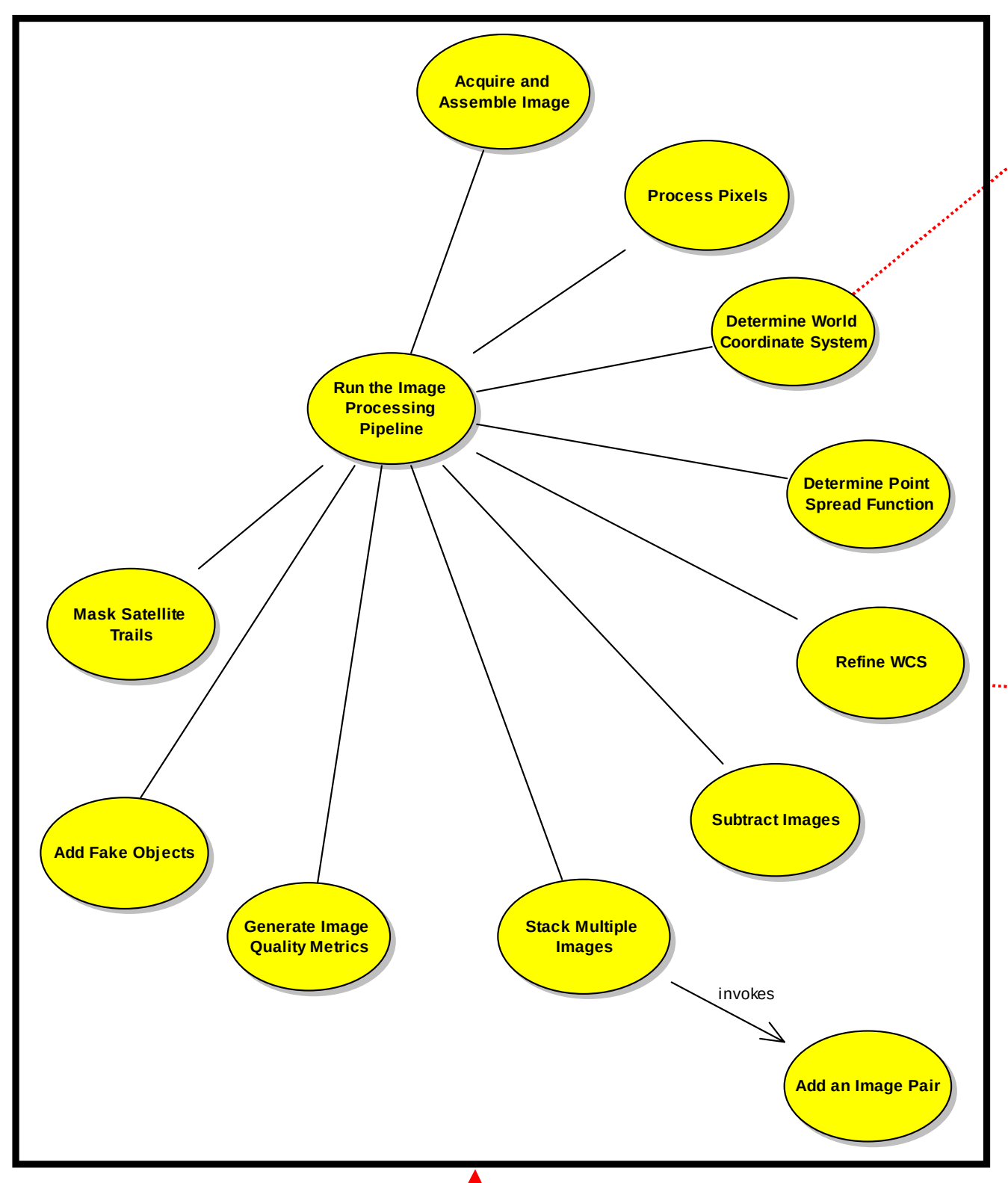
Large Synoptic Survey Telescope

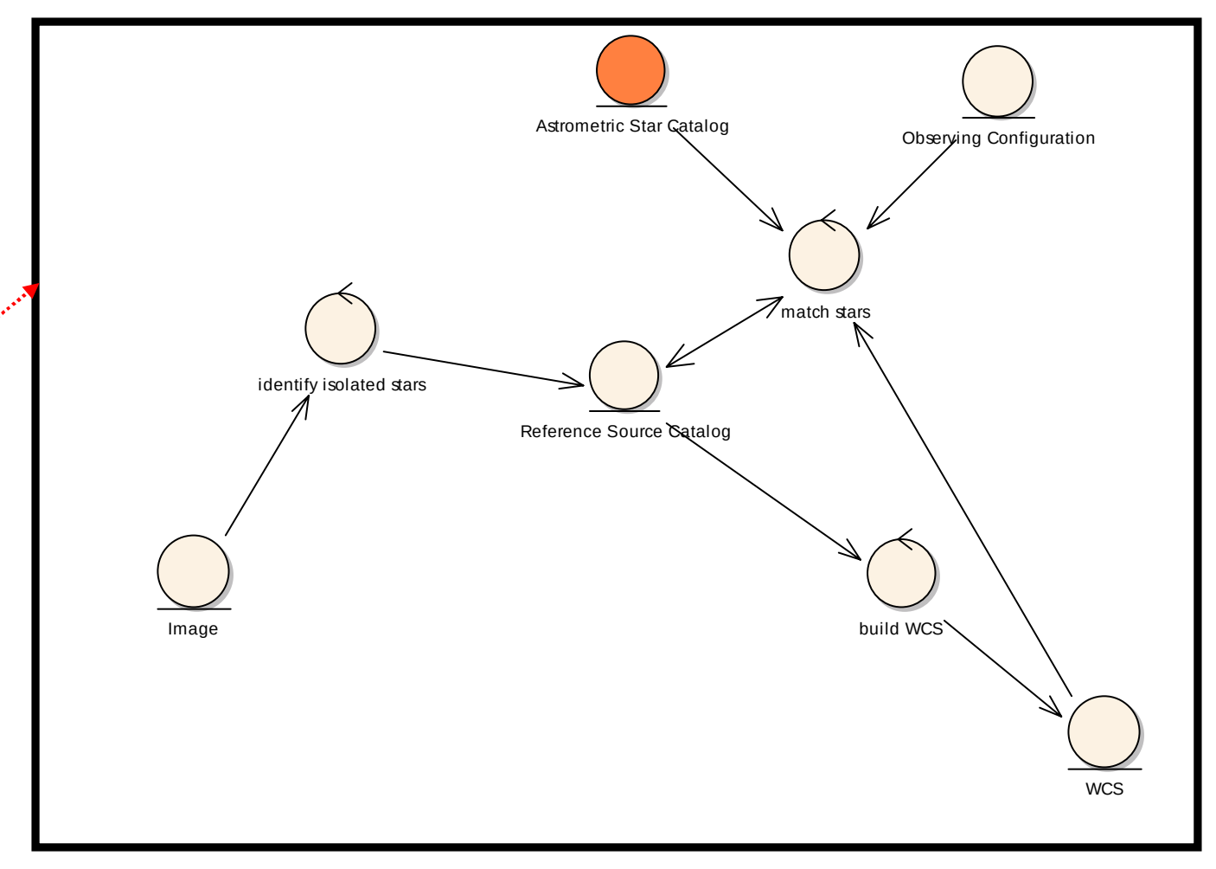
The LSST Data Processing Pipeline

T. Axelrod (Steward), A. Connolly (U. Pitt.), Z. Ivezic (U. Wash.), J. Kantor (LSSTC), R. Lupton (Princeton), R. Plante (NCSA), C. Stubbs (Harvard), D. Wittman (UCD)

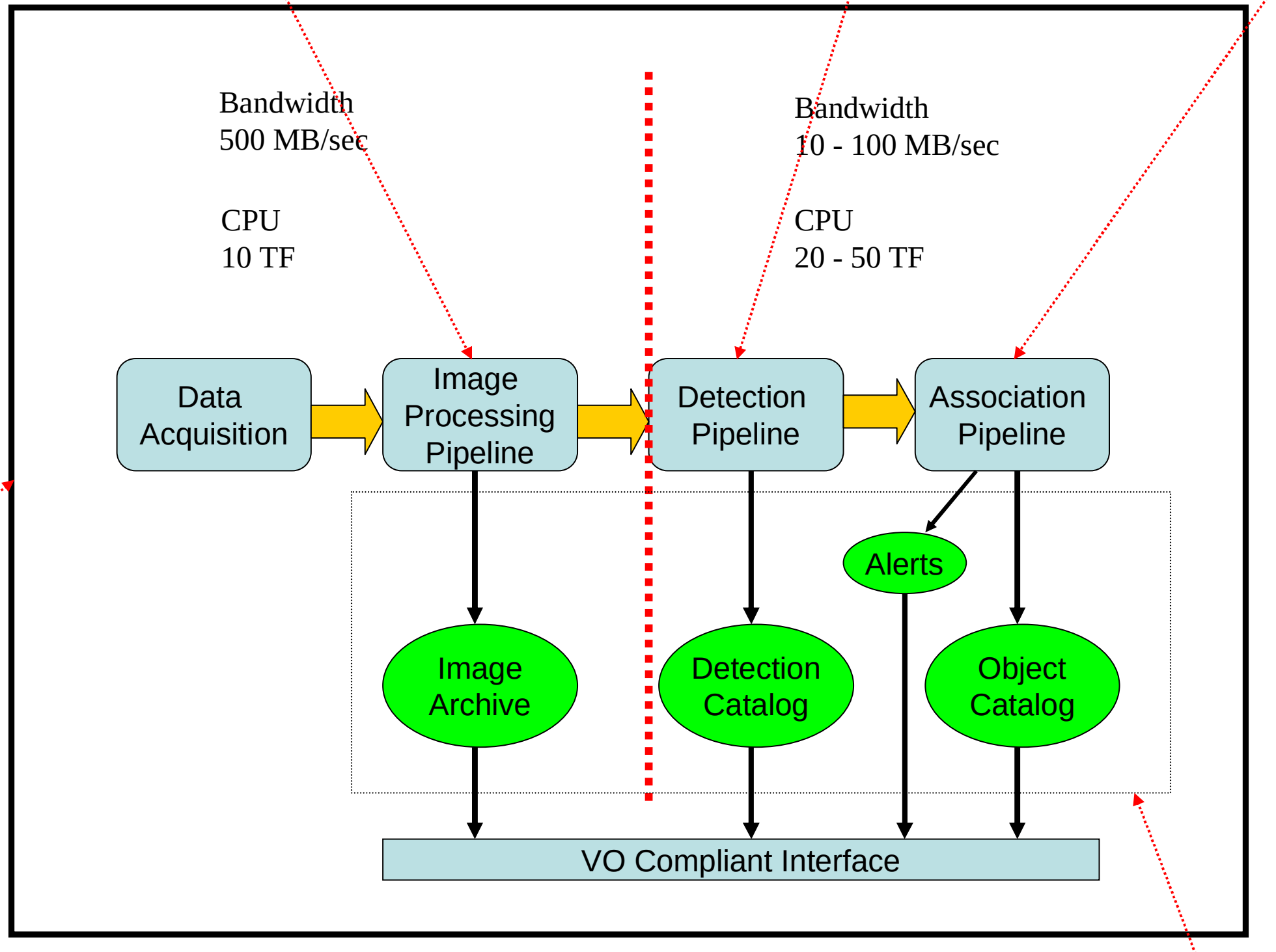
In science observing mode, the LSST telescope and camera system will deliver a 3.2 Gpixel image every 12 sec, a data rate of about 0.5 GByte/sec. The data processing pipeline must process this incoming data to produce the LSST's primary data products: the calibrated image archive; the detection catalog; the object catalog; and real-time alerts. Additionally, the quality of the incoming data must be rapidly assessed and fed back to the observatory control system. The pipeline must be flexible enough to allow addition of new processing stages and replacement of existing algorithms with improved ones, and must be robust in the face of the failure of hardware components such as disk drives. We present a preliminary design of the primary LSST data products, and of the data processing pipeline. The mapping of the pipeline onto computing hardware is discussed, along with estimates of the computational, I/O, and network bandwidths required.

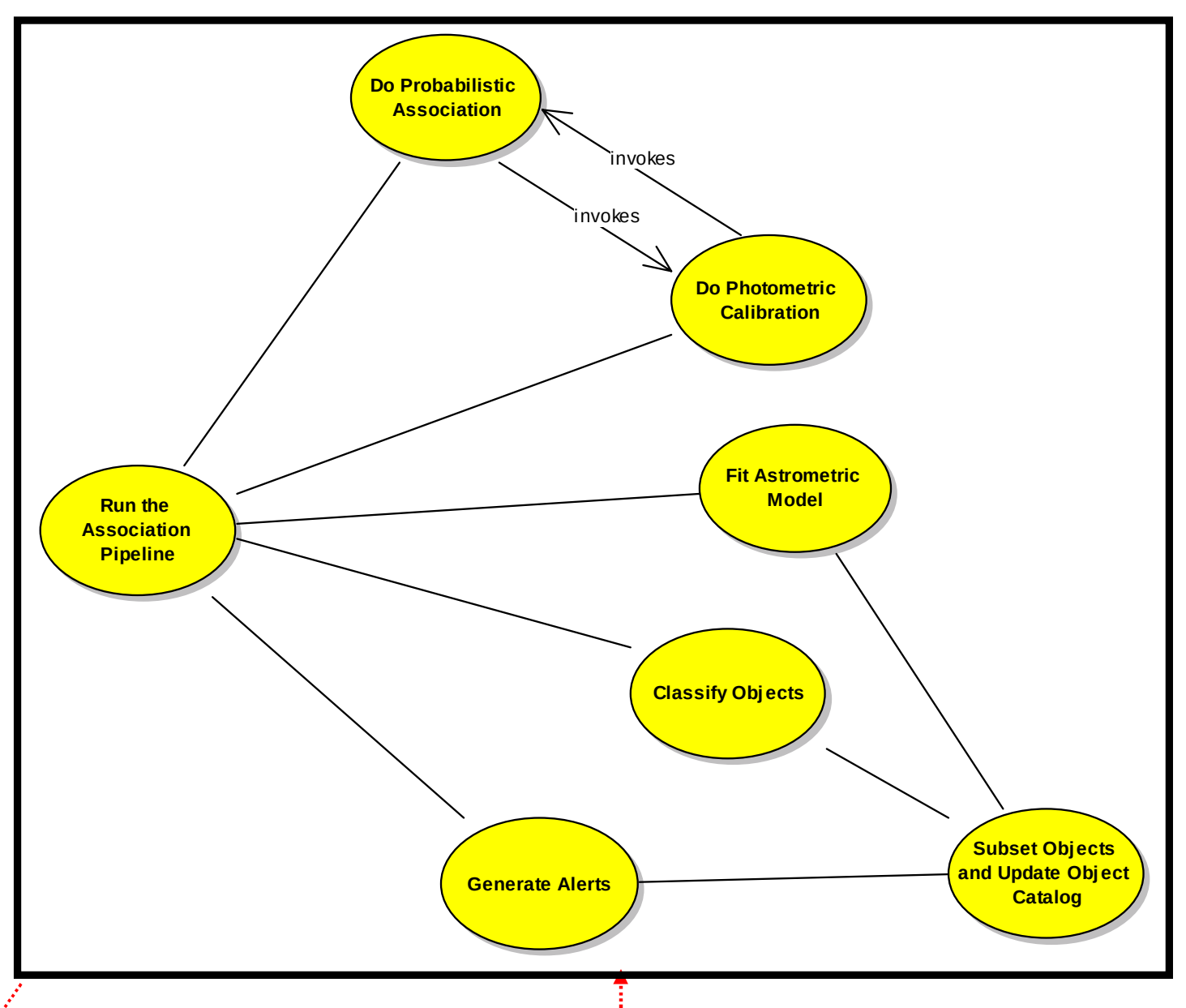

Acquire and Assemble Image
Process Pixels
Determine World Coordinate System
Run the Image Processing Pipeline
Determine Point Spread Function
Refine WCS
Mask Satellite Trails
Subtract Images
Add Fake Objects
Generate Image Quality Metrics
Stack Multiple Images
invokes
Add an Image Pair

The Image Processing Pipeline (IPP) is broadly responsible for taking raw images from the Camera and producing calibrated science images. The UML Usecase diagram (fig. 2) gives an idea of its scope. In addition to the usual crosstalk correct/debias/flatten procedures ("Process Pixels"), the PSF and WCS must be determined. This is driven by an astrometric and photometric reference catalog of stars. The IPP is also responsible for injecting fake objects for purposes of efficiency determination, and for detecting and masking satellite trails. When required, the IPP also produces subtracted and/or coadded images. The IPP also has the important function of assessing the quality of each image, and making this information available to the Observatory Control System (OCS). Quality parameters will cover camera performance, PSF quality, and atmospheric transparency



The Detection Pipeline (DP) is responsible for detecting sources in an image, which can be one of several types, including single science images, subtracted images and coadded images. Every detected source yields a new entry in the Detection Catalog containing all parameters measured for the source. These include position, magnitude, and a variety of shape parameters. The DP will normally run on images in which sources matching objects in the reference catalog have already been identified by the IPP. These sources are all bright, relatively isolated, and small in number. The DP's task is to go well beyond this, detecting all sources exceeding a specified signal-to-noise floor. Since nearly all LSST sources will be well below sky levels, achieving the best possible sky subtraction is a priority for the DP.




Do Probabilistic Association
invokes
invokes
Do Photometric Calibration
Run the Association Pipeline
Fit Astrometric Model
Classify Objects
Generate Alerts
Subset Objects and Update Object Catalog

The Association Pipeline (AP) is responsible for maintaining the Object Catalog (OCat) . The Detection Catalog (DCat) contains information about only a single epoch and filter, and may be dominated by false detections generated by noise. In contrast, the OCat is intended to contain dominantly information about real astronomical objects, and includes information from all relevant epochs and filters. Because many objects are moving, have time variable brightness, or both, the task of the AP is quite complex. In general, the association of detections from different times and spatial positions can be made only probabilistically, and it may be necessary to carry in the Object Catalog more than one possible association for a detection. To avoid exponential growth in size of the OCat, suitable pruning algorithms must be developed. This is an active area of research.
The parameters stored in the OCat for an object include both best estimates for time average properties, such as median magnitudes and colors, shape parameters, and astrometric parameters (or for solar system objects, orbital parameters), and the time dependent light curves in all available bands

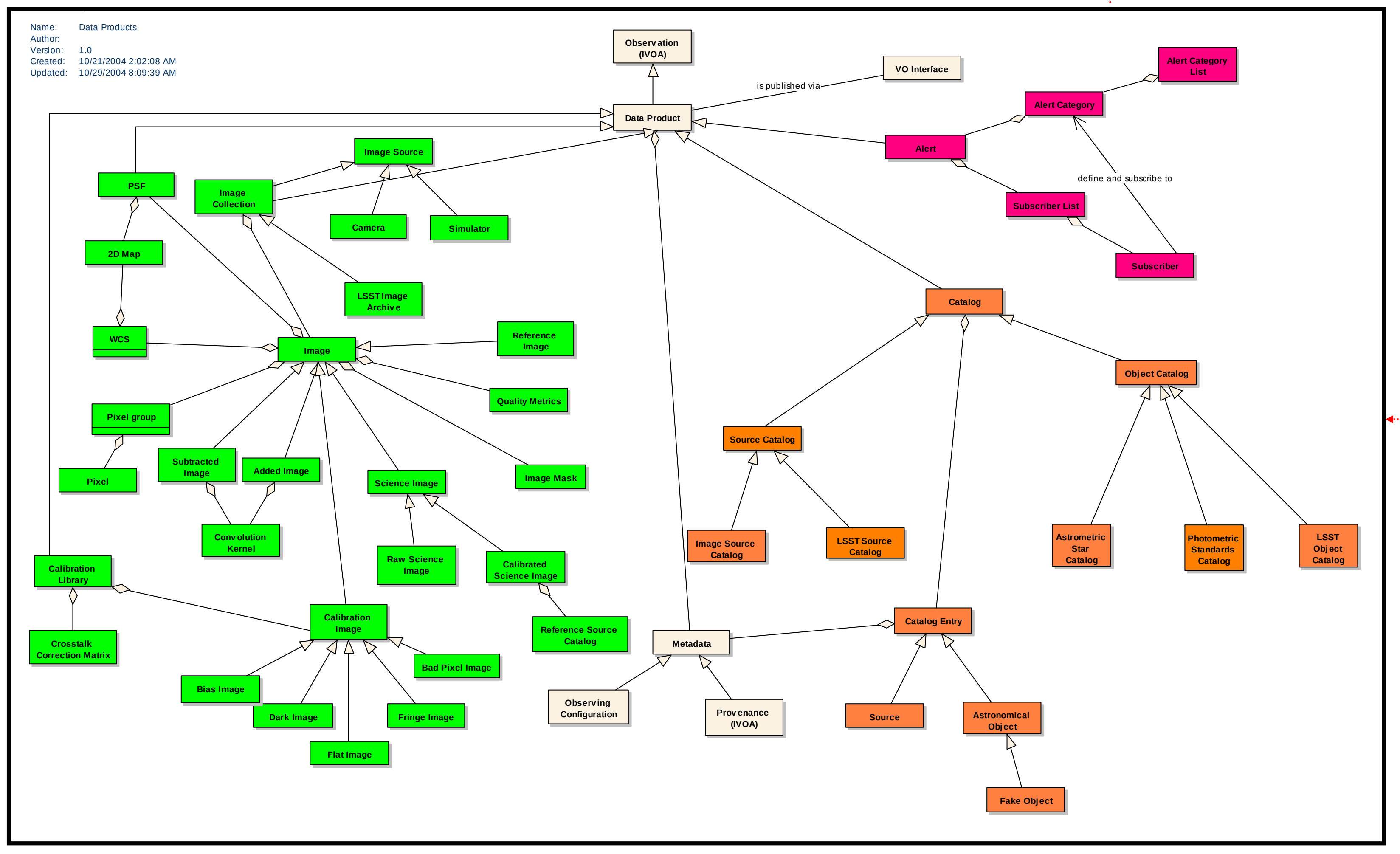
Computational Requirements

Two broad requirements determine the computational capacity required by the LSST data pipeline:

- The pipeline must not fall behind in processing observations during a night. This requires completely processing during 24 hours a set of images that arrive every 12 sec for at least 10 hours (3000 images / 24 hrs).
- The pipeline must have sufficient excess capacity to support reprocessing of past data. This one is difficult to assess in advance of operations, but experience with other surveys suggests that reprocessing will grow to perhaps 2x the nightly throughput.
- Rough throughput estimates are shown in Fig. 1

These determine the required throughput of the pipeline. There is an additional latency requirement, set by the need to generate alerts promptly in response to interesting transient events.

- The LSST pipeline can be highly parallel, applying large numbers of CPU's to many tasks.
- The throughput, P, and latency, t, are approximately related by P = N / t, where N is the number of CPU's. If parallelism is applied at the highest possible level (each observation is processed by its own cpu), the required throughput could be met by (for example) 1000 cpus, each taking t=8 hrs.
- The science requirements for t are still to be firmed up, but are likely to be closer to t=10 min. This implies that parallelism must be found within the processing of a single observation, for example at the CCD level.



Data Products

LSST will produce four classes of publicly available data products:

- Image Archive
- Detection Catalog
- Object Catalog
- Alerts

The relationships between these products, and some of their inner structure, are shown in Fig. 2

- The Image Archive contains highly processed images, such as image stacks and subtracted images, as well as raw and calibrated images.
- The Detection Catalog contains an entry for every detection generated by the Detection Pipeline. Depending on pipeline parameters such as the detection threshold, the Detection Catalog may contain large numbers of false detections.
- The Object Catalog is intended to contain information about real astronomical objects, including their time dependent brightnesses in the various filter bands, motion on the sky, shape parameters, and classification. It is likely to be the most heavily used, and most computationally demanding, of the data products.
- Alerts are notifications sent to subscribers of events that meet predefined criteria. Because the interests, and capacity to process alerts, vary widely among potential subscribers, we are committed to allowing alerts to be flexibly defined on a per-subscriber basis.

All publicly available LSST data products will be made available through a VO-compliant interface. Many aspects of this interface are still to be worked out in cooperation with the VO community.