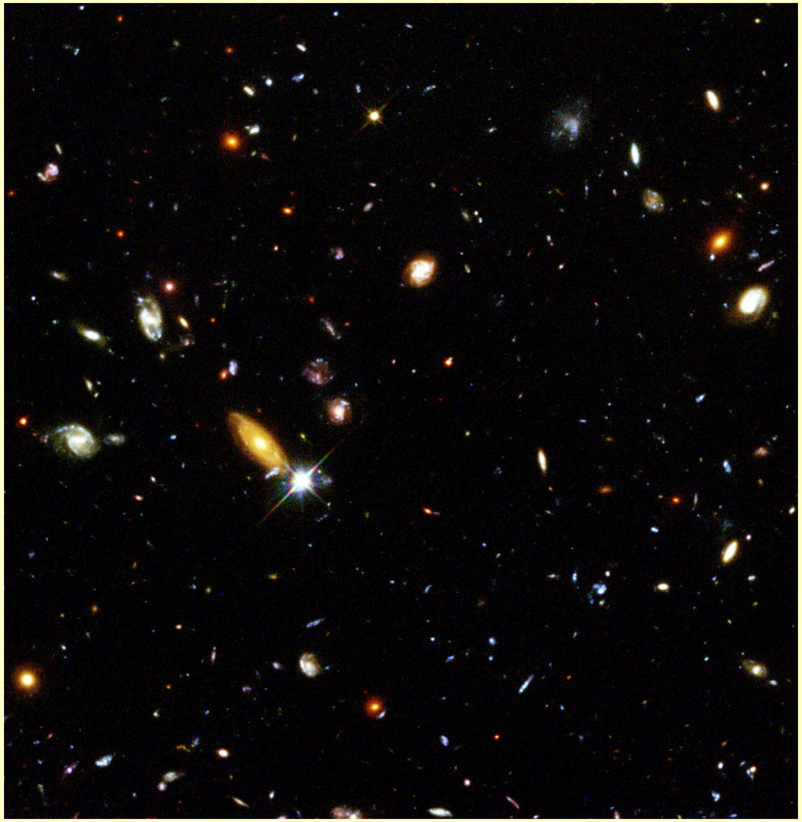
Data reduction for large surveys



Pierre Astier
LPNHE/IN2P3/CNRS
Universités Paris 6&7

Large imaging surveys : instruments

ground

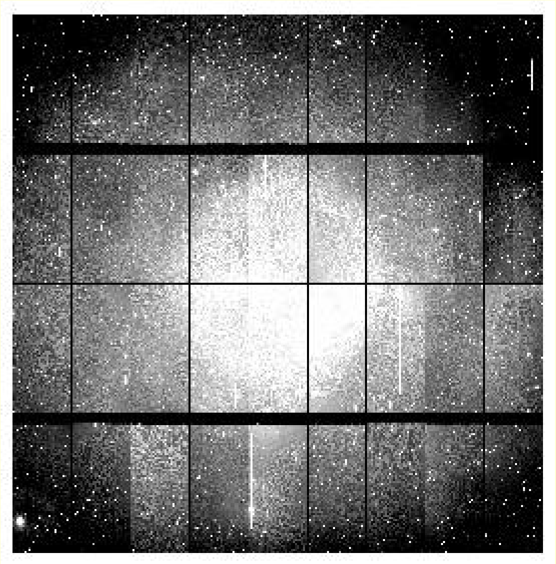
space

	FOV	diameter	first light	status	who/where
Megacam on CFHT	1 deg2	3.6m	2002	running	Mauna Kea
SDSS-III	7 deg2	2.5m	2008	running	Apache Point
VST @ ESO	1 deg2	2.6 m	2010	running	ESO/Paranal
VISTA @ ESO	1 deg2	4 m	2010	running	ESO
HyperSuprimeCam	~2 deg2	8 m	2012	funded	Japan/Subaru
Dark Energy Survey	2.9 deg2	CTIO-4m	2012	funded	Fermilab/CTIO
Pan StarsS	7 deg2	1.8 m	2009	funded	Univ. Hawaii
Pan StarsS 4	7 deg2	1.8 m x 4	??	not funded	Univ. Hawaii
LSST	10 deg2	8 m	2018	almost funded	NSF/DOE
WFIRST	0.7 deg2	1.5 m	2016(+)	On hold	NASA/DOE
Euclid	0.5 deg2	1.2 m	2017(+)	competing	ESA

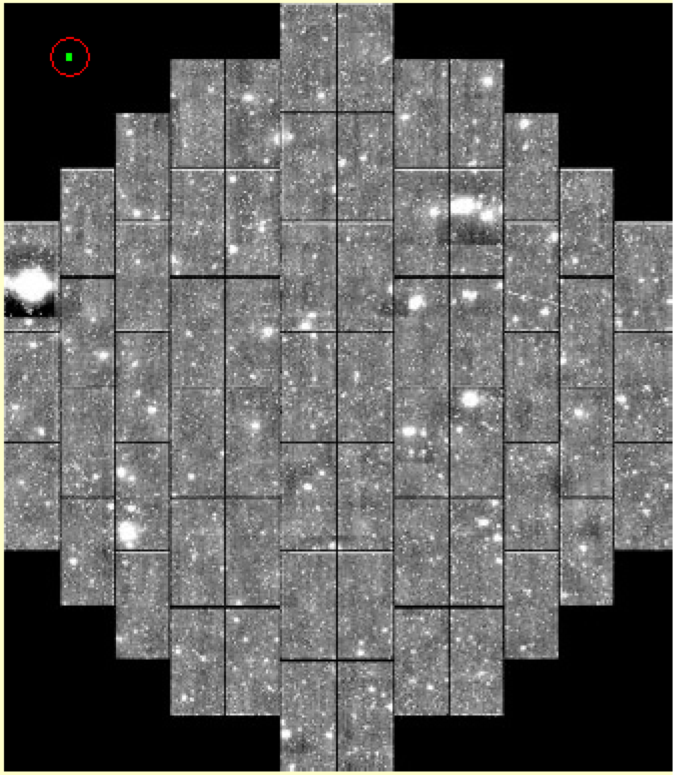
Large or very large projects which can address more than just dark energy !

Large imaging surveys : produced data

Megacam : 340 Mpixels



DEC : 520 Mpixels (sim.)



Wide field imaging cameras produce images (!) :

- Megacam (on CFHT, first light in 2002):

  30 images/hour *680 Mb * 7 hours/night *150 nights/year * 5 years

  → 100 Tb

  (similar amount for the Dark Energy Survey)

- EUCLID (2018 ? ESA space mission at L2)

  < 250 Gb/day (telemetry) * 365*5 years

  → 450 Tb

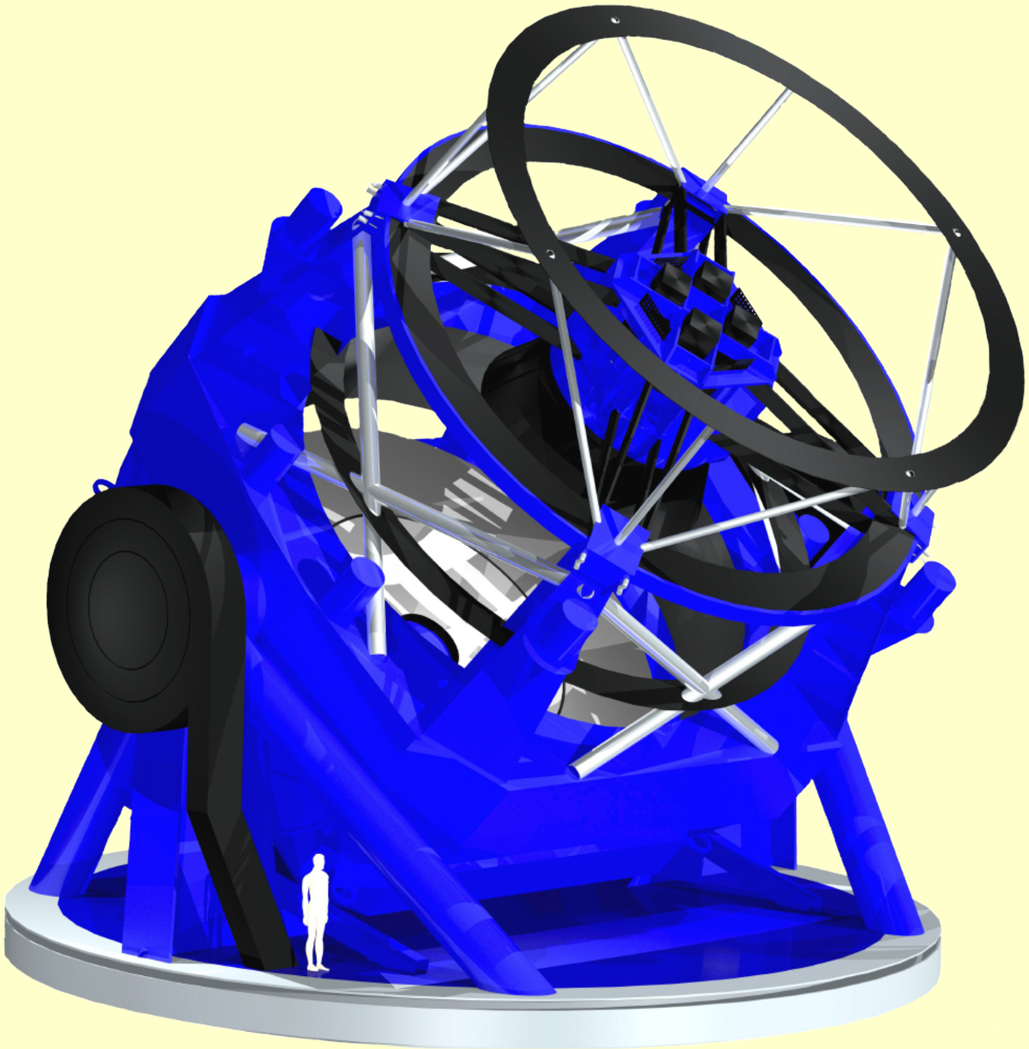
- LSST (2018 ? wide-field project in Chile) :

  180 images/hour * 6.4 Gb * 8 hours/night * 300 nights/y * 10 years

  → 27 Pb

LSST Concept

8-m class telescope dedicated to wide-field imaging in the visible



- 8.4 Meter Primary Aperture
- 10 sq degrees Field Of View
- 3.2 Gigapixel Camera
  - ~200x (4k x 4k) CCDs
  - Six Filters
- 1 image every 20 s.
- Raw images : 13 TB/night
- Covers the whole visible sky every 3 nights
- 10 years of operation
- To be built in Chile
- Funding essentially secured (mostly US)
- First light expected in 2018.

http://www.lsst.org

LSST science mission

- Dark Energy and the accelerating universe
- Map of the Milky Way
- Comprehensive census of Solar System objects
- .... and the unknown

Wide survey :

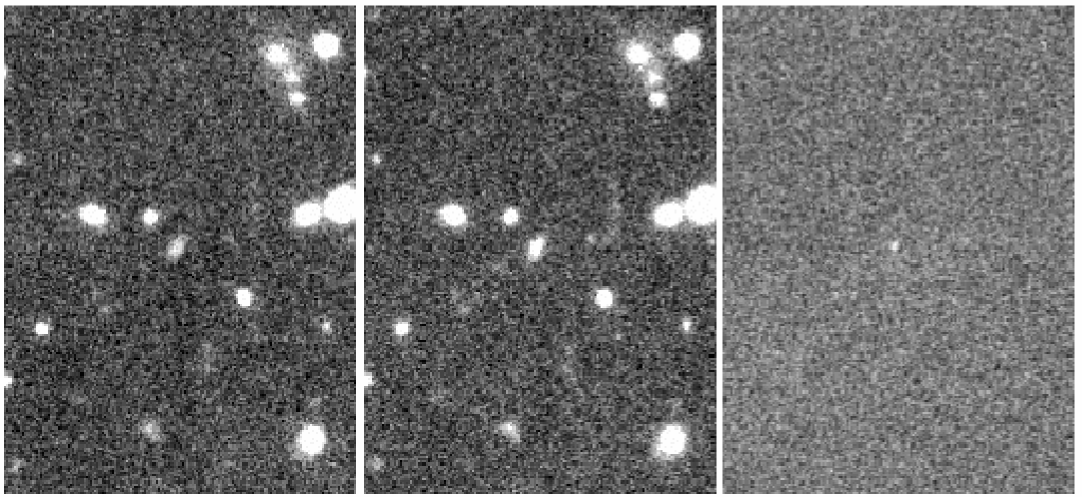
Whole (southern) sky in 6 bands with ~ 400 visits per band

Deep survey :

Repeated imaging (few days) of a few pointings, in 6 bands.

Key science data products :

- Alerts for transients (difference imaging)
- Image stacks
- Catalogs :
  - Static sky
  - Transients
  - Moving objects



Pierre Astier (30/05/2011 Aspera)

Wide survey characteristics

6-band Survey: *ugrizy* 320–1050 nm

- Sky area covered: 20,000 deg^2 0.2 arcsec / pixel
- Each 9.6 sq.deg FOV revisited >300 times/band
- Time resolution: >20 sec
- Limiting magnitude: 26.5 AB magnitude @10σ (24.5 in u)
  24 AB mag in 15 seconds
- Photometry precision: 0.01 mag requirement, 0.005 mag goal
- Galaxy density: 50 galaxies/sq.arcmin
- 3 billion galaxies with color redshifts
- Time domain: Log sampling, seconds – years

Sociology and data releases

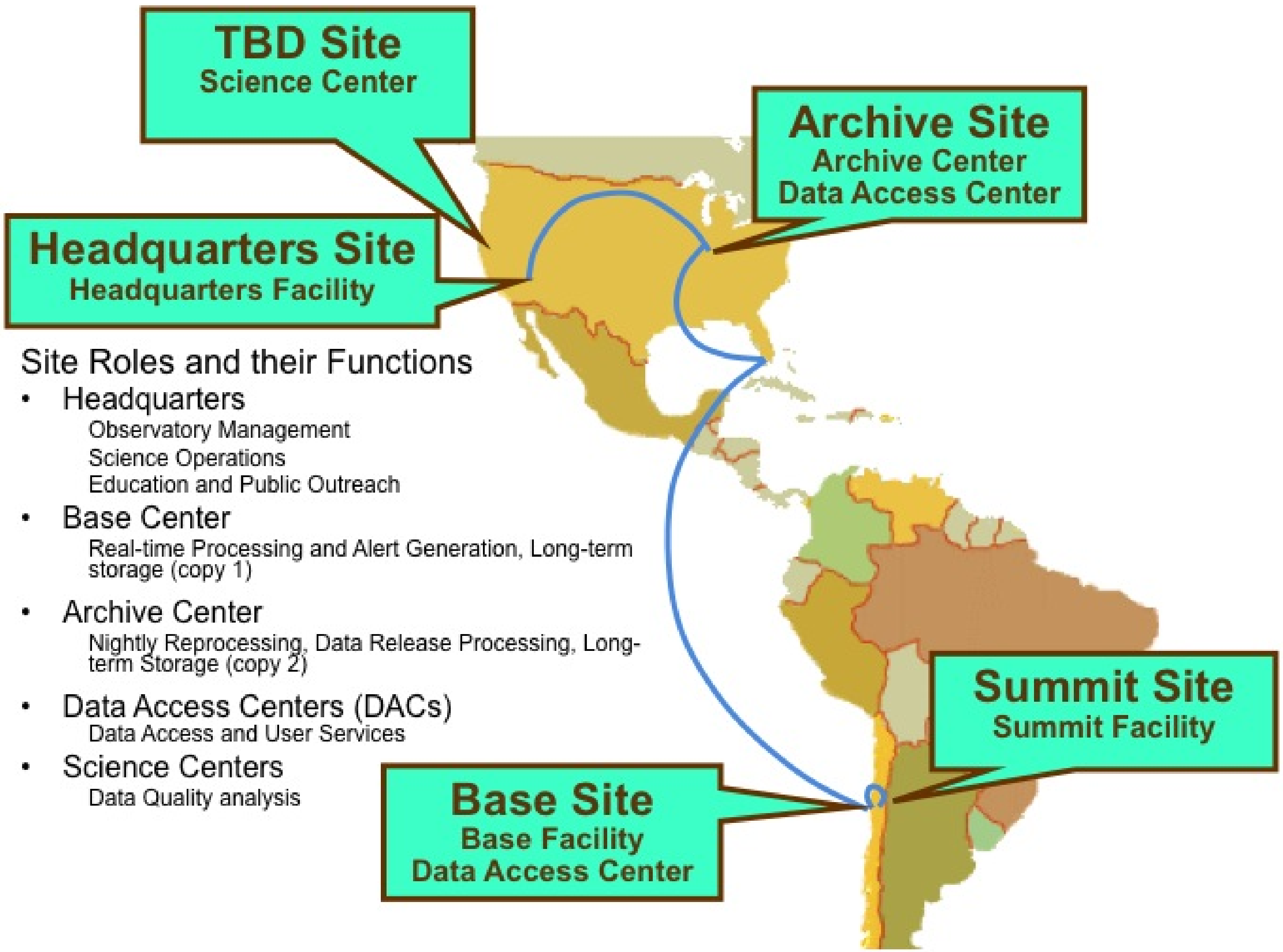
- LSST <u>has to</u> deliver a yearly (processed) data release to the public.
- The policy is the same as for SDSS (aka " The Sloan")
- The SDSS is nowadays the first source of citations for astrophysics publications.

- The "Data Management System" for LSST is firstly designed to meet this yearly data release goal....
- .... and the real time alerts for transients.

Data Management System Functions

- **Process image stream from camera to generate real-time transient alerts**
  - Difference image based
- **Periodically process entire set of survey data to produce a Data Release**
  - Self consistent set of data products, all processed with the same algorithms
  - Full survey depth; meets SRD requirements
- **Periodically produce calibration data products needed by other pipelines**
- **Make data available to scientists, with enough processing cycles and support to make it useful**

Data Management Sites and Functions



Site Roles and their Functions

- Headquarters
  - Observatory Management
  - Science Operations
  - Education and Public Outreach
- Base Center
  - Real-time Processing and Alert Generation, Long-term storage (copy 1)
- Archive Center
  - Nightly Reprocessing, Data Release Processing, Long-term Storage (copy 2)
- Data Access Centers (DACs)
  - Data Access and User Services
- Science Centers
  - Data Quality analysis

DMS Performance Requirements

Real-time alert latency	60 seconds
Nightly data generation rate: Raw pixel data from camera (24 hrs) Image through pipelines Archived images + metadata Catalogs (transient phenomena)	 15 TB (16 bit, science + calibration) 30 TB (32 bits) + 108 TB (32 bit) intermediate images 15 + 1 TB (32 bits compressed to 16 bits) 2 TB (32 bit compressed to 16 bits)
Catalog volume (average per release): Source Catalog Deep Object Catalog	 1.8 PB 0.1 PB
Yearly data archive rate (average): Images Catalogs Metadata	 10.6 PB 1.9 PB 1.9 PB
Total digital storage Summit / Telescope site La Serena Base Facility (Catalogs) Archive Center (Catalogs) Archive Center (Images) Archive Center Cache and Spare Data Access Centers (replication) Data Access Centers (end user space)	 100 TBytes (4 nights + spare capacity, fixed over 10 yrs) 106 PBytes (full image backup, 4 nights + spare capacity, 10 yrs) 38 PBytes (1 catalog release per year (2 in yr 1) over 10 yrs w/indices) 78 PBytes (total image archive) 5 PBytes (total over 10 yrs) 6 PBytes (total over 10 yrs) 12 PBytes (total over 10 yrs)
Nominal computational req'mnt At telescope site At base site At archive center At data access centers for users	 <1 TFlops 37 TFlops 100 TFlops (yr 1); 290 TFlops (yr 10) 57 TFlops (total all DACs)
Communications Bandwidth Telescope to base site Base site to archive Archive to Data Access Centers Data Access Centers to end users	 40 Gbits / sec 2.5 Gbits/sec avg, 10 Gbits/sec burst 10 Gbits/sec (total) 16 Gbits/sec (total)

J. Kantor et al, SPIE 7740-60 (2010)

LSST : storage and CPU

		Year 1	Year 10
Storage :	Images	~10 Pb	78 Pb (disks + tapes)
	Catalogs	~ 5 Pb	38 Pb (DB with indices)
(>= 2 copies)			

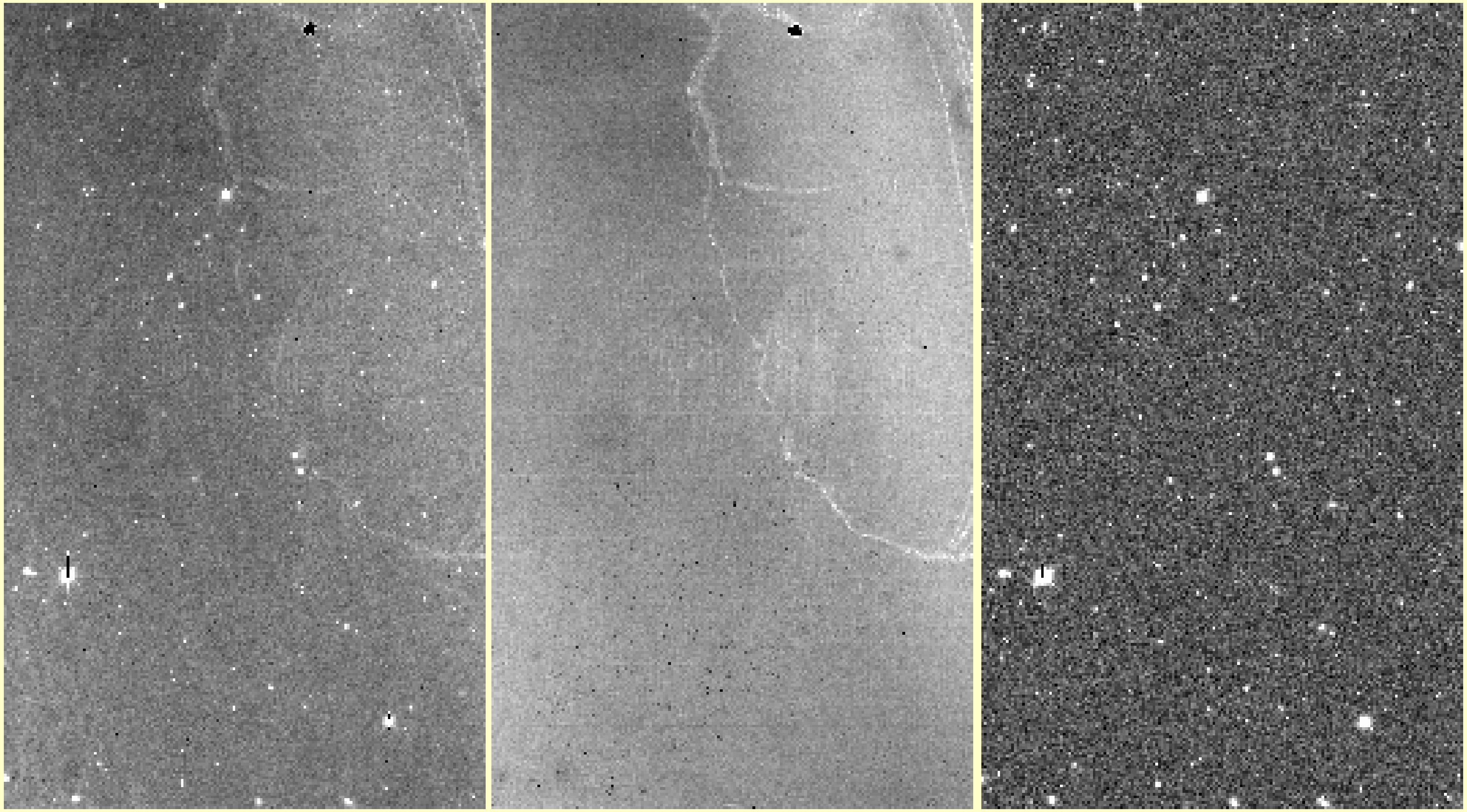
CPU : Transients (real time) ~30 TFlops

	Releases	100 TFlops	300 TFlops

(mainly image stacking + DB access,
stacking CPU increases with # images !)

Typical image processing : flatfielding

Raw image Flat Flat-fielded

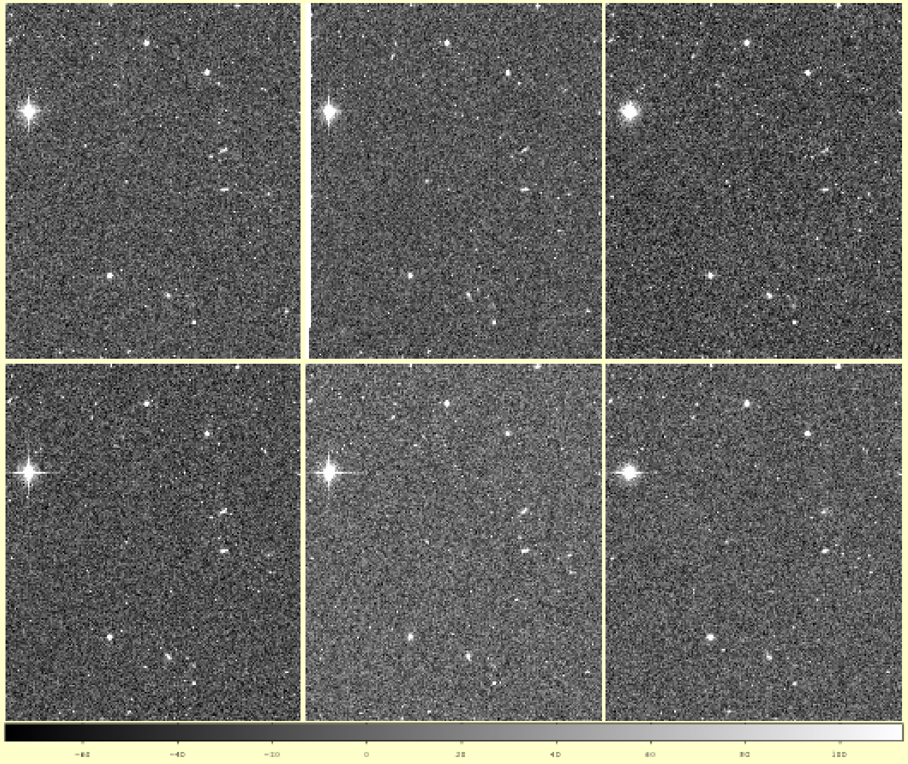


I/Os : ~ 60 Mb, CPU : ~ 0.5 s

Image Processing : stacking (1)

Successive exposures of the same field

From object catalogs
→ pixel-to-sky mapping
→resample to the same pixel grid



(Real data from Megacam@CFHT)

Image processing : stacking(2)

Combine (co-add) aligned pixels to produce the stack (same processing for weights)

~600 images of 300 s

300s integration

Resource budget for stacking

Example : stack 700 exposures on 1 square degree at (0.2”/pixel).

Resampling (for 10 Mpixels): 2s CPU 20 Mb input 80 Mb output
Resampling “intensity” = ~50 Mb/s (I/O's per CPU s)

Resampled image volume : 2 Tb /square degree of sky

Co-adding : CPU is ~ 20 % of resampling. I/O's are the same.
Co-adding “intensity” = ~250 Mb/ (CPU) s

For wide field image processing,
I/O efficiency now drives hardware and software.

LSST : Data Products

Processing Cadence	Image Category (files)	Catalog Category (database)	Alert Category (database)
Nightly	Raw science image Calibrated science image Subtracted science image Noise image Sky image Data quality analysis	Source catalog (from difference images) Object catalog (from difference images) Orbit catalog Data quality analysis	Transient alert (every 60 s) Moving object alert Data quality analysis
Data Release (Annual)	Stacked science image Template image Calibration image RGB JPEG Images Data quality analysis	Source catalog (from calibrated science images) Object catalog (optimally measured properties) Data quality analysis	Alert statistics & summaries Data quality analysis

LSST storage plans : what and where

Images :

- Raw images, calibration frames, and stacks are resident on disk (and indexed in DB).
- Image metadata in image headers (and DB ?)
- Everything else (mainly calibrated images) is generated “on demand”.

Catalogs:

- All in DB, with “geographic segmentation”.

Focus on Dark Energy measurements

- Weak lensing of galaxies.

  Two and three-point shear correlations in linear and non-linear gravitational regimes.
  → requires a measurement of the ellipticity of galaxies (and stars)

- Supernovae to z = 1 or more.

  High statistics Hubble diagram
  → requires the measurement of light curves

- Galaxies and cluster number densities as function of z.

  Can be done from catalogues (at first order)

- Baryon acoustic oscillations (BAOs).

  Can be done from catalogues (at first order)

Resampling images?

The current way to cosmic shear (e.g. Fu et al 2008) :

- Collect enough images to reach the required depth.
- Resample them to a common pixel grid
- Coadd aligned images
- Measure ellipticities of galaxies and stars on the stack.
- ....

This is not optimal :

- Resampling introduces correlations between neighbour pixels, which are not (practically) tractable.
- Resampling alters the shape of objects !
- Images with the best image quality are buried among mediocre exposures.
- Linearity issue : saturation affects bright stars in best images only.

Can we avoid resampling ?

YES !

- Still use stacks to find galaxies and get their rough position
- Simultaneous fit of ellipticity and position in a (large) set of images

→ involves the simultaneous (random) access to hundreds of images
→ has to be very efficient (there are typically 300 000 galaxies/sq. degree)
→ Interesting challenge

This “MultiFit” concept appears as the baseline in LSST TDR.

Resampling-free measurements : practical implementations

- Shape measurements ?
  I am not aware of any.

- Transients flux measurements? Yes : for supernovae
  - "Scene modeling" in SDSS-II supernovae survey (Holtzman et al, 2010)
  - PhD of Nicolas Fourmanoit (2010)
    → about 1h for per object over ~500 images
    → it take minutes just to load the image stamps and get ready to fit.

For "simultaneous fits" to hundreds of images, we seem to be far from the needed computational efficiency.

Summary

- I/O efficiency is probably the key issue for wide-field imaging data reduction systems (for large surveys).
- Lossless image compression helps at reducing the load.

- LSST volumes pose an interesting challenge for both hardware and software.

- Agencies impose a “general purpose” data reduction of large public surveys both for immediate use and legacy. (SDSS, VST@ESO, LSST, ...)

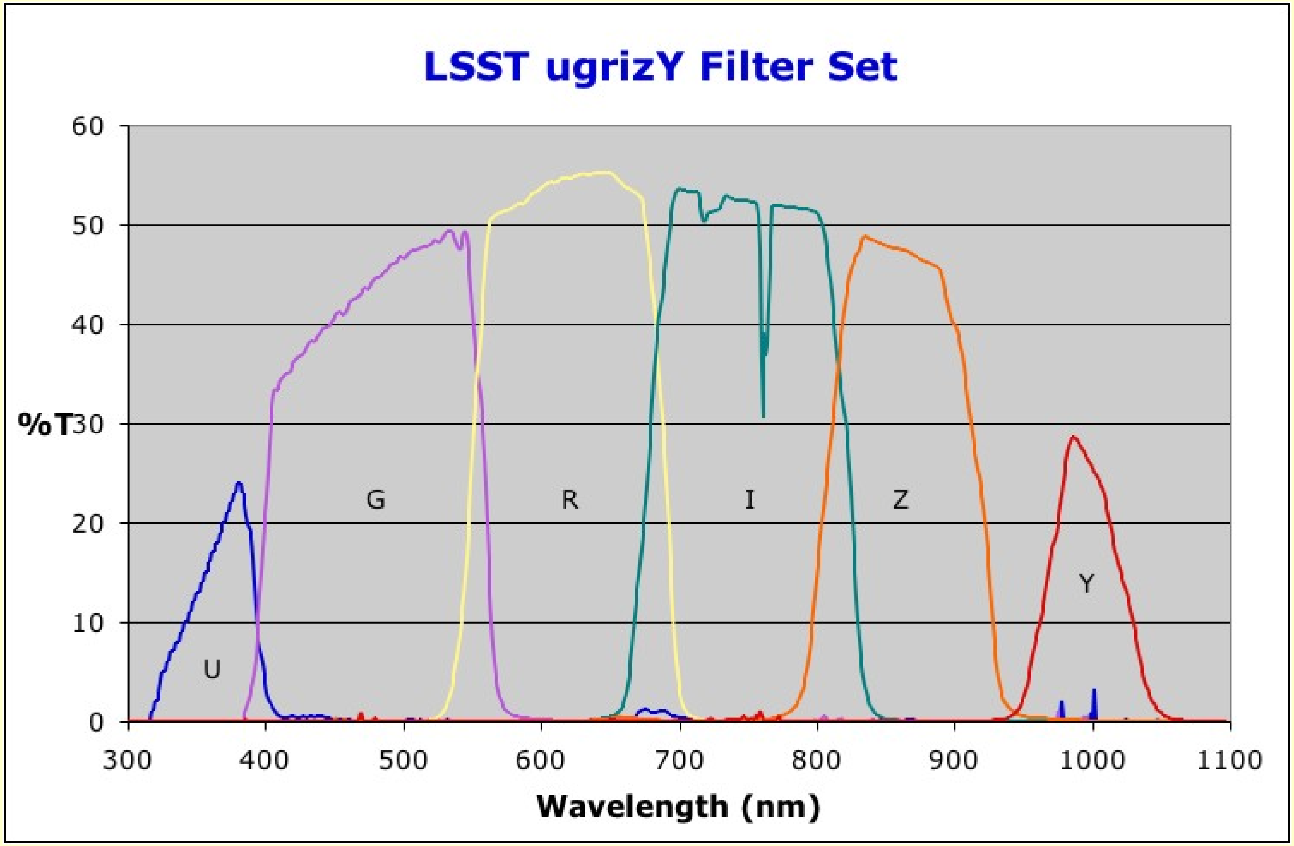
- Most of the fore-front analyses have anyway to carry out their own processing.

More Slides

Massively Parallel Astrophysics

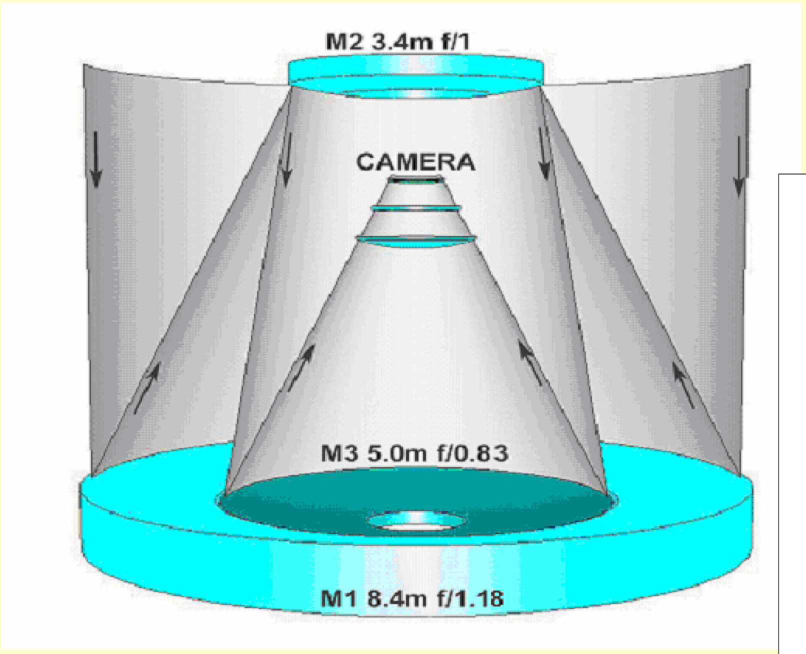
- Dark matter/dark energy via weak lensing
- Dark matter/dark energy via supernovae
- Dark Energy via Baryon Acoustic Oscillations
- Galactic Structure encompassing local group
- Dense astrometry over 20000 sq.deg: rare moving objects
- Gamma Ray Bursts and transients to high redshift
- Gravitational micro-lensing
- Strong galaxy & cluster lensing: physics of dark matter
- Multi-image lensed SN time delays: separate test of cosmology
- Variable stars/galaxies: black hole accretion
- QSO time delays vs z: independent test of dark energy
- Optical bursters to 25 mag: the unknown
- 6-band 27 mag photometric survey
- Solar System Probes: Earth-crossing asteroids, Comets
- Extragalactic stars

LSST filter set



System optical throughput analysis
Meets filter complement, performance requirements
Meets image depth, image quality requirements

Telescope Optics (relevant for IQ)



Paul-Baker Three-Mirror Optics

8.4 meter primary aperture.

3.5° FOV with f/1.23 beam and 0.20" plate scale.

PSF controlled over full FOV.