3.2 Computers and Data Management

Jared Males
Nirav Merchant, Eric Lyons

Overview

- This describes our computer system design and the closely related data management plan.
- Requirements:
  - Computer-system:
    - Perform real-time calculations with minimum latency, no more than 1/4 frame (125 microseconds)
    - Perform near-real-time loop maintenance (gain optimization, filter determination
    - Instrument control: stages, filter wheels, camera readouts
    - Support user interfaces: AO operations and Science control
  - Data Management Plan
    - Minimum requirement: save all science camera images, along with appropriate metadata
    - Goal: save ALL data, including AO system telemetry at full rate.

FLOPS Budget

- Main Loop: matrix-vector-multiply (MVM) reconstruction:
  - (2*MN – M) FLOPS per calculation
    - M = 14400 (max pixels assuming no-slope calculation*) *Note: Slope calculation consumes FLOPS too
    - N = 2048 + 97 (total actuators being controlled)
    - 3700 Hz => 0.23 TFLOPS
    - This gives 1 frame delay due to calculation
    - 1 TFLOP capacity = 1/4 frame delay
    - All of this is highly naive, we won't achieve these rates
  - NVIDIA GTX 1080 Ti: 11 TFLOPS
    - 1/48 frame delay due to calc. (naive)
    - ~$700 each --> we can meet the MVM requirement with minimum possible delay due to reconstruction
- Other Main Loop Tasks:
  - PyWFS image processing
  - Integration and Filtering (control law)
  - DM safety checks, applying commands, etc.

Other FLOPS Heavy Tasks

- Gain Optimization:
  - Need PSDs of each mode (~2200)
  - Optimum technique: 30 FFTs of 3700 points (1 sec), averaged
  - Probably need rolling average with frequent updates => basically 2200 FFTs calculate eachsec
  - Gain optimization includes numerical minimization
- LQG
  - Kalman filter system determination (line identification)
- Predictive Control:
  - Poyneer Method: layer identification and filter determination
  - EOF method: SVDs of size ~10k X 2200
  - LP method: efficient Levinson Recursion with gain opt
  - Both EOF and LP result in large filters (time history)
- Other loops
  - LOWFS (same tasks, just fewer pixels and actuators)
  - FPWFS (similar tasks)
  - Pupil alignment, PyWFS pupil stabilization

Computer Inventory

- Computer Needs:
  - Real-Time control Computer (RTC)
    - PyWFS reconstruction, HODM and LODM control
    - Various gain optimization and filter calculation tasks
    - Real-time telemetry
  - Instrument Control Computer (ICC)
    - LOWFS reconstruction and LOWFS-DM control
    - Science camera acquisition (2 cameras)
    - Motion control (> 26 DOF)
    - TCS Telemetry monitoring
  - AO Operations Computer (AOC)
    - In control room, custom workstation for AO operations
    - AO operator workstation
    - Server for Astronomer User Interface
- FLOPS Requirement
  - Main loop requires 1 1080 Ti class GPU
  - SCExAO Experience: ~4x 1080 Ti GPUs can handle everything, but Pred. Con. may require more
  - Managing real-time processes is easier with more GPUs, even if each one not 100% loaded

Data Management Requirements

- Goal: save everything for later use
  - Long term system analysis (e.g. see Bailey+, SPIE 2016)
  - Data reduction

Table 1:: RTC Data Rate and Capacity Requirements.

Source	Name	Type	Bits /pix	Size [pixels]	Rate [frames/sec]	Rate [B/sec]	Notes
PyWFS	WFS Images	`ushort`	14	14400	3700	93247400	Assumed stored @14 bits
LODM	LODM Commands	`single`	32	97	2000	780000	
HODM	HODM Commands	`single`	32	2048	3700	30317800	
Misc.	Various	`uchar`	8	10485760	1	52428810	A 50 MB/sec margin.
			MB/sec	TB/hr	TB/night	TB/run	
		Total:	168.6	0.58	5.79	17.4	10 hr/night x 3/run

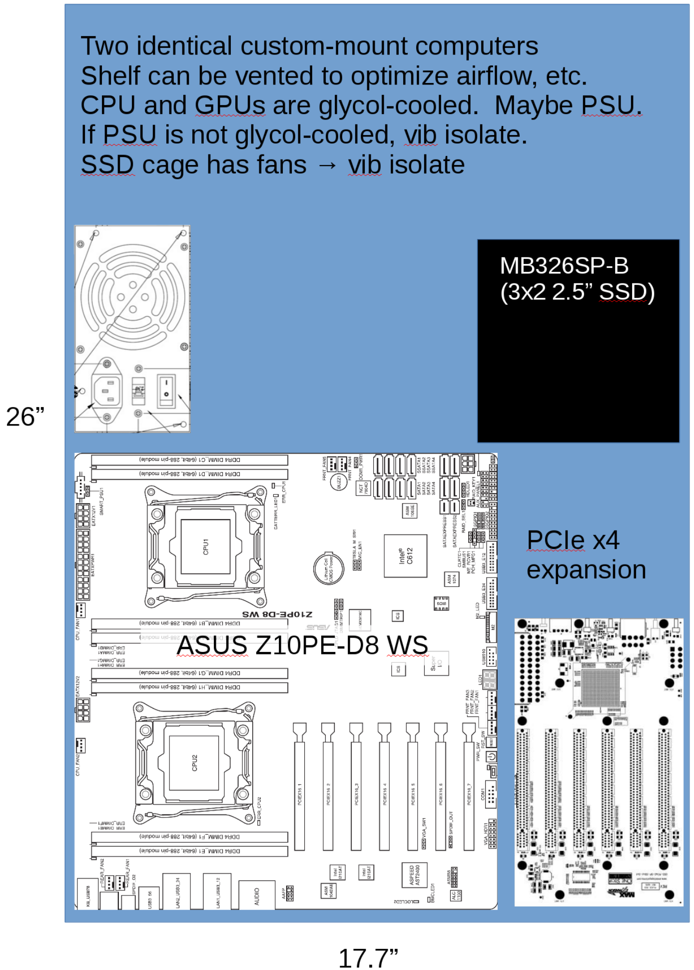
Table 2:: ICC Data Rate and Capacity Requirements.

Source	Name	Type	Bits /pix	Size [pixels]	Rate [frames/sec]	Rate [B/sec]	Notes
SCI-1	Science image	`ushort`	16	1048576	26	54526004	
SCI-2	Science image	`ushort`	16	1048576	26	54526004	
TCS Telem	Various	`uchar`	8	1	261.9	383.9	Avg rate, see App. C.
Pupil Alignment	Pupil Images	`ushort`	16	4096	1	8194	
LOWFS	LOWFS Images	`ushort`	16	262144	56	29360240	
LOWFSDM	LOWFS Commands	`single`	32	97	2000	780000	
Misc.	Various	`uchar`	8	10485760	1	52428810	50 MB/sec margin.
			MB/sec	TB/hr	TB/night	TB/run	
		Total:	130.8	0.45	4.49	13.47	10 hr/night X 3/run

Data Management Requirements

- Requirements:
  - Write speeds: > 170 MB/sec (mostly sequential write)
  - Nightly Capacity: > 6 TB each for RTC and ICC (10 hr night)
  - Run Capacity: > 31 TB total (3 nights)
  - Must get this back to UA, store, and make it accessible.
- Following design uses today's technology and prices
  - Shows we can achieve it
  - But... we won't buy anything until we need it (1-2 years)
  - That means it will be cheaper and/or easier to achieve

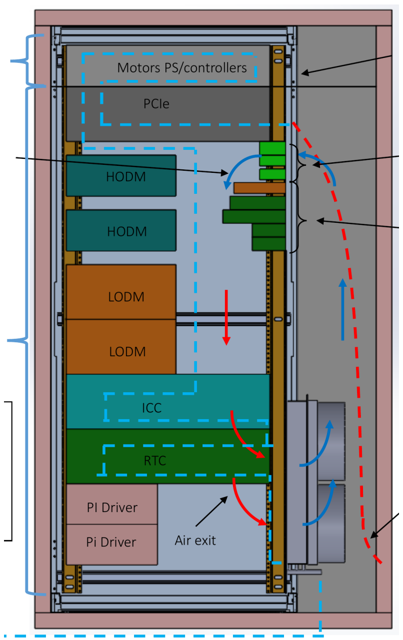
Custom Rack Mount Computers

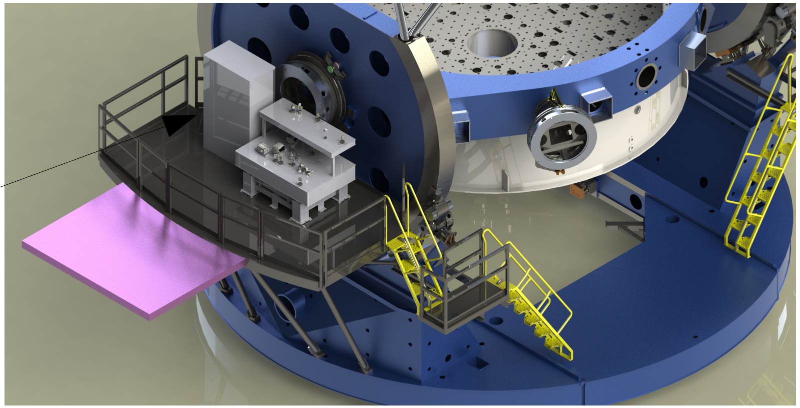


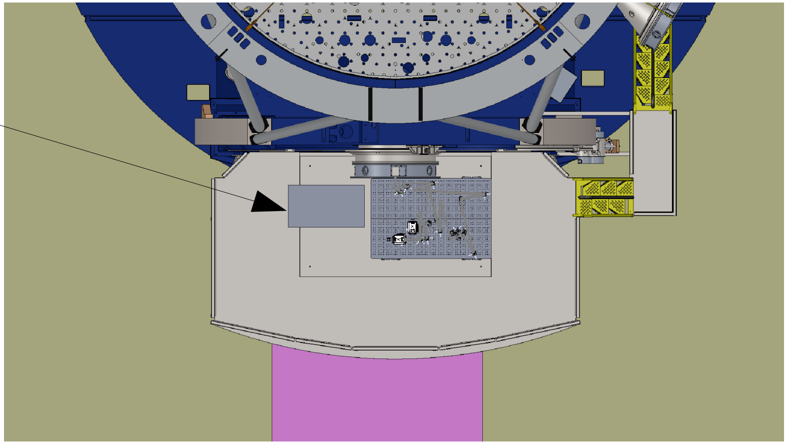
- RTC and ICC will be on platform next to instrument
- No case, we will drill our own ATX hole pattern (etc) in an Aluminum plate (a rack shelf)
  - Max space and airflow and ease of cabling
  - What we did for VisAO and it works great
- Specs on following slides
- PCIe expansion card gives extra slots
- CPU and GPUs are direct liquid cooled
  - See thermal design by Dan Alfred

Electronics Rack

Electronics rack on platform contains RTC, ICC, and PCIe expansion, plus many other controllers and drivers.







For details see part 3.1

RTC Specs

Component		Manuf.	Part No.	No.	Cost Each.	Total Cost	Notes
ATX 12"x13"	Motherboard	ASUS	Z10PE-D8 WS	1	$554.08	$554.08	
	Processor (x2)	Intel	i7 6950x (10 core)	2	$1,649	$3,298.00	Liquid Cooled, 20 cores total
ATX	PS	Quiet Power	Quiet 1050W PC ATX PSU	1	$89.99	$89.99	Fan--> vibration isolated mount
	RAM	G.Skill	8x16 GB DDR 2400	1	$759	$759.00	128 GB total
	HDD Cage		MB326SP-B	1	$80.00	$80.00	Holds 6 2.5" SSDs
OS	HDD	Crucial	CT525MX300SSD1	2	$149.90	$299.80	525GB ea, RAID-1, 525GB total
Storage	HDD	Samsung	MZ-75E4T0B/AM	4	$1,480	$5,918.20	4TB ea, 4 RAID-5, 12TB total
PCIe Slots:							
PCIe Slot 1 (x16 3.0 CPU-1) PCIe Slot 2 (x8 3.0 CPU-1)	GPU	NVIDIA	1080 Ti	1	$699.99	$699.99	Liquid Cooled. If expanded, this is HIC slot
PCIe Slot 3 (x16 3.0 CPU-1) PCIe Slot 4 (x8 3.0 CPU-1)	GPU	NVIDIA	1080 Ti	1	$699.99	$699.99	Liquid Cooled. If expanded, this is HIC slot
PCIe Slot 5 (x16 3.0 CPU-2) PCIe Slot 6 (x8 3.0 CPU-2)	GPU	NVIDIA	1080 Ti	1	$699.99	$699.99	Liquid Cooled
PCIe Slot 7 (x16 CPU-2)	Expansion Card	Cyclone Microsystems	PCIe2-437	1	$709	$709.00	
PCIe Expansion:							
PCIe Slot 1 (x4 2.0)	HOWFS F.G. (x4	Matrox	Solios eV-CLF	1	$0	$0	Part of OCAM purchase
PCIe Slot 2 (x4 2.0)	HODM (x1)	BMC	HVA Interface Card	1	$0	$0	Included in BMC DM cost
PCIe Slot 3 (x4 2.0)	LODM	Alpao	DM Inteface Card	1	$0	$0.00	Included in Alpao DM cost
PCIe Slot 4 (x4 2.0)	10 GBs Net	Small Tree	P2E10G-2-T	1	$599	$599.00	
					Total	$14,407.04	Assumes no overhead

RTC PCIe Expansion

- COTS solution gives 6 more double-wide GTX slots
  - CUBIX Xpander 8 (have to use 2 slots to connect)

Component		Manuf.	Part No.	No.	Cost Each.	Total Cost	Notes
Rack Module	5U	Cubix Xpander	Rackmount 8 Gen 3	1	$8,799	$8,799	Possible upgrade, thermal is sized for this
Slot 1	GPU	NVIDIA	1080 Ti	1	$0	$0	Moved from Slot 1
Slot 2	GPU	NVIDIA	1080 Ti	1	$0	$0	Moved from Slot 2
Slot 3	GPU	NVIDIA	1080 Ti	1	$700	$700	Liquid Cooled
Slot 4	GPU	NVIDIA	1080 Ti	1	$700	$700	Liquid Cooled
Slot 5	GPU	NVIDIA	1080 Ti	1	$700	$700	Liquid Cooled
Slot 6	GPU	NVIDIA	1080 Ti	1	$700	$700	Liquid Cooled
Slot 7	GPU	NVIDIA	1080 Ti	1	$700	$700	Liquid Cooled
Slot 8	GPU	NVIDIA	1080 Ti	1	$700	$700	Liquid Cooled
					Total:	$12,999	

We have budgeted for this (financially, thermally, and spatially), but it is an upgrade we do not plan to start with.

ICC Specs

Component		Manuf.	Part No.	No.	Cost Each.	Total Cost	Notes
ATX 12"x13"	Motherboard	ASUS	Z10PE-D8 WS	1	$554.08	$554.08	
	Processor (x2)	Intel	i7 6950x (10 core)	2	$1,649	$3,298.00	Liquid Cooled, 20 cores total
ATX	PS	Quiet Power	Quiet 1050W PC ATX PSU	1	$89.99	$89.99	Fan--> vibration isolated mount
	RAM	G.Skill	8x16 GB DDR 2400	1	$759	$759.00	128 GB total
	HDD Cage		MB326SP-B	1	$80.00	$80.00	Holds 6 2.5" SSDs
OS	HDD	Crucial	CT525MX300SSD1	2	$149.90	$299.80	525GB ea, RAID-1, 525GB total
Storage	HDD	Samsung	MZ-75E4T0B/AM	4	$1,480	$5,918.20	4TB ea, RAID-5, 12TB total
PCIe Slots:							
PCIe Slot 1 (x16 3.0 CPU-1)							
PCIe Slot 2 (x8 3.0 CPU-1)	GPU	NVIDIA	1080 Ti	1	$699.99	$699.99	For LOWFS Recon.
PCIe Slot 3 (x16 3.0 CPU-1)	SCI-1 F.G. (x4)	Matrox	Solios eV-CLB	1	$925.00	$925.00	Andor 888
PCIe Slot 4 (x8 3.0 CPU-1)	SCI-2 F.G. (x4)	Matrox	Solios eV-CLB	1	$925.00	$925.00	Andor 888
PCIe Slot 5 (x16 3.0 CPU-2)	LOWFS F.G. (x4)	Matrox	Solios eV-CLB	1	$925.00	$925.00	Andor 897
PCIe Slot 6 (x8 3.0 CPU-2)							
PCIe Slot 7 (x16 CPU-2)	Expansion Card	Cyclone Microsystems	PCIe2-437	1	$709	$709.00	
PCIe Expansion:							
PCIe Slot 1 (x4 2.0)							
PCIe Slot 2 (x4 2.0)							
PCIe Slot 3 (x4 2.0)	LODM	Alpao	DM Inteface Card	1	$0	$0.00	Included in Alpao DM cost
PCIe Slot 4 (x4 2.0)	10 GBs Net	Small Tree	P2E10G-2-T	1	$599	$599.00	
					Total	$15,782.06	Assumes no overhead

AOC Specs

Component		Manuf.	Part No.	No.	Cost Each.	Total Cost	Notes
ATX 12"x13"	Motherboard	ASUS	Z10PE-D8 WS	1	$554.08	$554.08	
	Processor (x2)	Intel	i7 6950x (10 core)	2	$1,649	$3,298.00	Liquid Cooled, 20 cores total
ATX	PS	Quiet Power	Quiet 1050W PC ATX PSU	1	$89.99	$89.99	Fan--> vibration isolated mount
	RAM	G.Skill	8x16 GB DDR 2400	1	$759	$759.00	128 GB total
	HDD Cage		MB326SP-B	1	$80.00	$80.00	Holds 6 2.5" SSDs
OS	HDD	Crucial	CT525MX300SSD1	2	$149.90	$299.80	525GB ea, RAID-1, 525GB total
Storage	HDD	Samsung	MZ-75E1T0B/AM	4	$355	$1,419.80	1TB ea, RAID-5, 3TB total
PCIe Slots:							
PCIe Slot 1 (x16 3.0 CPU-1)							
PCIe Slot 2 (x8 3.0 CPU-1)	GPU	NVIDIA	1080 Ti	1	$699.99	$699.99	For display purposes, 2 monitors
PCIe Slot 3 (x16 3.0 CPU-1)							
PCIe Slot 4 (x8 3.0 CPU-1)	GPU	NVIDIA	1080 Ti	1	$699.99	$699.99	For display purposes, 2 monitors
PCIe Slot 5 (x16 3.0 CPU-2)	10 GBs Net	Small Tree	P2E10G-2-T	1	$599	$599.00	
PCIe Slot 6 (x8 3.0 CPU-2)							
PCIe Slot 7 (x16 CPU-2)							
					Total	$8,499.65	Assumes no overhead

– This will be in the control room, so in a workstation case with 4 monitors, keyboard, mouse, etc.

Specification Summary

- Using SSD Drives:
  - 4x 4TB in RAID 5 in RTC and ICC each:
    - (1-1/N) = 12 TB capacity on each of RTC and ICC
    - Single drive sequential write speed: 520 MB/sec (only need 32% of this)
    - This meets the nightly data management spec
- 20 true CPU cores per machine, 128 GB RAM
- On platform: 10 NVIDIA 1080 Ti
  - 113 TFLOPS from 35,840 CUDA cores

How Do We Get It Home?

- Extensive network transfer testing from LCO to Tucson with existing MagAO data shows that we can't do it.
  - Can be as low as ~100 GB a day (depends on load)
  - So it could take 310 days to transfer a 31 TB run's worth
- Solution: we pack it home in 2 of these:

5x 8TB RAID Array. 8.5" X 4.6" X 6.8".

Just drives, no computer

RAID 5: 32 TB
– enough for 1 run
– Unload RTC and ICC to this each day.

$2395 each, 2x for redundancy

- Alternative to shipment: with collaborators Merchant and Lyons (UA/CyVerse) we are exploring using a host in Chile for these units
  - Instead of transporting all the way home, they could be delivered to a site in La Serena or Santiago where network transfer will be possible

How Do We Store It At UA?

- We buy a 136 TB RAID machine, which will be hosted by CyVerse (not in Steward)
- Connected to CyVerse iRODS. Interface through CyVerse
  - Currently stores and manages 2.5 peta-bytes of data
- Automagically gives us distribution to users, with access rights, can enforce propietary periods, etc.
- ~ $20 K each
  - 1 at UA, 1 at TACC for offsite backup
- Can hold 4 runs worth
  - We may only save non-science telemetry for a set period of time

- Ongoing hosting/management cost: 10 hrs, ~$720/year.

Computing Budget

- Main Components:
  - RTC: $14,407
  - PCIe Xpansion: $12,999
  - ICC: $15,782
  - AOC: $8450
  - Transport: 2x $2395 = $4790
  - Storage: 2x $20k = $40k
  - Various Spares: ~$5k
- Total: $101,428

Reality

- It is unlikely that we will operate at full-rate 10 hours/night
  - With pointing overheads, etc, this probably has ~2.5 hr margin.
- Data specs assumed full-frame and full-rate for all cameras. LOWFS, especially, will use small windows and therefore much smaller rates most of the time
- Someday, runs might be longer than 3 nights
  - But space and speed will continue to get cheaper
- **This design can meet our needs today, and has capacity margin**
  - PCIe expansion may not be needed (GPU capacity is ever expanding)
  - More space will be available for same money in the future.
  - Can cut off-site backup if space needed at UA

Can We Reduce This Much Data?

- We have developed a cloud-based data reduction system through our collaboration with Nirav Merchant, Eric Lyons, and Asher Haug-Baltzell (CyVerse / BIO5)
  - Distributed computing solution to image processing / reduction problem: Findr
  - See: http://adsabs.harvard.edu/abs/2016SPIE.9913E..0FH and http://adsabs.harvard.edu/abs/2017ApJ...836..223W
  - We recently obtained 10^6 CPU hours on NSF Jetstream: <u>resources are available.</u>

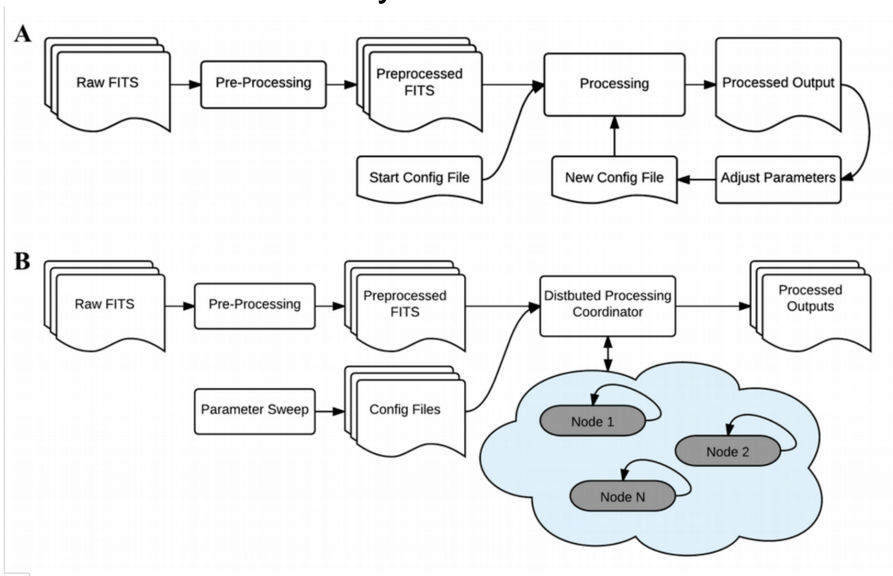
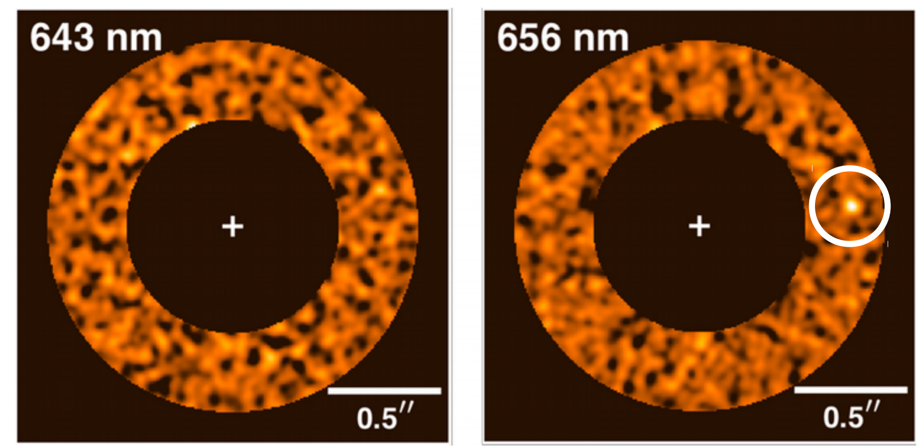


Figure 1. PSF subtraction workflows before (A) and after (B) cloud distribution. (A) Before cloud distribution, determining optimum parameters for KLIP reduction was a slow, manually intensive process. After each reduction, parameters would be adjusted and results compared until a high signal-to-noise ratio was achieved. (B) Distribution of the reduction process to cloud computing resources leads to a simplified workflow and enables full sweeps investigating all parameter combinations. This workflow is faster, less prone to bias, requires less manual input, and leads to statistically sound conclusions.



Images of GQ Lup B reduced with Findr on the NSF Chameleon cloud system. See Haug-Baltzell+ (2016) for details, and Wu+ (2017) for the interpretation.

Detailed Data Management Plan

Here we describe in detail how the data collected by the MagAO-X project will be managed.

1. Products of the Research
The following data products will be generated from MagAO-X related laboratory testing: (1) deformable mirror (DM) influence functions; (2) DM flat measurements; (3) pyramid wavefront sensor (PyWFS) characterization; and (4) detector characterization. On-sky engineering data will include: (1) PyWFS pupil images; (2) adaptive optics (AO) system telemetry; and (3) science camera images (for Strehl ratio measurements, etc.). These data will be made available using the NSF-funded CyVerse (formerly iPlant) infrastructure described below. All developed techniques and engineering results for the MagAO-X project will be made publicly available through one (or more) PASP papers and SPIE conference papers. These will be made publicly available on arxiv.org, and we will make pre-prints available directly from the project website (https://visao.as.arizona.edu).

Science observations enabled by this project will produce raw images and spectra, as well as AO system telemetry. These will be analyzed and used to produce scientific publications by the responsible astronomer. All scientific publications based on MagAO-X data will be made publicly available on arxiv.org, and we will make pre-prints available directly from the project website (https://visao.as.arizona.edu).

2. Data and Metadata Quantity and Format

All raw data will be saved in HDF5 files, including science images. This is a departure from the norm in astronomy, however the standard FITS (Pence et al., 2010) format is quite outdated and has limited metadata description capabilities. Furthermore, we expect HDF5 to provide better performance in terms of access speed (reading and writing). To support legacy data processing systems, we will follow the HDFITS standard (Price et al, 2015). This standard is designed for converting from HDF5 to FITS and provides python code to do so. This includes generation of FITS header metadata.

We will follow the same standards used for VisAO for our metadata collection and storage. Metadata describing the AO system configuration, weather, science camera configuration, and other details relevant to observations will be stored in appropriate HDF5 attributes, and published in FITS headers when converted.

3. Access to Data and Data Sharing Practices and Policies
Laboratory and Engineering Data: All laboratory and on-telescope-engineering data (whether daytime testing or night-time observations) generated for this project will be made publicly available after a three-month period. This is primarily to give the team time to analyze the data and perform quality control. Access to the data will be provided using the NSF-funded CyVerse infrastructure. CyVerse provides several resources for scalable data management with connections to large-scale and distributed computing resources, namely:

- **The Data Store (iDS):** The iDS is the foundation for data management, and is built on the integrated rule-oriented data system (iRODS) and delivers cloud data storage/transfer accessible through command line, software (iDrop), and web-based systems. Advantages of iDS include: (1) scalable, distributed, and redundant storage; (2) the appearance of a single file system to the user that nonetheless consists of easily expanded distributed data servers; (3) parallel file transfers of up to 1 Gb per second; and (4) strong support for metadata. All data generated by this project will be made publicly available through the iDS.
- **CyVerse's Authentication Service (iAS):** iAS, built on the secure identity system Shibboleth, validates researchers on CyVerse's systems, allows researchers to move among CyVerse's resources, and permits other systems to validate users. Systems federated with CyVerse's Authentication Service can then retrieve data from iDS, analyze them, and then send the analytical results back for downstream visualization.
- **CyVerse Atmosphere:** This is CyVerse's cloud computing platform and is based on OpenStack for managing virtual machines (VM) images and instances. Any complicated software stack developed in this project that cannot be integrated in CyVerse's Discovery Environment will be made available on a VM for other researchers to use.
- **CyVerse Discovery Environment (DE):** This is CyVerse's rich web-based system that allows users to manage their data in the iDS and run analyses on CyVerse's computing grid. All command-line driven workflows will be integrated in the DE for other researchers to use.
- **Powered by CyVerse:** This program offers technical assistance to analytical software systems that wish to utilize CyVerse's CI to power their bioinformatics platforms.
- **CyVerse Science APIs:** This service will be a gateway for CyVerse and third-party platforms to register web-based APIs for data exchange and services.

Science Observations: Prior to publication, by default the raw data from science observations enabled by this project will be subject to a 24-month proprietary period during which the responsible astronomer will have exclusive access. The CyVerse system will automatically enforce this proprietary period by leveraging "rules" used by the Data Store (part of the iRODS architecture). Exceptions to this policy will be made based on astronomer requests and the policies of the Magellan partner institutions. Published data will be made available upon request, with approval by the responsible astronomer, using the CyVerse infrastructure.

4. Policies and provision for re-use, re-distribution and products of derivatives

Raw data access will come with a simple request to cite the relevant MagAO-X-based publication if used in a public product. Reproductions of published images made available on the MagAO-X project website will always include citations to the primary source, with standard academic citation practices applied. All software generated by this project will be released under the MIT license for use without restriction, and integrated in the appropriate CyVerse resource for ease of reuse by the scientific community. No reach-through rights or intellectual property rights will be claimed on the outcomes of this research, including associated data, software, and hardware designs.

5. Archiving of Data

We will utilize the CyVerse infrastructure at the University of Arizona (UA) to store data for the life of the MagAO-X system and beyond. All data will be saved in FITS format, preserving all relevant metadata in the headers. Management of data at UA will be coordinated through three phases; initial, near term, and long term. The specific practices will depend on several factors, especially the available technology and associated costs, data quantity, and long-term public access. **Initial** data management will facilitate robust storage and later analysis phases. At least two copies of all data files will reside on redundant disk systems; a working copy will be backed up by at least one archive copy. All analyses will occur using the working copy. The primary goal of these considerations is to preserve raw data in case of catastrophic hardware failure. **Near-term** data management will support the project's computational analyses and facilitate public access. Once data are made public, they will be made available for public anonymous access through a specific area of the CyVerse Data Store. **Long-term** the data will be safely stored for later use. The Data Store replicates data between computers at UA and the Texas Advanced Computing Center (TACC). Currently, each site has at least 2 petabytes storage.

Access permissions will be managed automatically. Metadata recording the owning astronomer will be used to assign user permissions and provide a searchable database. Data will automatically become public at the end of the proprietary period.

6. Documentation

All data sets and software generated by this project will be documented and linked to using a wiki provided by UA, CyVerse, or another publicly available online documentation site. This site will also serve as a coordination hub between groups for documenting how data were handled, pre-processed, post-processed, published, and made publicly available.

References

Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A.; Stobie, E. "Definition of the Flexible Image Transport System (FITS), version 3.0", Astronomy and Astrophysics, Volume 524, id.A42, 40 pp.
http://adsabs.harvard.edu/abs/2010A%26A...524A..42P

Price, D. C., Barsdell, B. R., Greenhill, L. J. "HDFITS: Porting the FITS data model to HDF5", Astronomy and Computing, Volume 12, p. 212-220.
http://adsabs.harvard.edu/abs/2015A%26C....12..212P