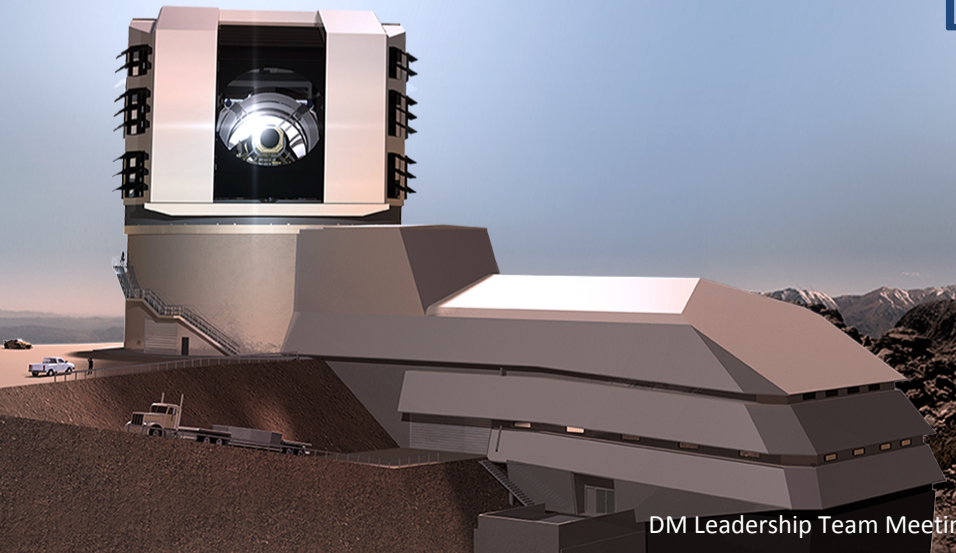
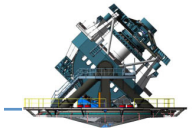
Lossy Compression WG Status

Robert Gruendl

DM Leadership Team Meeting
November 2, 2017



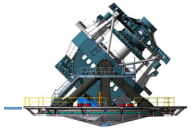


Lossy Compression WG Charge (LDM-582)

This WG was chartered to understand whether it is possible to store a broader set of release products than originally planned, so that they might be more readily available (i.e., bypassing tape storage or on-the-fly reprocessing). This is only viable if a significant degree of compression is possible and the images are useful for a wide range of science use cases.

The actual charge (LDM-582) has the following:

- Define criteria for "science-usable" lossy-compressed processed images for LSST image types,
- Collect compression algorithm candidates (preferring existing "off-the-shelf" tools/libraries),
- Evaluate their compression ratios (at "science-usable" quality),
- Evaluate constraints on processing that usage of compression may impose (e.g., avoidance of repeated re-compressions),
- Quantify the savings from application of lossy compression, in the context of the LSST Sizing Model (LDM-144)
- Make recommendations on which image types to lossy-compress, the algorithms to apply, and the description of processing constraints these would impose.



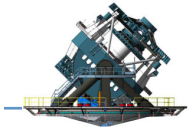
WG Membership

- Robert Gruendl (NCSA; Chair),
- Paul Price (Princeton),
- Bob Armstrong (Princeton),
- Krzysztof Findeisen (UW),
- Sophie Reed (Princeton),
- Eric Morganson (DES/NCSA; observer),
- Ben Emmons (EPO Tucson; observer)

- Completion date for the charge is November 30, 2017
- WG Chair reports to DM Project Manager and SSST weekly

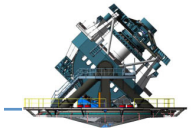
Plan of Attack

- Examine SRD for amount of loss that can be endured, pipelines for strategic points where compression may be useful.
- Measure algorithmic performance (a utility developed for HSC processing exists that partially addresses this to compress image
  - Deploy tests first as stand-alone
  - Eventually deploy within a pipeline for larger scale tests
- Examine images to measure added noise and check for significant artifacts (simple difference between original and compressed image)
- Compare catalogs products generated from never-compressed and compressed images.
- Extend image/catalog tests to examine COADDed products constructed from never-compressed and compressed images



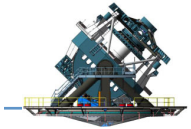
Current "Progress"

- WG Convened in September 2017
- Currently the most likely candidate is a RICE algorithm (or a variant)
- Initial test set will use HSC data
  - Some expansion will be necessary to obtain non-cosmological survey cases (e.g., nebulosity & crowded fields). Either from DECam or HSC
- Initial plan for catalog measurements to be examined are:
  - Position, flux (along with associated uncertainties). This code exists
  - Sizes (currently thought to be adequate to cover shapes)
  - Reed is charged with adding add sizes (and to make addition of further quantities relatively easy)



Current "Progress"

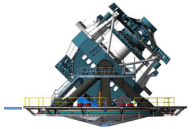
- Summary of AP pipeline (Findeisen)
  - Primary product where compression might be viable/useful is the 30-day store of PVI products (used for pre-covery/forced photometry)
  - This would yield only modest storage savings
- Summary of DRP products is not yet complete
  - Current most likely candidate would be PVI image products (with a goal that compressed versions can serve rather than re-reduction for most user cases).
  - COADD products are the second most likely product to be considered.
  - Current best suggestion is that measurements made in DRP would occur prior to compression
  - This dues opens an issue that the user community be properly informed about the usage of these products (and the reproducablity of DRP processing measurements)



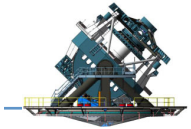
Current Prognosis

Progress has been slow at best

- WG has capable hands but only modest availability for work
- Addition of Reed is meant to correct that but much still hinges upon non-LSST effort (mostly Price's cross-section)
- Preliminary results are still likely within the current charge's timeframe (Nov 30) but robustness/completeness are becoming less and less likely.
- At this time I think it likely the WG can have a good estimate for the suitability of a RICE type algorithm along with a methodology to test its suitability more widely (e.g. on a wider range of products). This may be enough for planning/estimates going forward (i.e. a benchmark to be beat by alternative candidate algorithms)
- It is very unclear where the effort might come from to investigate a variety of algorithms.



Extra's showing a synopsis of an earlier study



The DES Experience with FPACK (RICE)

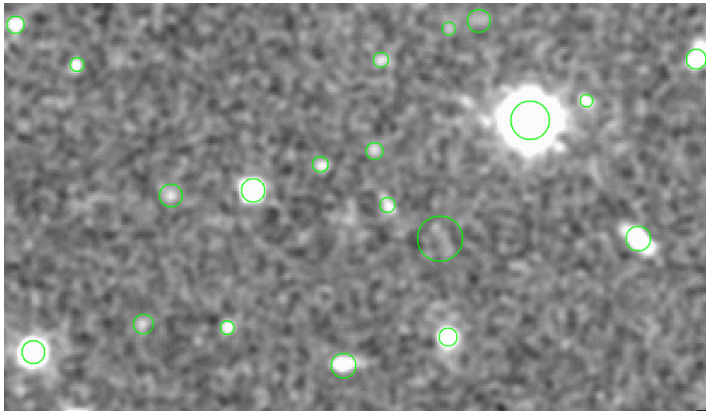
DES Uses Lossy Compression:

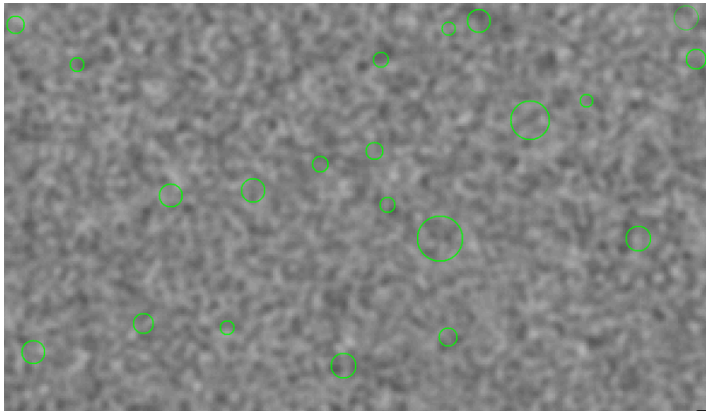
- Each CCD and COADD image is comprised of SCI, WGT, and MSK HDUs (FITS format)
- SCI and WGT use RICE compression with a directive to identically preserve zero's
- MSK is ushort (16bit unsigned integers) so uses PLIO (lossless)

The quantization factor (q) used by RICE was set to 16 after a series of experiments examining the compression factor achieved along with the resulting image and catalog products.

Image Performance

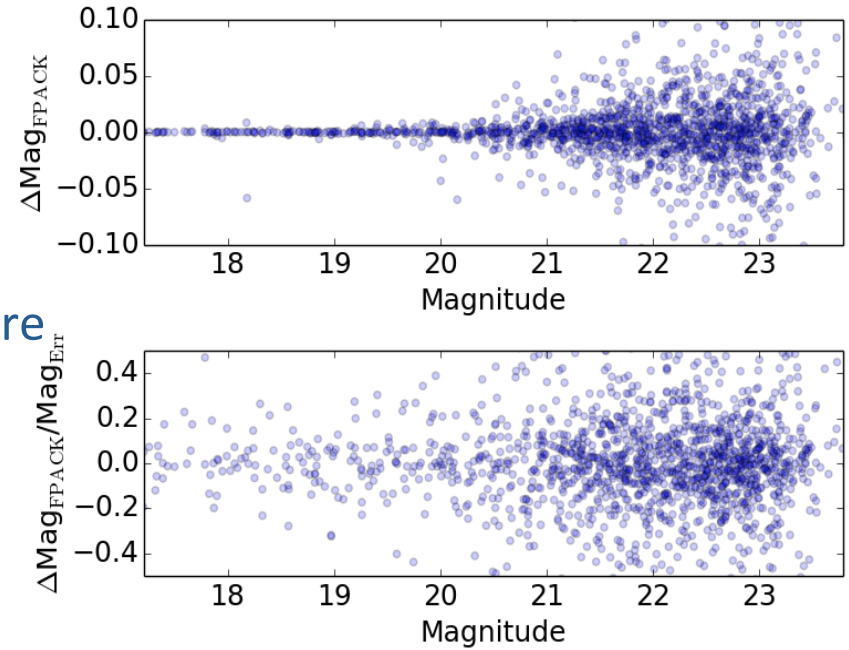
- FPACK adds noise:
  $(12q^2)^{-1/2}$ RMS_{image}
- For q = 4, this is 7.2%
- We calculate 8.3%
- 0.3% when added in quadrature

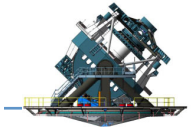




Catalog Performance

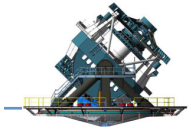
- Detrend image & mask artifacts
- Catalog objects
- FPACKed and unpacked image
- Made another catalog
- Difference
  - RMS ~ 0.2*Mag_{Err} for q = 4
  - Increases Mag_{Err} 2% in quadrature
- Lose ~1% of sources
- Gain ~1% new sources
- Similar population to outliers





Varying q

q	RMS_{Diff}/RMS_{Image}	RMS(ΔMag)	$RMS(\Delta Mag/Mag_{Err})$	Outliers	Compression
4	0.083	0.021	0.20	11	7.7
8	0.041	0.012	0.12	4	6.5
16	0.021	0.0057	0.055	5	5.6
32	0.010	0.0031	0.027	0	4.9
64	0.005	0.0017	0.015	0	4.4



Guidelines

- Delete notes and save this template as your presentation
- No drop shadows
- Only use the font Calibri
- Please do not alter template