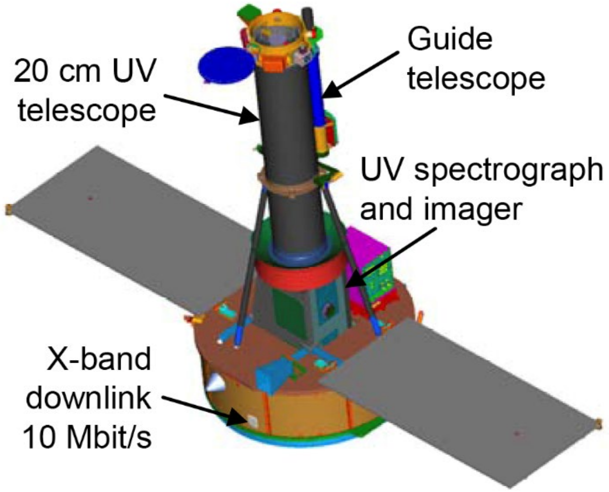
IRIS vs MUSE

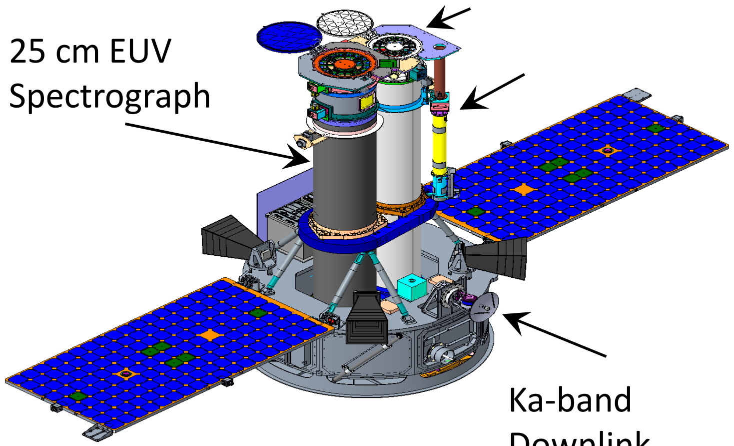
IRIS

- 20cm UV telescope
- Slit-jaw imager & slit-spectrograph
- Four 2061x1056 pixel CCDs, windowed readout, avg. cadence 4s
- 20GB/day (uncompressed Level 2)
- Launch 2013
- 12 downlinks/day

MUSE

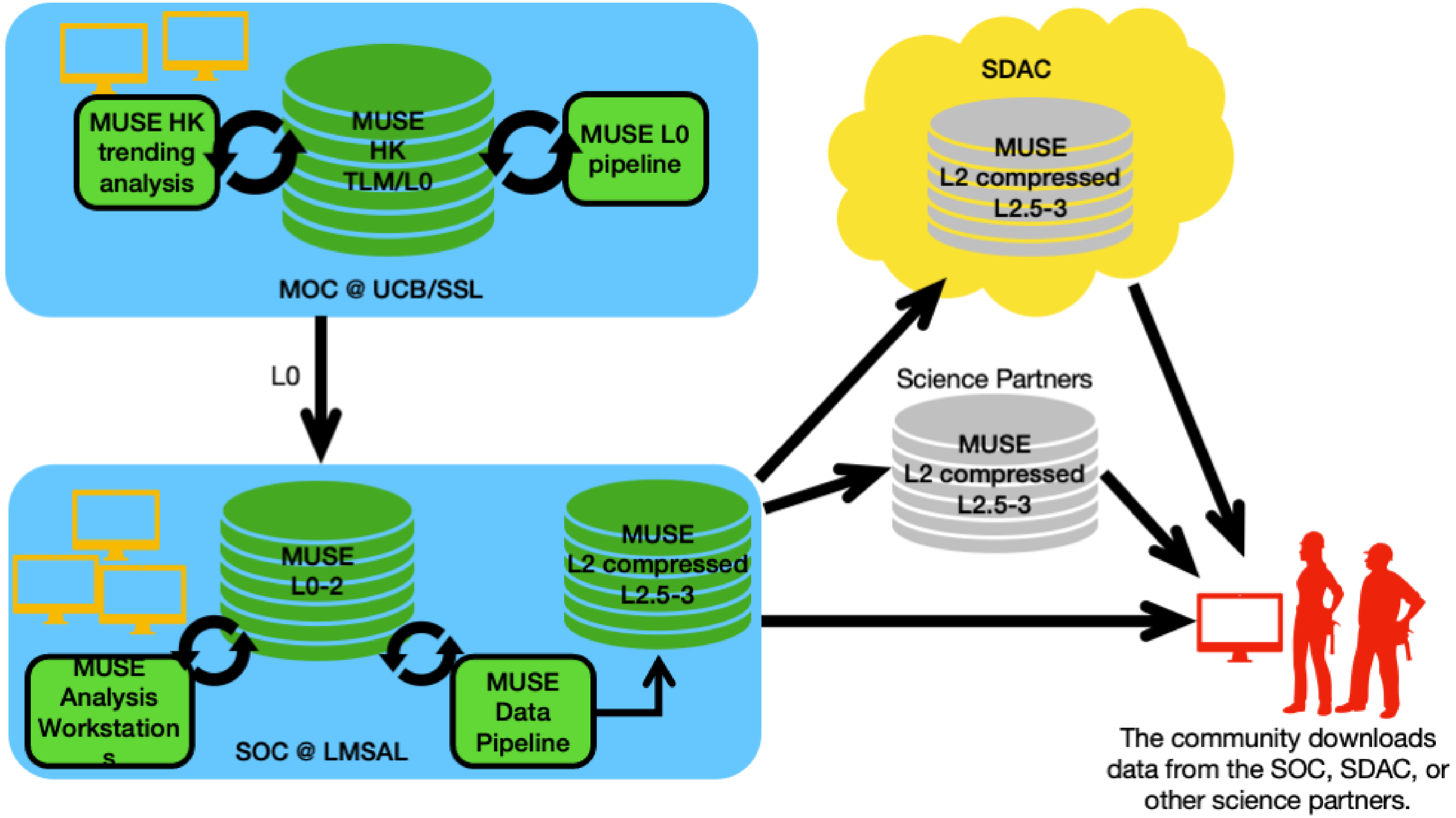
- 25cm EUV telescope for multi-slit spectrograph (SG)
- 20cm EUV telescope for context imager (CI)
- SG: three 1024x1024 CCDs, avg. cadence 1s
- CI: 4096x2048 CCD, avg. cadence 4s
- 844 GB/day (uncompressed Level 2)
- Launch 2027
- 8 downlinks/day





Baseline Plan for MUSE Data Flow

- The Misson Operations Center (MOC) at University of California, Berkeley Space Sciences Laboratory receives the telemetry data
  - Creates L0 data
  - Trends HK data
- The Science Operations Center (SOC) receives L0 data
  - Creates calibrated, higher level data products including movies, media, etc.
  - Analyses and verifies data products
  - Provides data products to NASA Solar Data Analysis Center's (SDAC) cloud and other science partners
  - Serves the data directly to the community



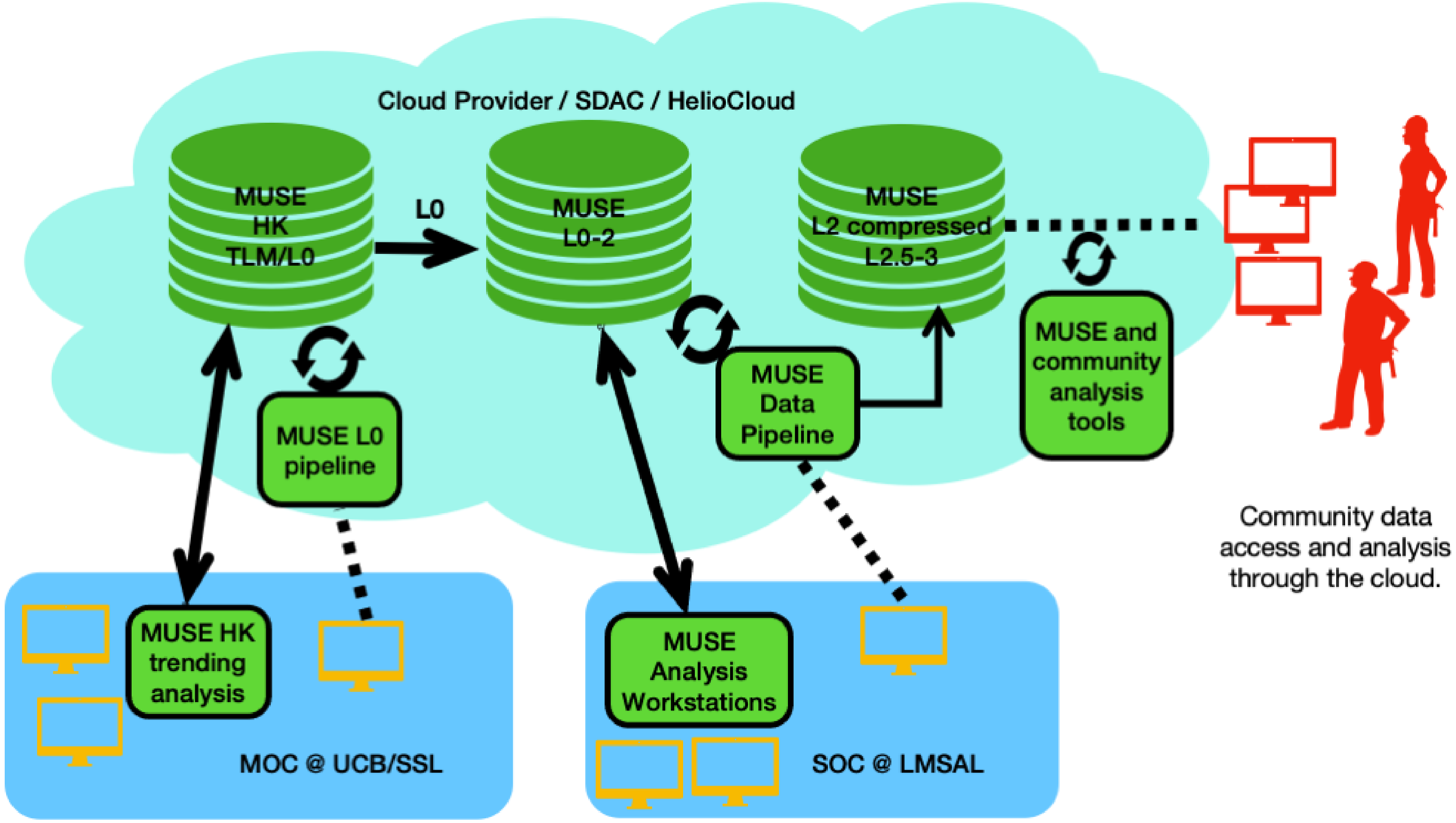
Potential Cloud-based Plan for MUSE

MOC places telemetry into Cloud storage and creates Level 0.

The SOC runs the MUSE pipeline in the cloud.

Processes are monitored from the MOC and SOC.

- No local storage nor compute servers needed
- Data is directly available to the community in a HelioCloud environment.



Considerations for Cloud Storage and Computing

Local

- Storage
  - Fixed cost for existing storage after moderate increase of capacity.
  - Monthly license cost comparable to cloud archive storage (infrequent access i.e. , >90d)?
  - Infrastructure and distribution channels are already in place.
- Computing
  - Only maintenance costs after initial upgrade of some compute servers.
  - Infrastructure and pipeline management are already in place.
  - Ideal for the constant load of pipeline processing of data.

Cloud

- Storage
  - Costs ~ volume/month
    volume increases linearly with time.
  - Total cost increase quadratically with mission duration:
    - *Problematic for longer, cost-constrained extended missions*
  - Tiered storage lowers the costs considerably.
  - *Data egress costs unpredictable.*
- Computing
  - Constant cost/month for daily pipeline processing.
  - Advantage of cloud computing is scalability and on-demand resource availability.
  - Ideal for the reprocessing of mission data.
  - Easy access for the community via NASA's HelioCloud project.

Extended missions favor local solution.

Hybrid Approaches

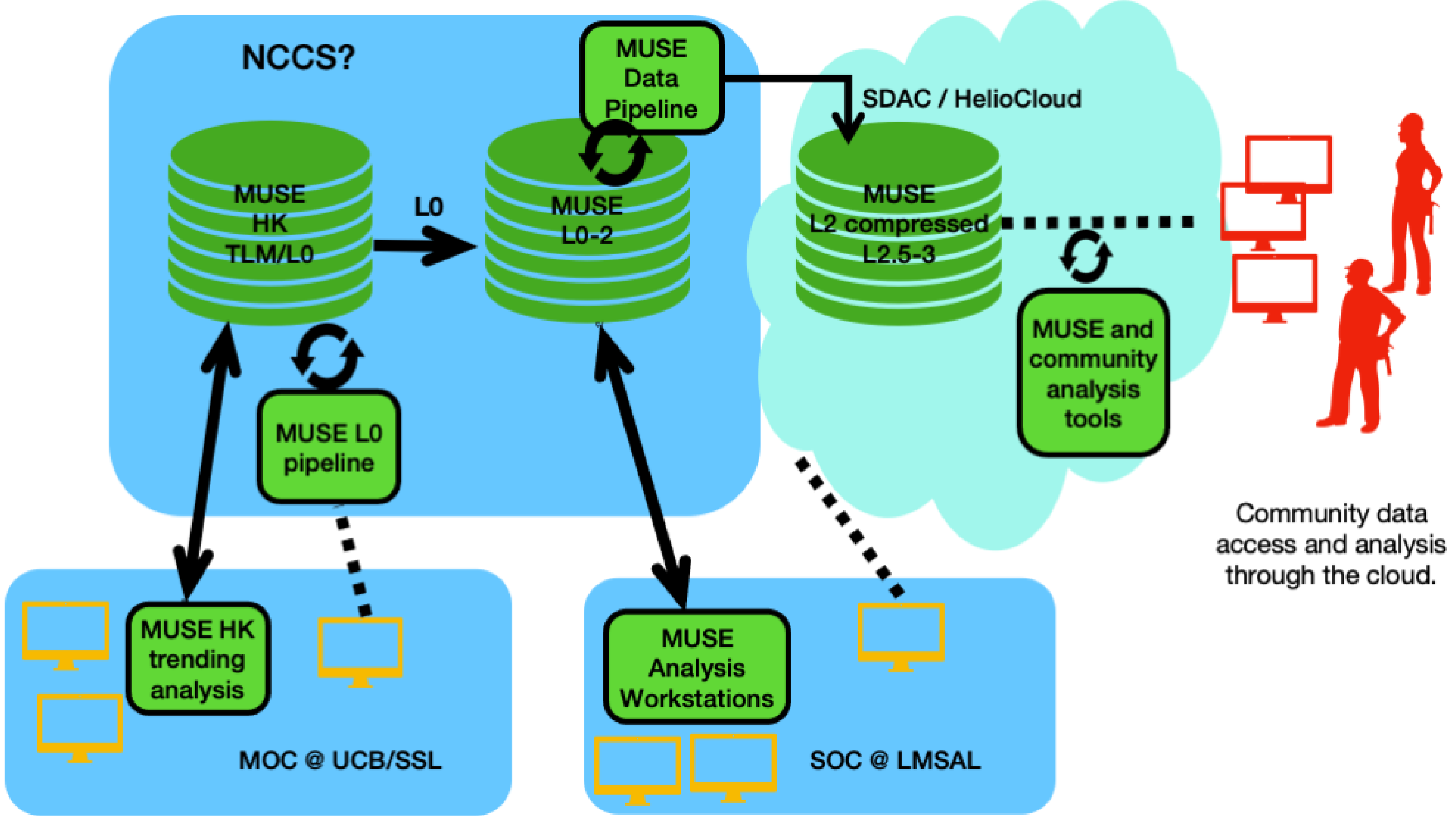
- Cloud provider?
  - Commercial project-funded.
  - NASA provided, also for extended missions.
- Lower data levels local, higher level in the cloud.
  - Like baseline but create and store data levels > 2 in the cloud.
  - Only cloud access for higher data levels.

Level	Description	Rate [GB/day]	Rate [TB/year]
Raw	Mission Data	167 (max)	60
0	Image Data	389	139
1	Pixel Corrected Image Data	389	139
2-Uncompressed	Recast Data Products	844	301
2-Media	Quicklook images and movies made from Level 2	36	13
2-Compressed	Compressed Level 2 for public downloads	389	139
2.5	Spectral Disambiguation Maps	9	3
3-Uncompressed	Moment maps from fitting or SDC	323	115
3- Compressed	Compressed Level 3 for public downloads	32	12
3-Media	Quick look images and movies from Level 3	7	2

NCCS Based Processing?

A potential alternative approach would be to perform some data processing for MUSE at National Center for Climate Simulations (NCCS).

- Ideal for re-processing large amounts of data after a pipeline update.
- Daily processing may be a too large commitment.
- Would require transfer of significant amount of data between SOC, cloud and NCCS, adding complexity.



Summary

- IRIS and MUSE are creating similar kind of data, but MUSE has a 40x higher data rate.
- Baseline plan for MUSE data management during prime mission:
  - Local storage and processing, distribution of prime mission data also via SDAC cloud.
  - Local data management is feasible.
  - Relies on well-tested procedures and existing infrastructure (after partial update).
  - Costs well understood.
  - *Easy and risk free.*

- Evaluating Cloud storage and processing for MUSE.
  - Easy scalability for reprocessing.
  - Development of new procedures for cloud data access, processing, and distribution.
  - Easy community access via the HelioCloud, no large downloads required.
  - *Associated with rising long-term costs and benefits for the community for large data projects.*

- Alternatives:
  - (Re-)Processing at NCCS.
  - Cloud processing and distribution for higher data products.

- Extended mission approach TBD!

LOCKHEED MARTIN

Backup: Existing Data Processing Center at LMSAL

- Facility:
  - Located in the climate-controlled facility at Lockheed Martin's Solar and Astrophysics Laboratory in Palo Alto, CA.
  - Important components on Uninterruptible Power Supply (UPS) and building back-up power.
- Current Components:
  - Main data processing cluster has >1000 CPU cores.
  - 2,200 TB file server.
  - High-speed (32 Gbps) interconnect between data processing cluster and file-server.
  - Database and backup systems to support processing.
  - Dedicated database and web server for external data access.
  - 10 Gbps link to CENIC/Internet2.
- Currently being used for IRIS, Atmospheric Imaging Assembly (AIA), and Hinode.
- Updating aging hardware increases the storage and computing capabilities matching the increasing mission requirements.

- Still a cost to the project!