Overview of ngVLA Computing Concept

Jeff Kern

Atacama Large Millimeter/submillimeter Array
Karl G. Jansky Very Large Array
Robert C. Byrd Green Bank Telescope
Very Long Baseline Array

Boundary Conditions

- We know how to build and operate large geographically distributed interferometers.
  - Combined VLA, ALMA, and VLBA experience
  - PI Driven Operations model

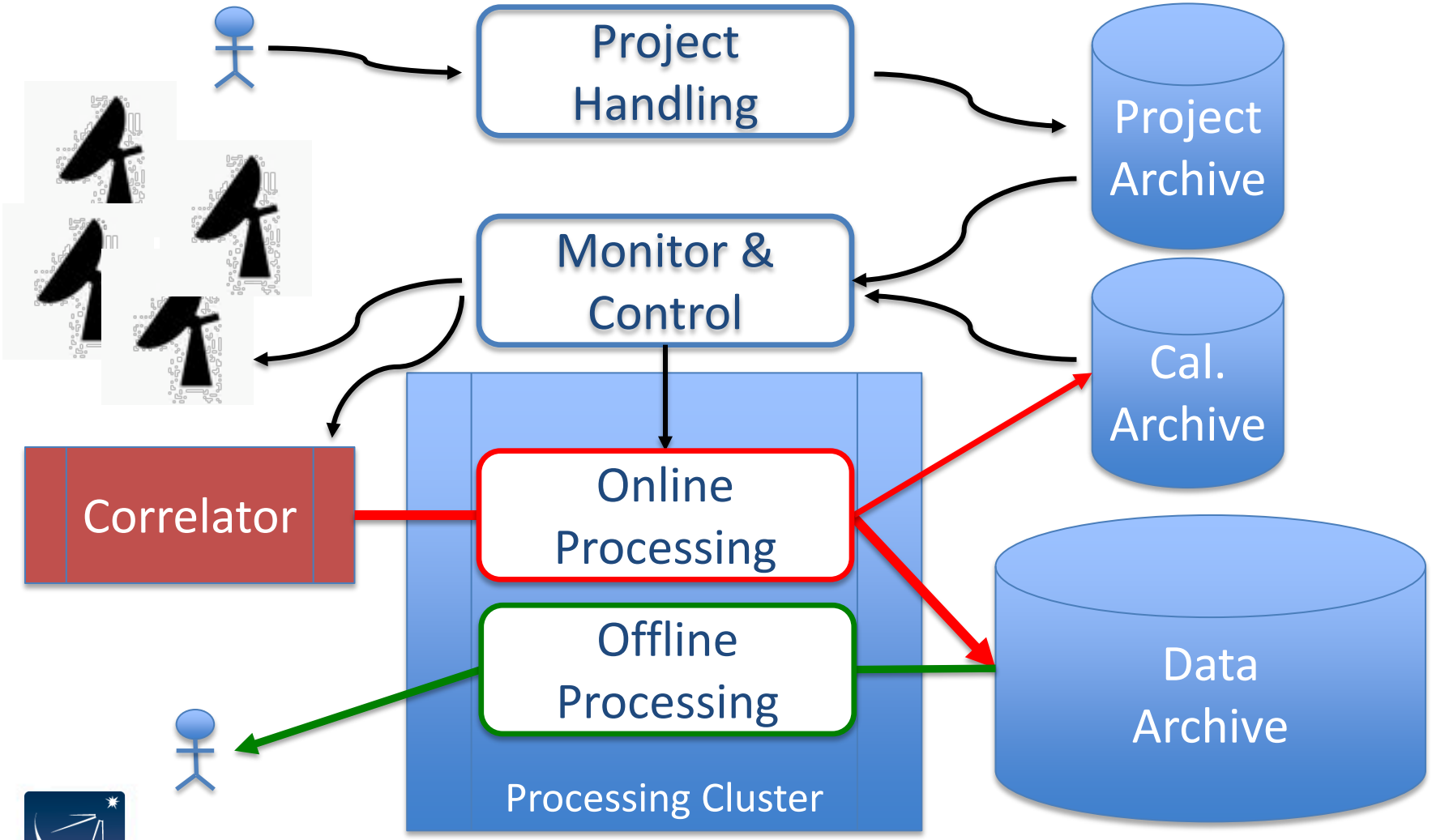
- What is unique about the ngVLA
  - Designed and implemented to minimize operations costs.
  - High Data Rate
    - Delivering raw data to the PI is not an option
    - Community will expect Science Ready Data Products (SRDP)

Some Corollaries

- Bounded Ops Cost + SRDP
  - Cannot afford army of data reducers
  - Calibration must be an observatory responsibility
    - Not possibility to rescue data taken improperly.
  - Ok to throw away good data as long as bad data does not make it to the archive

- PI Driven Operations
  - Need full suite of project lifecycle tools
  - Flexible data reduction system required
    - Imaging and calibration will not be done at home institution

Software Architecture



Project Handling

- Very similar to existing tools and processes.
  - Largest challenge will be resisting the urge to start from scratch.

- Possible collaboration / reuse options:
  - NRAO Suite (starting refurbishment now)
  - ALMA rethinking OT / connections
  - SKA TM in similar phase

Monitor and Control

Challenge is to support simultaneous maintenance and operation of hundreds of antennas.

- Key Concepts:
  - Top tier hardware systems (Antenna, Correlator,…) are stateless
    - Should be able to operate independently (Testing, Maintenance)
  - Local monitoring and fault tree analysis
    - Distribute computation away from central bottleneck
  - Data assumed to be faulty unless assertion that it is good received

Data Processing

- Key question: What is the primary archived product.

- Assumptions:
  - Sampling at Time/Bandwidth smearing limits for 95% response (β=0.5)
  - Disk cost based on AWS Glacier Price with assumed 24 month halving time, does not include data transfer fees or S3 storage for processing
  - 75% Observing efficiency
  - Assuming flat data rate.
  - Assuming single pixel feeds
- Notes:
  - Timeline for operations is single largest uncertainty (Currently using 2030)

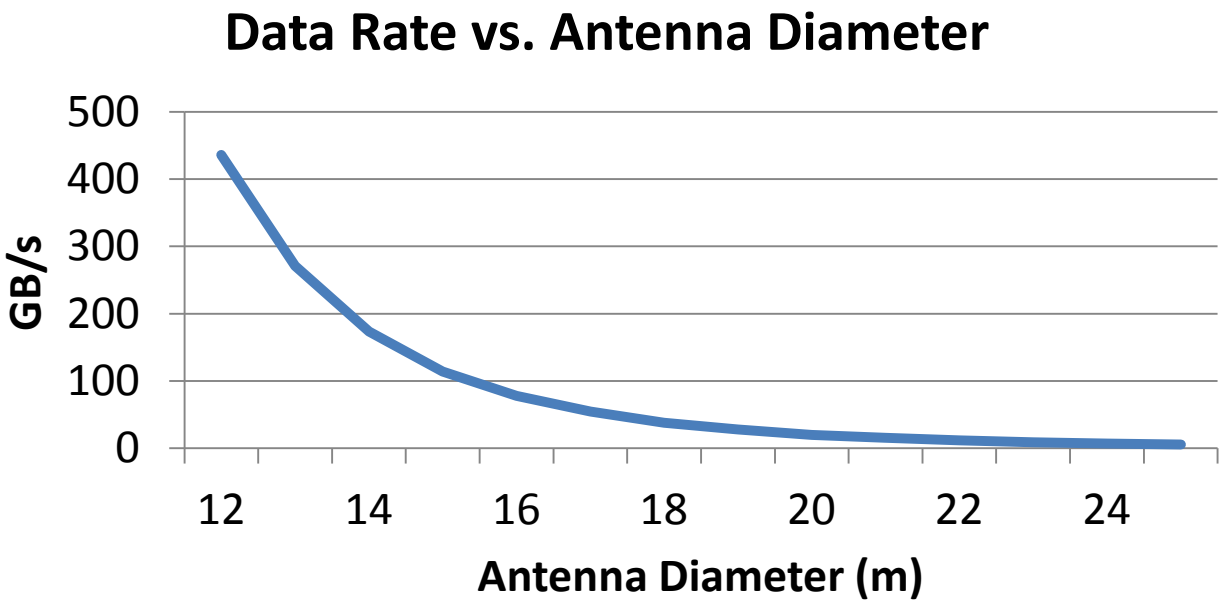
Baseline Cost Model (2030)

		2 GHZ	10 GHz	30 GHz	80 GHz	100 GHz
Band Top	115.00 GHz	3.00 GHz	14.00 GHz	40.00 GHz	95.00 GHz	115.0 GHz
Band Bottom	1.00 GHz	1.00 GHz	6.00 GHz	20.00 GHz	65.00 GHz	85.00 GHz
Resolution (mas)		140	28	9.2	3.5	2.8
FOV (arcmin)		29	5.9	2	0.6	0.51
Integration Time		0.35	0.35	0.35	0.35	0.35
Channel Width		0.03 MHz	0.15 MHz	0.51 MHz	1.67 MHz	2.18 MHz
Number of Channels:	78,000	78,000	52,000	39,000	18,000	13,765
GB/s	38.4	74.6	49.7	37.3	17.2	13.2
PB/yr	867.1	1684.3	1122.9	842.2	388.7	297.2
Storage Cost	632,841	1,229,338	819,559	614,669	283,693	216,947
Pixels per Plane:	2.39E+10	4.00E+10	2.78E+10	2.25E+10	1.51E+10	1.38E+10

Input on design decisions

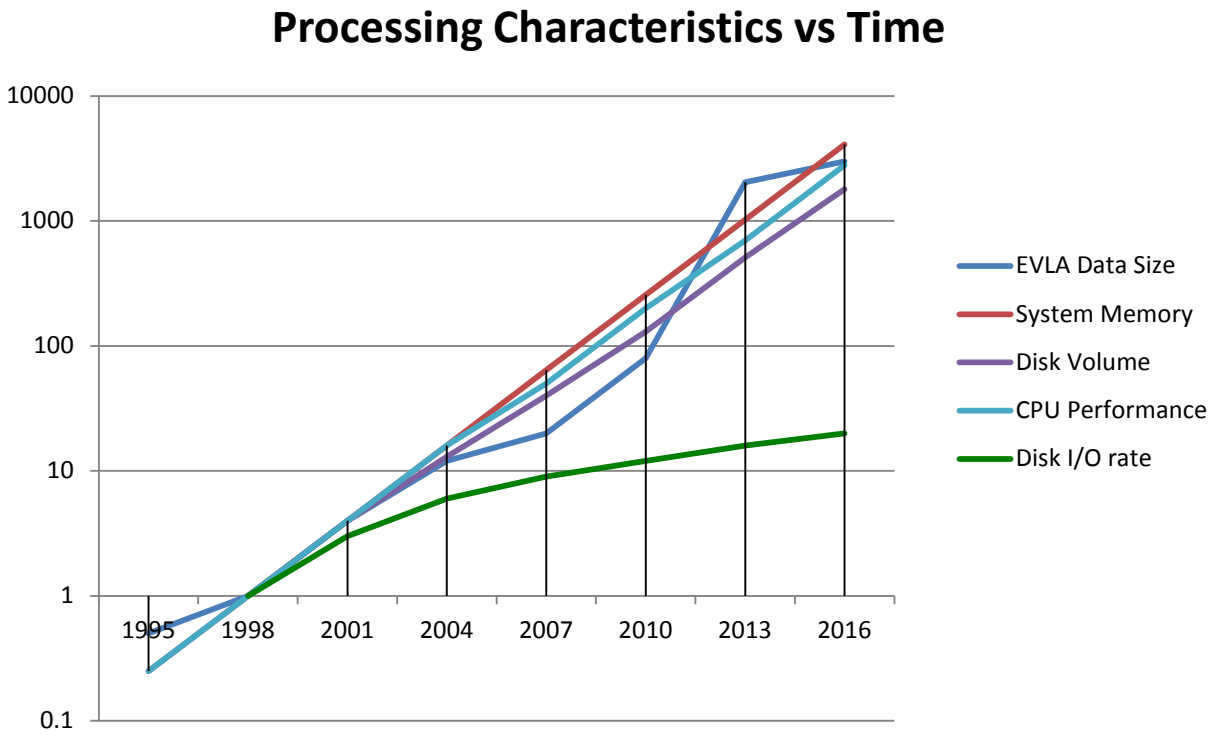
- Going to wideband feeds (3 feeds rather than 5) increases the data rates and volumes by approximately 3x.

- Data rate is also a strong function of antenna diameter
  - Assuming constant collecting area



Data Access (Disk I/O)

- Pure storage is only part of the problem, current processing techniques require multiple passes through the data
  - Unlike many other aspects of the processing equation, time alone will not solve the data access problem.



Data Processing Approach

- PI Driven approach means we need flexible data processing
  - But PI will not be able to reduce at home
  - Pipeline / SRDP required (plus PI driven reprocessing)

- Reuse CASA algorithms, rethink underlying engine
  - parallel, thread safe, etc.
  - Require CASA (or CASA like system) for commissioning

- Premium on data access
  - Process directly on object store, no "working format"
  - Separate target and calibration data, emphasis on using online telcal results
  - Algorithmic work to localize data, at all stages of processing.

Processing Resources

- ngVLA has a nominal data rate ~1000x JVLA

- Calibration scales linearly with data rate (more or less)
  - Provided no additional passes through the data
  - Moore's Law addresses most of the growth

- Imaging is more difficult
  - Images are ~ 400x larger
  - Algorithmic complexity (? See later talk)

- Model will need to be remote batch processing, either on dedicated local hardware or on cloud resources.

Conclusion:

- ngVLA does not require a revolutionary approach to computing.
- Practical to store the visibility data (although more expensive than previous telescopes)
  - Mature data processing capabilities must be part of baseline delivery.
  - Data processing will require changes in methodology to avoid repeated IO.
- Data volumes represent a significant factor in design considerations.

The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.

www.nrao.edu • science.nrao.edu