EVLA Data Processing PDR

Overview

Tim Cornwell, NRAO
EVLA: Data Management

- EVLA has sub-contracted EVLA data management to NRAO Data Management group
- End-to-end processing needs being addressed by DM End-to-end (e2e) project
- Data reduction needs being addressed by DM AIPS++ project
Principal EVLA Subsystems
End-to-end goals

• Streamline observer access to NRAO telescopes
  – End to end management from proposal to science
  – Cross-Observatory consistency

• Greatly improve data products to users of NRAO radio telescopes
  – Provide original, calibrated, and auxiliary data, default images and processing scripts
  – Improve monitoring of instrument behavior

• Greatly improve archive access
  – On-line access to archives of contemporary and historical images, surveys, catalogs, etc.
  – Technical and scientific data mining via web and NVO

To reach these goals, initiated End-to-end Project in July 2001
e2e requirements and scope

- Extensive discussion of *first pass* scientific requirements with Scientific Working Group
  - Captured in e2e project book:  
    http://www.nrao.edu/e2e/documents/e2eprojectbook.doc
  - Proceeding on basis of current requirements
  - Description of workflow from proposal to observing script
    - Converted to high level architecture and data flow
- Refine scientific requirements at end of phase 1 (July 2002)
- Commit to design and scope at end of phase 2 (April 2003)
  - First e2e advisory group meeting ~ April 2003
- Spending ~ 15% of budget on planning
  - Good way to mitigate against risk
e2e development

- Current staff
  - John Benson, Tim Cornwell, Boyd Waters, Honglin Ye
  - Lindsey Davis (IRAF, NOAO – to join in Sept, funded by ALMA), another later
  - Doug Tody (IRAF, NOAO – to join in Sept, part of large NSF-funded collaboration)

- Use spiral development model
  - Develop in 9 month phases
    - Get requirements, plan, design, implement, test
    - Review requirements, plan, design, implement, test…..
  - Five year development plan consisting of 7 phases
  - Add new staff incrementally

- Three important principles
  1. Keep it simple
  2. Reuse as much as possible
  3. Deliver new capabilities soon and often
The Observing System provides feedback to the Real-Time Scheduler to allow for real-time adjustments. The Real-Time Scheduler then forms a 'Queue of Control Scripts' (e.g., crd files) from a 'Scenario'. The Observing System provides feedback to the Real-Time Scheduler to allow for real-time adjustments. The Observing System may also raise events via Monitor Data.
A "Scenario" is an ordered list of Projects.

The Real-Time Scheduler produces a Queue of Control Scripts (e.g. crd files) from a Scenario.

The Observing System provides feedback to the Real-Time Scheduler by reporting the Control Scripts "as observed". The Observing System may also raise events via Monitor Data. TBD.
## Overall e2e architecture

<table>
<thead>
<tr>
<th>Package</th>
<th>How?</th>
<th>Priority</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational Model</td>
<td>Document</td>
<td>High</td>
<td>First version</td>
</tr>
<tr>
<td>Proposal Submission Toolkit</td>
<td>Web form or Java-based tool</td>
<td>Medium</td>
<td>Investigation</td>
</tr>
<tr>
<td>Proposal Management Toolkit</td>
<td>Java-based tools plus database</td>
<td>Medium</td>
<td>Investigation</td>
</tr>
<tr>
<td>Telescope Simulation Toolkit</td>
<td>AIPS++ tools</td>
<td>High</td>
<td>Deferred</td>
</tr>
<tr>
<td>Observation Evaluation Toolkit</td>
<td>AIPS++ tools</td>
<td>Medium</td>
<td>Deferred</td>
</tr>
<tr>
<td>Observation Scripting Toolkit</td>
<td>GBT Observe, GUI editor</td>
<td>High</td>
<td>Investigation</td>
</tr>
<tr>
<td>Remote Observing Toolkit</td>
<td>Java, AIPS++ tools</td>
<td>Low</td>
<td>Deferred</td>
</tr>
<tr>
<td>Observation Scheduling Toolkit</td>
<td>OMS + local adaptations</td>
<td>Low</td>
<td>Investigations</td>
</tr>
<tr>
<td>Archive Toolkit</td>
<td>AIPS++ tables + AIPS++ tools</td>
<td>High</td>
<td>Prototyping</td>
</tr>
<tr>
<td>Pipeline Toolkit</td>
<td>Production rule software, AIPS++ tools</td>
<td>High</td>
<td>Prototyping</td>
</tr>
<tr>
<td>Pipeline heuristics</td>
<td>Glish scripts as production rules</td>
<td>High</td>
<td>Prototyping</td>
</tr>
<tr>
<td>Calibration source toolkit</td>
<td>Ingres db + Java</td>
<td>High</td>
<td>In development</td>
</tr>
</tbody>
</table>

Data flow
Telescopes and projects

- e2e will be retrofitted to all NRAO telescopes (GBT, VLA, VLBA)
- VLA
  - Putting archive on-line now, working towards pipeline processing
- EVLA
  - Sub-contracted to deliver entire e2e system for EVLA (for 18 FTE-years)
  - Close interaction with EVLA project team at all levels
- VLBA
  - Will start moving archive to disk after VLA archive
  - VLBA pipeline processing once AIPS++ can handle it
- GBT
  - Designing archive facility for deployment in GBT early 2003
  - Watching re-engineering of observing script generation
- ALMA
  - Sub-contracted to develop pipeline (framework only) and post-processing
  - Start development July 2002
  - ALMA has own equivalent to all parts of e2e
  - Trying for reuse if possible (e.g. Observation Scripting GUI from ALMA)
From NRAO to the National Virtual Observatory

- Produce images and catalogs from well-documented pipeline processing
- Images and catalogs available via NVO access tools
- All radio data stays within NRAO
- Other wavebands have similar relationships to NVO
- Built using web services and grid computing
ALMA has subcontracted development of offline processing and pipeline framework to NRAO

**e2e:**
- Must deliver pipeline framework
- No other re-use planned
- Proposal submission, observation scripting will be different

**AIPS++:**
- ALMA processing requirements documents being finalized
- AIPS++ in baseline plan
- AIPS++/ALMA tests under way to test compatibility
- ALMA representative (Gianni Raffi) recently joined AIPS++ Executive Committee
e2e timescales

- Customer requirements
  - ALMA development, Phase II starts this year, runs to 2006
  - GBT archive facility by end of proprietary period (early 2003)
  - NSF funding for archive work Sept 2001 – Sept 2003
  - Project book (http://www.nrao.edu/e2e) contains scientific requirements as currently understood

- First cycle of development (ended July 15, 2002)
  - Prototyped VLA archive and pipeline software and facility
  - Started loading VLA archive to disk
  - Improved support for VLA/VLBA calibrator database
  - Design for proposal submission and management

- Second cycle of development (ends in Q2 2003)
  - GBT archive facility
  - Thorough testing of archive and pipeline for VLA
  - Development of prototype observation scripting and scheduling
  - First advisory committee meeting

- End of overall generic development (2006)
  - Working archives, pipelines, ancillary software for VLA, VLBA, GBT
  - First generation for EVLA, ALMA

- Move onto EVLA and ALMA specific development (2006+)
## EVLA critical dates

<table>
<thead>
<tr>
<th>Correlator to Archive</th>
<th>Due date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from CBE</td>
<td>Q3 2003</td>
<td>Desirable</td>
</tr>
<tr>
<td>Test correlator prototype</td>
<td>Q4 2005</td>
<td>Desirable</td>
</tr>
<tr>
<td>Start test first correlator subset at VLA</td>
<td>Q4 2006</td>
<td>Desirable</td>
</tr>
<tr>
<td>First science with correlator subset</td>
<td>Q2 2007</td>
<td>Highly desirable</td>
</tr>
<tr>
<td>New correlator operational</td>
<td>Q1 2009</td>
<td>Required</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M&amp;C to Archive</th>
<th>Due date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchtests monitor data</td>
<td>Q1 2003</td>
<td>Desirable</td>
</tr>
<tr>
<td>Prototype system on EVLA test antenna</td>
<td>Q2 2003</td>
<td>Desirable</td>
</tr>
<tr>
<td>Start observing in transition mode</td>
<td>Q2 2004</td>
<td>Required</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scheduling to and from M&amp;C System</th>
<th>Due date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start test first correlator subset</td>
<td>Q4 2006</td>
<td>Highly desirable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post Processing</th>
<th>Due date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test first correlator subset</td>
<td>Q4 2006</td>
<td>Highly desirable</td>
</tr>
<tr>
<td>New correlator operational</td>
<td>Q1 2009</td>
<td>Required</td>
</tr>
</tbody>
</table>
Costing, schedule, deliverables, etc.

- Plan is to develop design in all e2e areas to level required to cost the project by end of development cycle 2 (April 2003)
- At that point, e2e commits to requirements, costing, schedule, deliverables
- Scope adjustments will be made at beginning of development cycles as agreed with EVLA
e2e resources

- ALMA numbers estimated by ALMA computing management
  - Seem to be in line with other ground based projects but considerably less than space based
- e2e numbers based upon straw man designs, reuse
- e2e scope will be adjusted to fit resources (~ 55 FTE-years)
- Neither constitute a detailed bottom-up derivation of resources from requirements

<table>
<thead>
<tr>
<th>Effort (FTE-years)</th>
<th>ALMA</th>
<th>e2e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposal Handling Software</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Scheduling Software</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>Pipeline</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>Data Archive</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>55</td>
</tr>
</tbody>
</table>
De-scoping options

• De-scoping occurs first within toolkits via priorities set by EVLA project
  – Potentially large de-scoping available here
• Next toolkits can be removed
• e2e is committed to provide Pipeline for ALMA
  – Pipeline requires Observation Scripting, Observation Scheduling, Archive
• Core architecture can survive removal of:
  – Telescope Simulation
  – Observation Evaluation
  – Remote Observing
• Spiral development allows these de-scopes to be made incrementally (at the beginning of each development cycle)
AIPS++ resources

• Expect roughly the same level of effort from AIPS++ on EVLA as on VLA currently

• Total effort ~ 10 FTE-years from 2003 to 2009

• Addressing EVLA-specific processing issues
EVLA-specific post processing

- Mostly well-understood and in place
  - AIPS++ package: can reduce VLA data end-to-end
  - BUT final requirements yet to be set
- EVLA-specific areas requiring more development
  - New modes of processing (next slide)
  - Very large data volumes
    - Automated flagging schemes
- Performance issues
  - Ensure that AIPS++ is efficient and fast enough (compare to AIPS)
    - AIPS++/AIPS speed ratio ~ 1 +1/-0.5 (with some outliers!)
  - Develop parallelized applications (e.g. imaging, calibration)
    - Well in progress in collaboration with NCSA
  - Develop location independent computing (a.k.a. Grid computing)
    - e.g. transparent access to archive and pipelines from remote locations
### Examples of EVLA hard processing problems

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast-slew mosaicing</td>
<td>~10ms data sampling rate. Remove sliding primary beam.</td>
</tr>
<tr>
<td>Full bandwidth synthesis</td>
<td>Deconvolve wide bandwidths while accounting for spectral index, polarization, rotation measures, opacity, etc.</td>
</tr>
<tr>
<td>Full-beam high-fidelity polarization imaging</td>
<td>Correction of time- and angle-dependent beam polarization.</td>
</tr>
<tr>
<td>High fidelity imaging</td>
<td>Image and deconvolve at ~ 10^7. Currently about ~ 100 away from this in best possible cases.</td>
</tr>
<tr>
<td>Wide-angle full-beam imaging</td>
<td>Huge images, fast data sampling rates, many imaging facets to accommodate non-coplanar baselines</td>
</tr>
<tr>
<td>RFI mitigation</td>
<td>Removal of RFI post-correlation – requires high data rates</td>
</tr>
</tbody>
</table>
## e2e status

<table>
<thead>
<tr>
<th>Package</th>
<th>Status</th>
<th>Who will present</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational Model</td>
<td>First version</td>
<td>Described in project book</td>
</tr>
<tr>
<td>Proposal Submission Toolkit</td>
<td>Design complete</td>
<td>Honglin</td>
</tr>
<tr>
<td>Proposal Management Toolkit</td>
<td>Design complete</td>
<td>Honglin</td>
</tr>
<tr>
<td>Telescope Simulation Toolkit</td>
<td>Design concept exists</td>
<td>Described in project book</td>
</tr>
<tr>
<td>Observation Evaluation Toolkit</td>
<td>Design concept exists</td>
<td>Described in project book</td>
</tr>
<tr>
<td>Observation Scripting Toolkit</td>
<td>Design concept exists</td>
<td>Boyd</td>
</tr>
<tr>
<td>Remote Observing Toolkit</td>
<td>No design yet</td>
<td>Tim</td>
</tr>
<tr>
<td>Observation Scheduling Toolkit</td>
<td>Design concept exists</td>
<td>Boyd</td>
</tr>
<tr>
<td>Archive Toolkit</td>
<td>Prototype complete</td>
<td>John</td>
</tr>
<tr>
<td>Pipeline Toolkit</td>
<td>Prototype complete</td>
<td>Tim</td>
</tr>
<tr>
<td>Pipeline heuristics</td>
<td>Prototype complete</td>
<td>Tim</td>
</tr>
<tr>
<td>Calibration source toolkit</td>
<td>First version complete</td>
<td>Honglin</td>
</tr>
</tbody>
</table>
Risks

- Creeping scope
  - Requires project discipline
  - *e.g.* scientific requirements for post-processing soon
- Lack of engagement by scientific staff
  - Work with DM Project Scientist (Dale Frail), DMSWG
- Observation scripting too hard
  - Develop incrementally
- Pipeline processing cannot be made to work for significant fraction of observations
  - Prototype on VLA: will require some changes to current practices
- Archive = Operational morass
  - Need automation and management staff soon
- Repeat of AIPS++
Lessons learned in AIPS++ project

- Software development:
  - Start new software development projects with realistic expectations
  - Control scope: initial requirements were developed without a reliable costing process
  - Management of distributed software projects is especially demanding
  - Establish firm staffing commitments
  - Continual refinement of processes important: moved to spiral development

- Package deployment:
  - Demonstrate scientific completeness: establishing threads of completeness by matching representative data to reduction scripts
  - User testing is vital: formed active, large Observatory-wide test group
  - Robustness: identifying and fixing defects as submitted
  - Performance must be regularly monitored: established benchmark suite, scheduled regular profiling, targeting known cases of poor performance
  - User interface design is very demanding: conducted one-on-one testing and group surveys
  - Documentation forms a gateway to the package: enlisted help of scientists in writing documentation
  - Training is best way to introduce new users to AIPS++: presenting tutorials to small groups

- Lessons learned applied across the Observatory, ALMA, e2e
Specific changes adopted by e2e

- **Spiral model**
  - Short development cycle
  - Deliver early and often

- **Involvement of scientists**
  - Set specifications at beginning of cycle 1
  - AOC scientists tested and advised on Calibrator Source Toolkit
  - Will review and change specifications at beginning of cycle 2
  - Dale Frail will be DM Project Scientist
  - Will be involved in pipeline development, testing of archive and proposal handling during cycle 2
  - Advisory Group meeting at end of cycle 2

- **Commit to requirements, plan, costing, schedule**
  - Design and development phase (first two cycles) ending in April 2003
  - Schedule, *etc.* then set