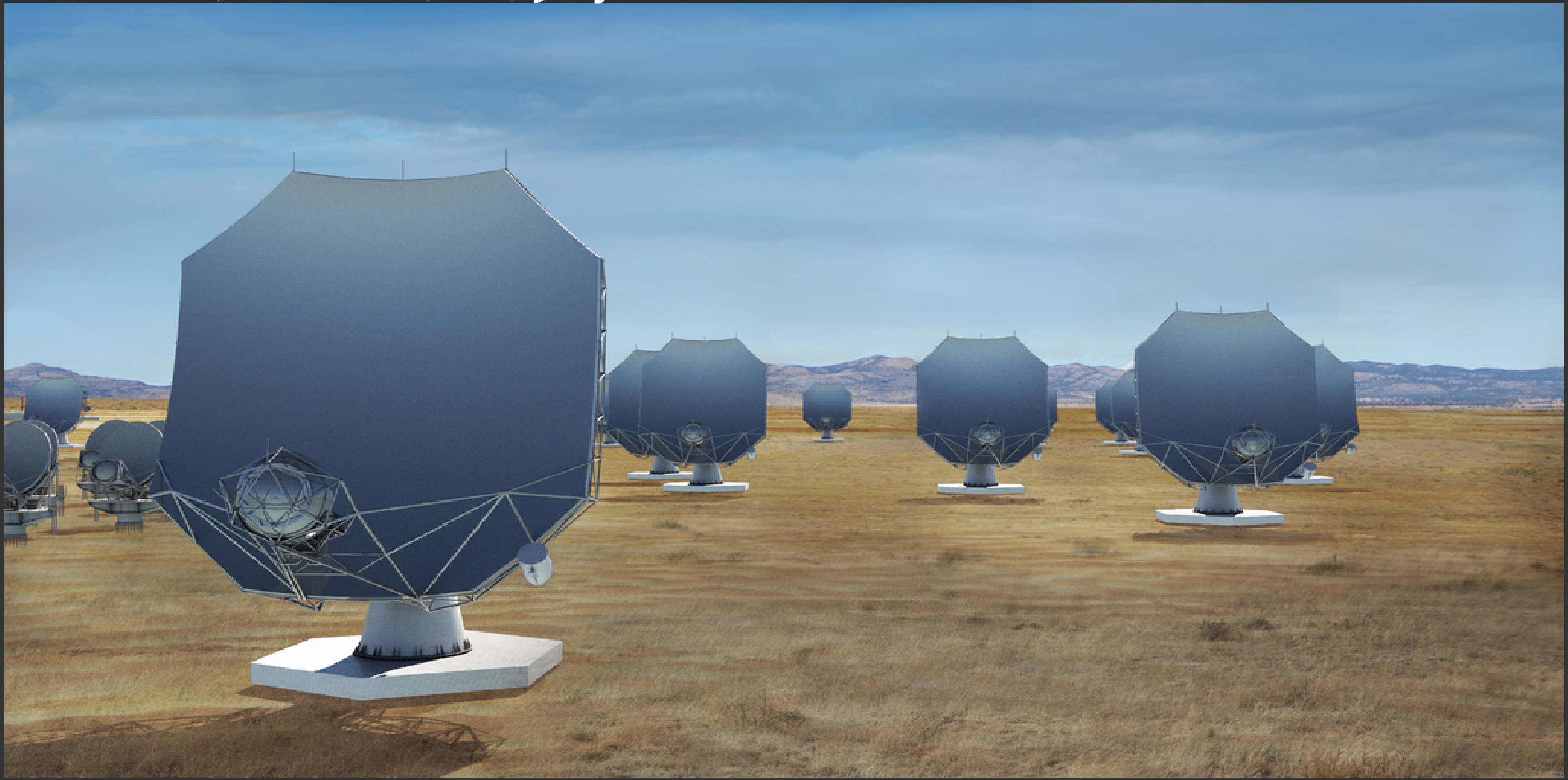
Radio Astronomical Imaging:
TeraBytes, PetaFLOPS & Algorithms

HTC 2024, Madison, WI, July 8th 2024



S. Bhatnagar

Algorithms R&D Group,
National Radio Astronomy Observatory,
Socorro, NM, USA

Introduction

- Lead for the NRAO Algorithms R&D Group
- NRAO: A NSF funded national observatory to build and operate large radio astronomy facilities: VLA, ALMA, VLBA, Greenbank Observatory

- Builds and maintains scientific software for calibration and image reconstruction
  - Widely used in the RA community internationally

- This talk: Overview of the RA data processing: What? Why? How?

  Work done with CHTC/PATh, NRP: Status, challenges, future
- Technical talk (remotely) by *Felipe Madsen*:

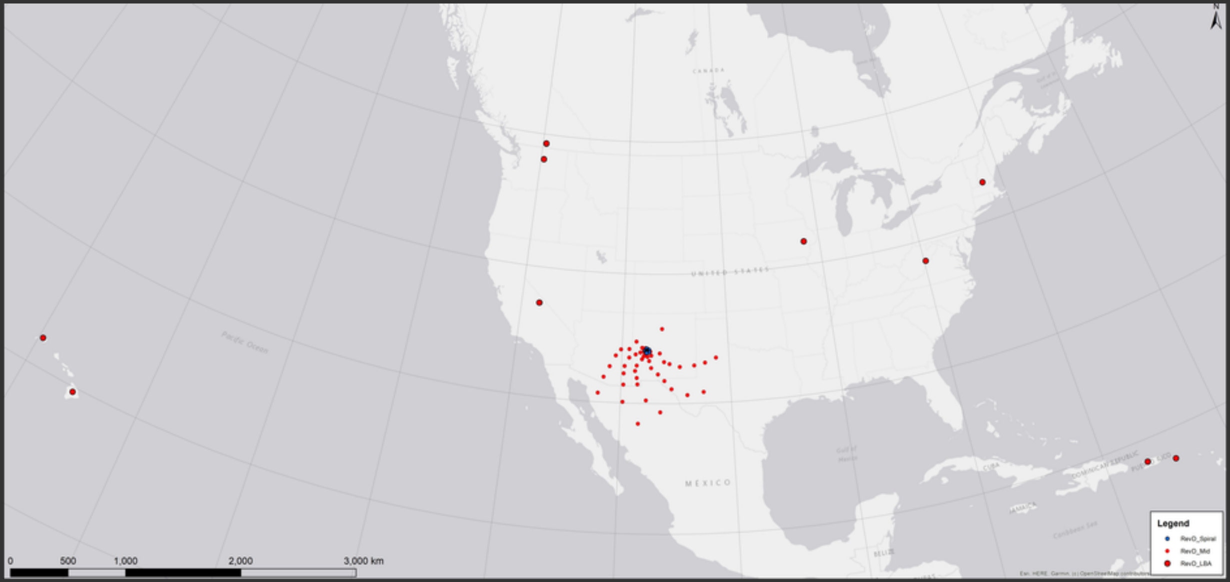
  Date: Thur, the 11th , 11:15 AM

  Title: *Implementation of NRAO's imaging workflow on HTCondor*

The Very Large Array (NM, USA)

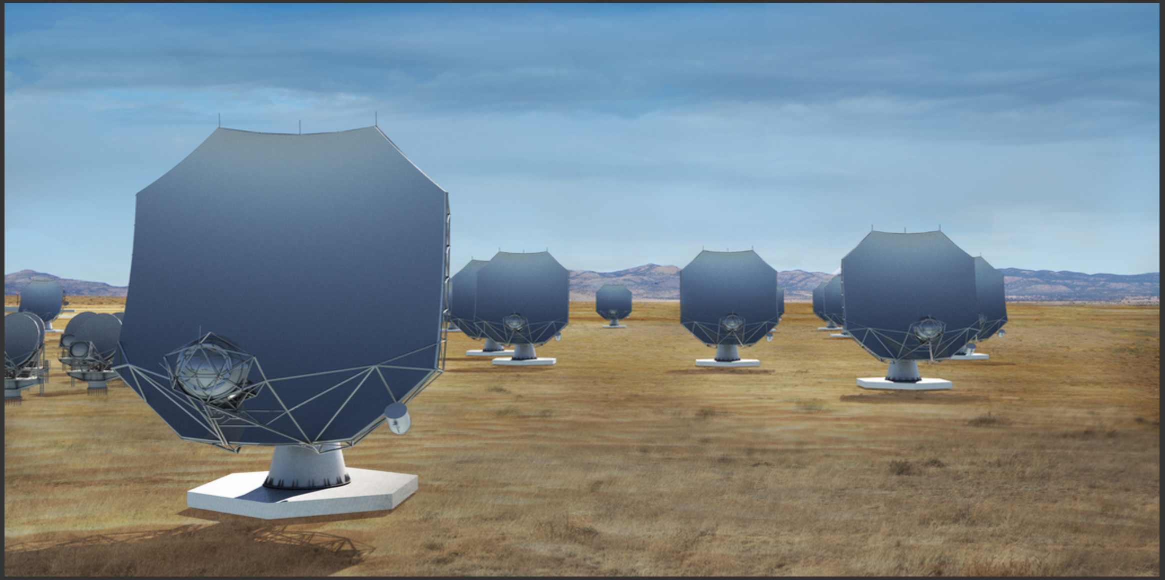
- 27 antennas
- Antennas movable on rails
  1 – 27 Km radius
- Spread over
  27 Km radius
- Size of the “lens”
  30 Km
- Frequency range
  300 MHz – 50 GHz

The next-generation VLA (ngVLA)



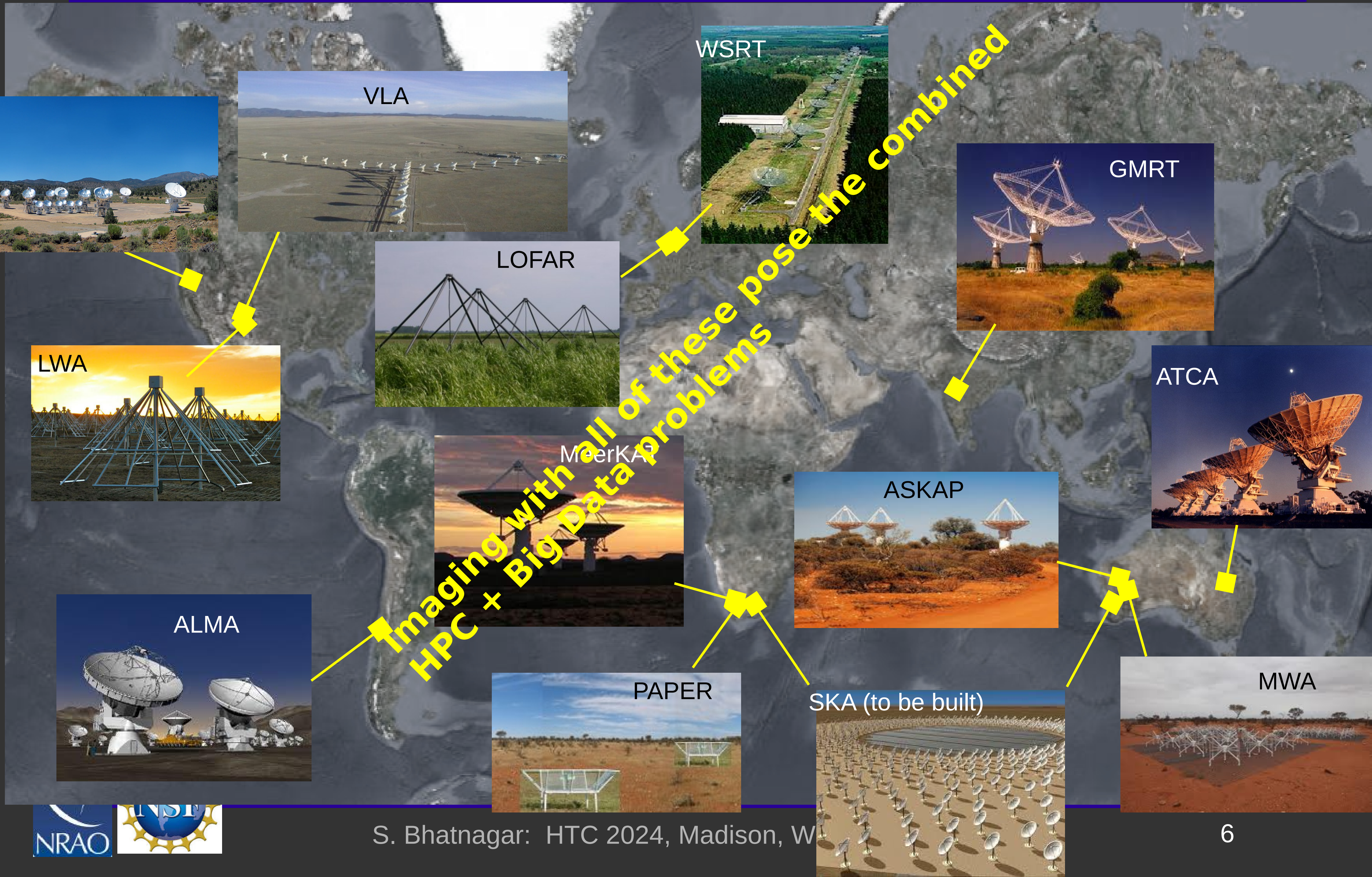
300 antennas in NM,UT,AZ,TX,MX

- ~300 antennas
- Spread across 1000s Km
- Frequency range 1 GHz – 110 GHz



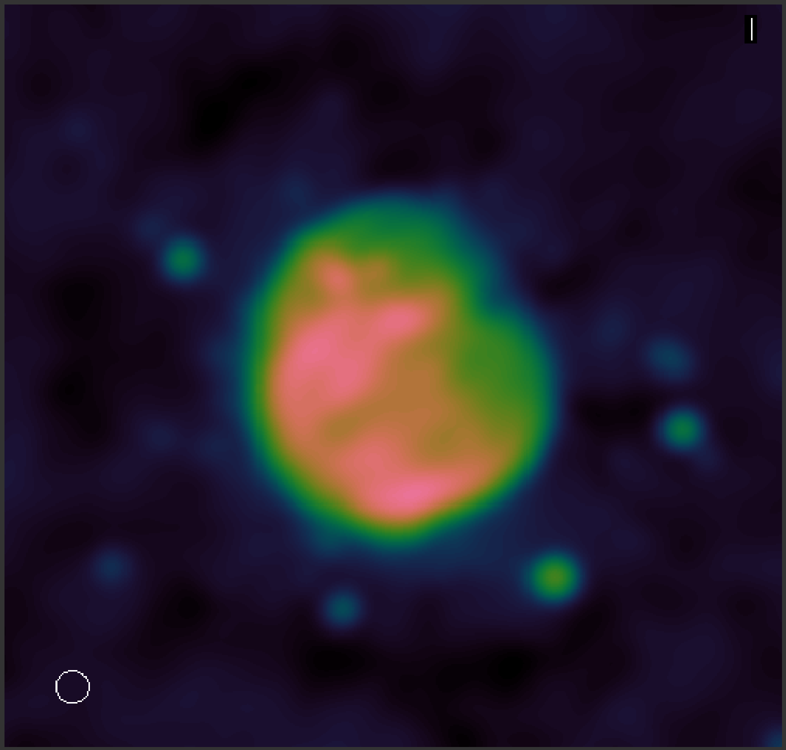
Other RA Observatories in the world

Other RA Observatories in the world



Aperture Synthesis Imaging: Why?

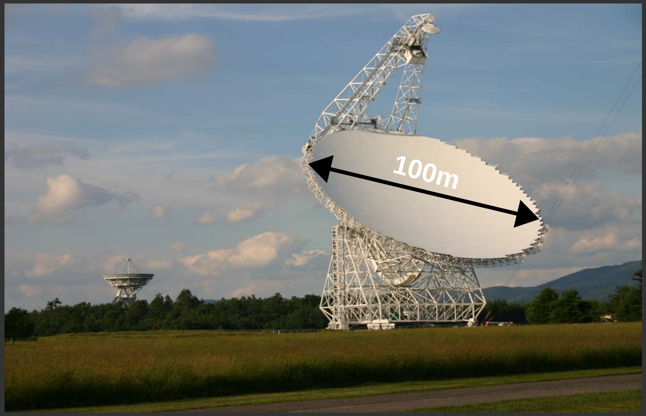
- Single dish Resolution too low for many scientific investigations
  - Limited collecting area + resolution limits sensitivity at low frequencies



Single dish resolving power

$$\frac{Wavelength}{\text{Dish Diameter}}$$

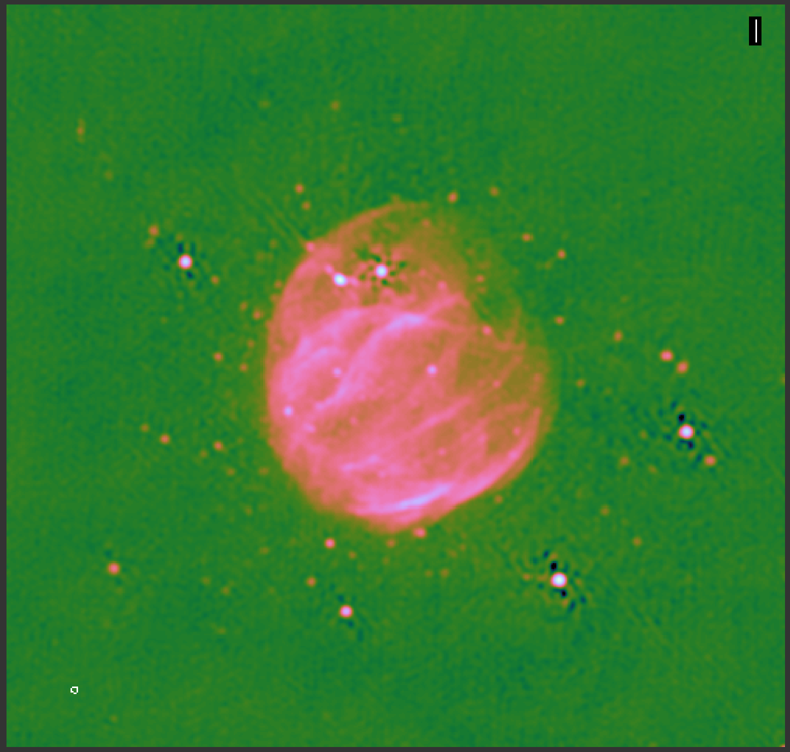
Biggest steerable single dish
= 100 m



Greenbank Observatory, WV

Aperture Synthesis Imaging: Why?

- Resolution determined by the max. separation between antennas
  - Sensitivity determined by the number and size of antennas

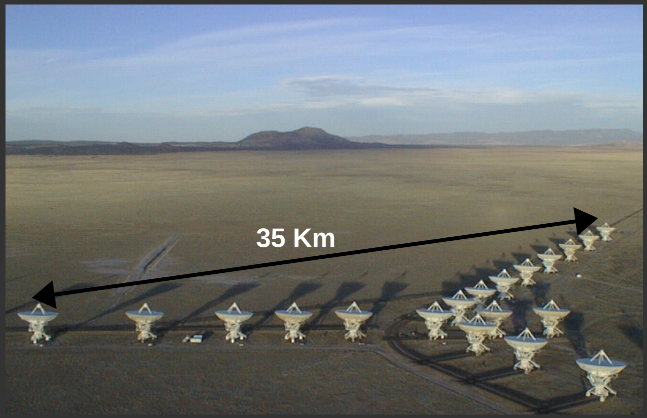


Synthesis Array resolving power

$$\frac{\textit{Wavelength}}{\text{Max. separation between antennas}}$$

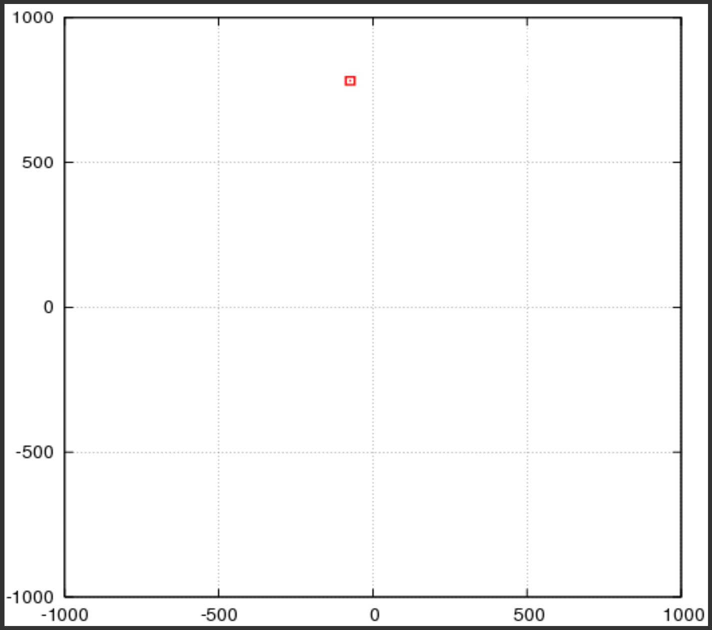
Max. separation in VLA
= 35 km

Resolution: ~ 350x better



Aperture Synthesis Imaging: How?

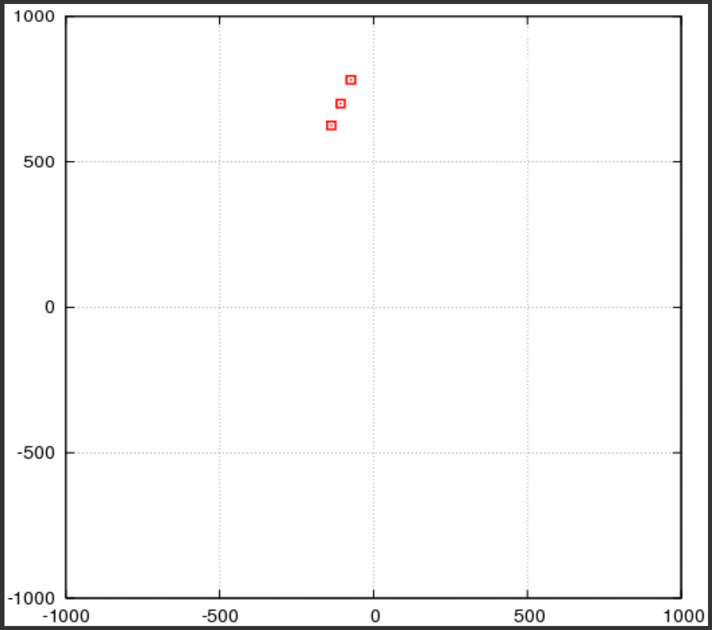
- An indirect imaging technique that collects data in the Fourier domain
  - Many antennas separated by 10s – 100s Km
  - Each pair of antennas measure **one** Fourier Component



  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

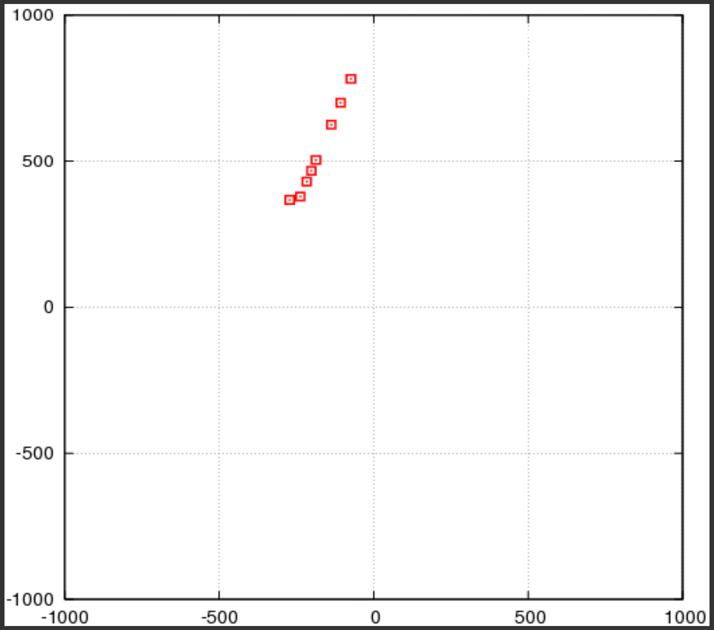
- An indirect imaging technique that collects data in the Fourier domain
  - Many antennas separated by 10s – 100s Km
  - Each pair of antennas measure **another** Fourier Component



  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

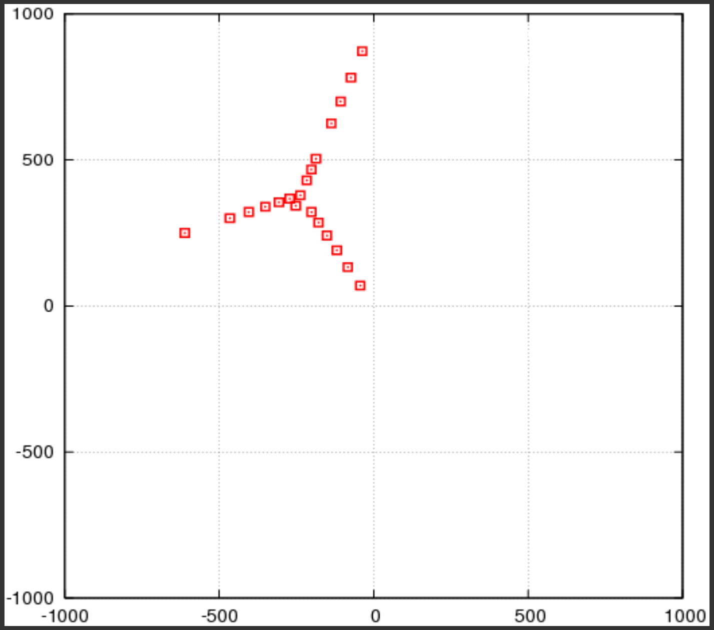
- An indirect imaging technique that collects data in the Fourier domain
  - Many antennas separated by 10s – 100s Km
  - Each pair of antennas measure **another (one)** Fourier Component

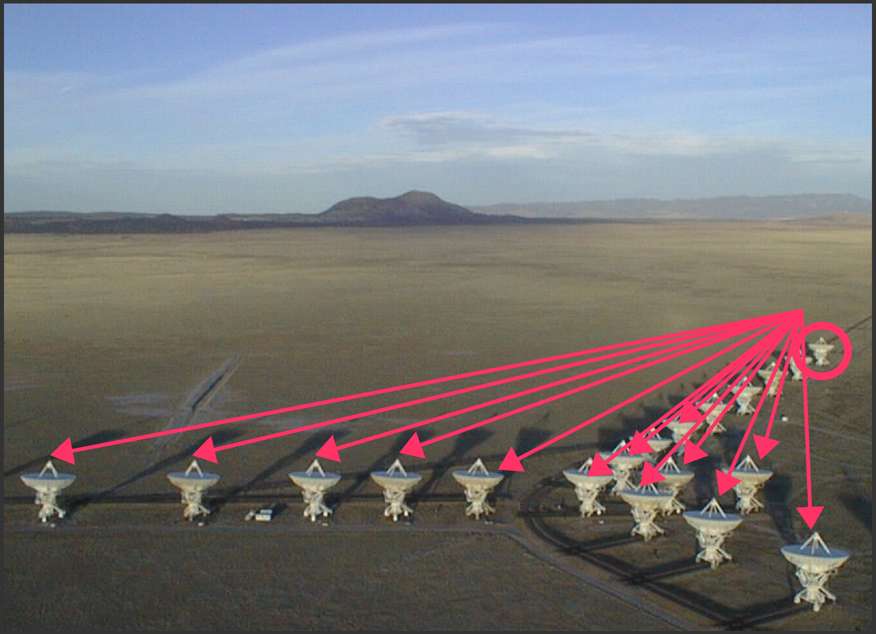


  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

- An indirect imaging technique that collects data in the Fourier domain
  - Many antennas separated by 10s – 100s Km
  - **All** pairs with **one** antenna measure N-1 Fourier Component = **26**

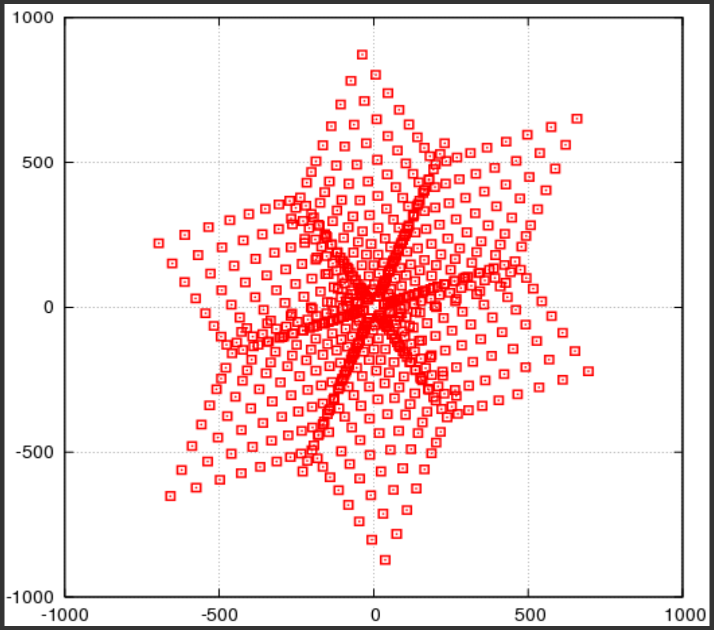


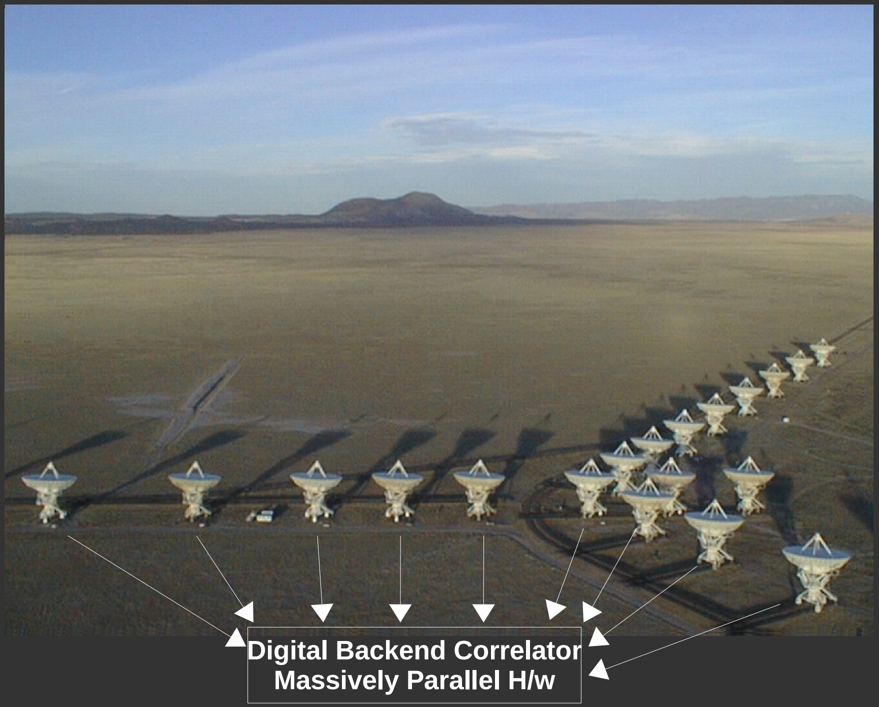


  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

- An indirect imaging technique that collects data in the Fourier domain
  - Many antennas separated by 10s – 100s Km
  - **All** pairs with **all** antenna measure N(N-1)/2 Fourier Component = **351**



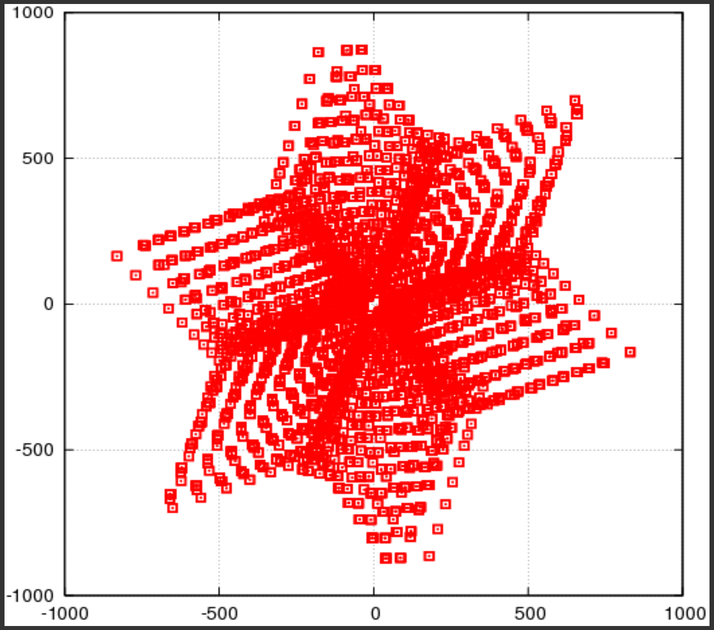


  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

- Aperture Synthesis
  - Use **Earth Rotation Synthesis** to fill the Fourier plane
  - **All** pairs with **all** antenna measures N(N-1)/2 Fourier Component

  - Measure N(N-1)/2 x 2 Fourier components over 2 integration time = **702**

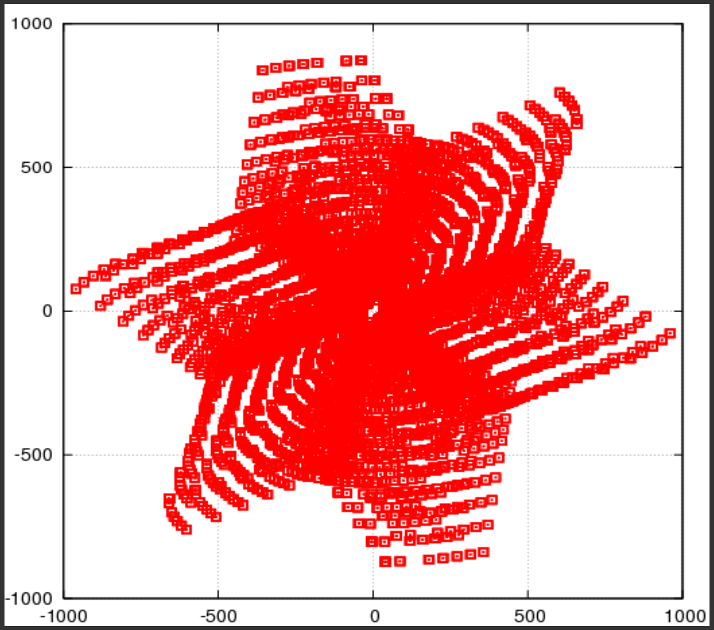


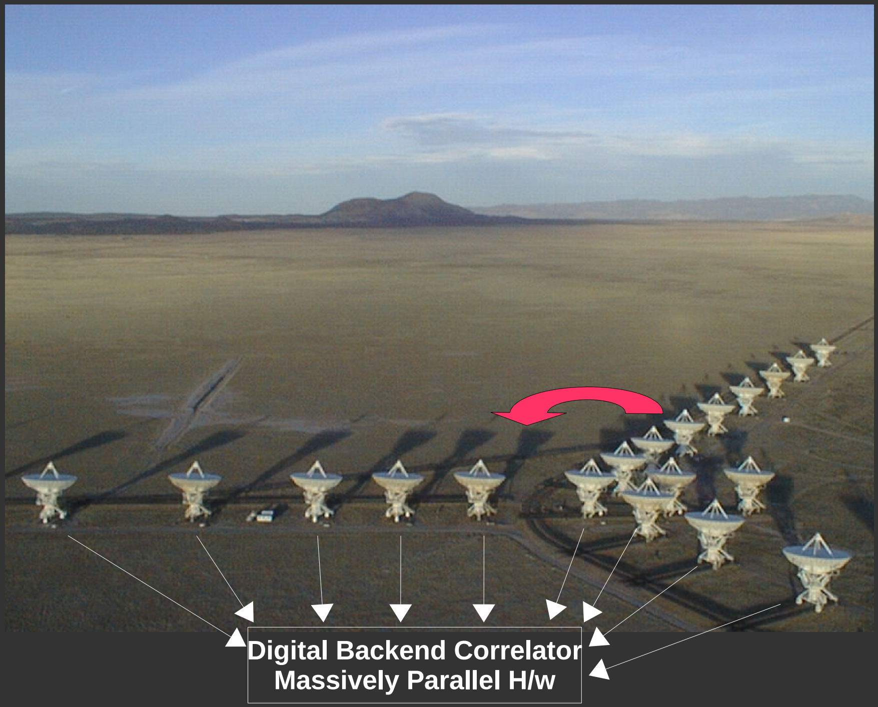
Digital Backend Correlator
Massively Parallel H/w

  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

- Aperture Synthesis
  - Use **Earth Rotation Synthesis** to fill the Fourier plane
  - **All** pairs with **all** antenna measures N(N-1)/2 Fourier Component
  - Measure N(N-1)/2 x 10 Fourier components over 10 integrations = **7020**

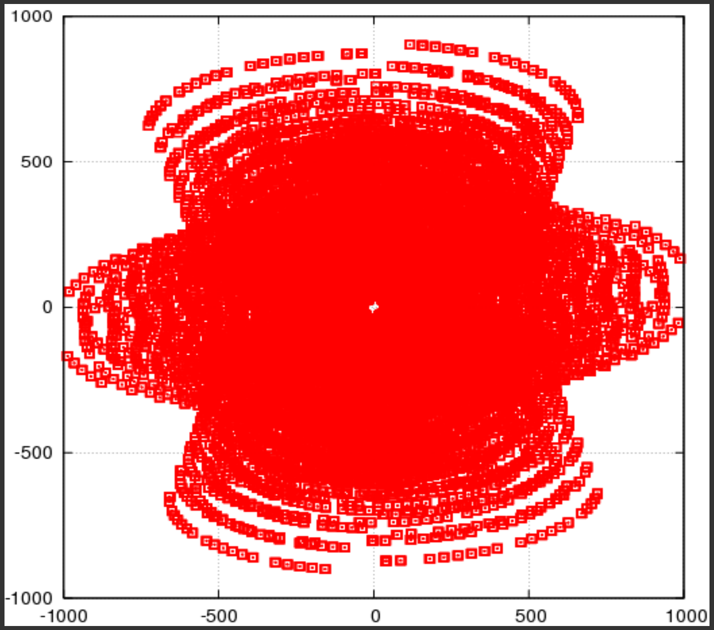


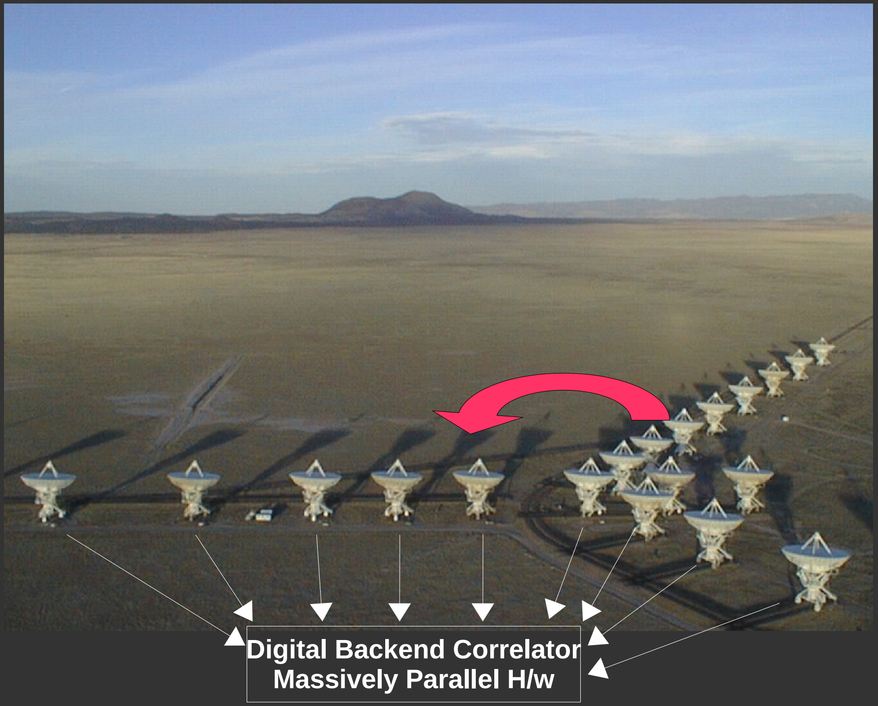


  - Synthesized aperture equal to the largest separation between antennas

Aperture Synthesis Imaging: How?

- Aperture Synthesis
  - Use **Earth Rotation Synthesis** to fill the Fourier plane
  - **All** pairs with **all** antenna measures N(N-1)/2 Fourier Component
  - Fourier Components measured over 10 hr: **O(10^{12})**

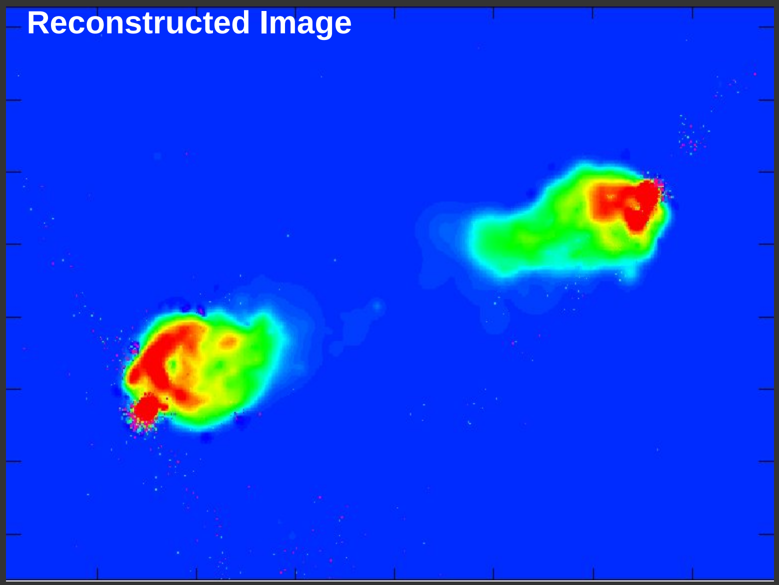


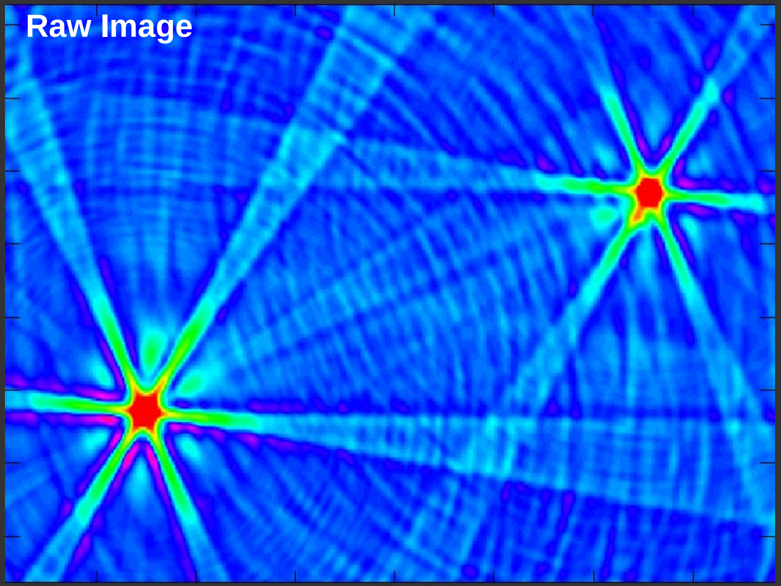


- Data Size: 10s TB now, 100s TB with ngVLA   ExaBytes for SKA-class telescopes
  - Data not on a regular grid.

Interferometric Imaging

- Raw image (FT of the raw data) is dynamic range limited





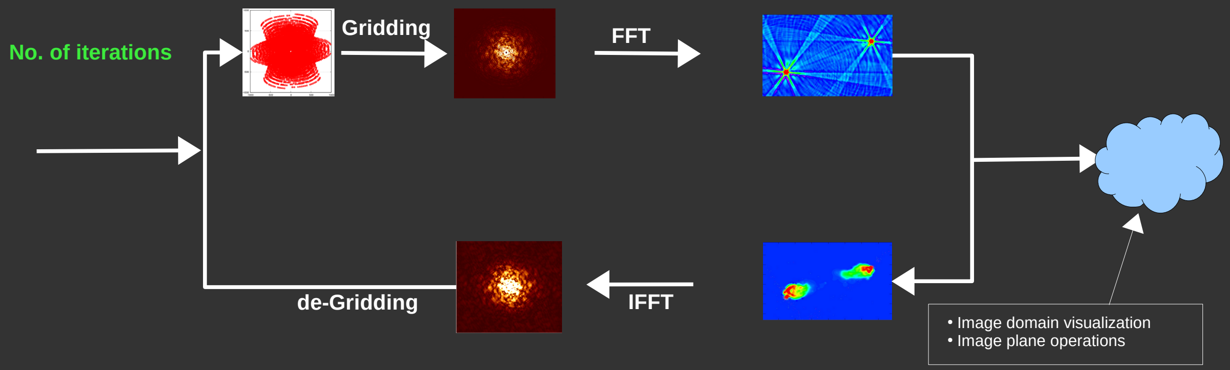
Dynamic range: > 1 : 1000, 000

Dynamic range: 1 : 1000

- Processing: Remove telescope artifacts to reconstruct the sky brightness
- RA image reconstruction is a High-Performance-Computing-using-Big-Data problem
- Using dHTC + distributed data collection. Hard to even know everything that can go wrong!

System level description

- Typical data processing workflow

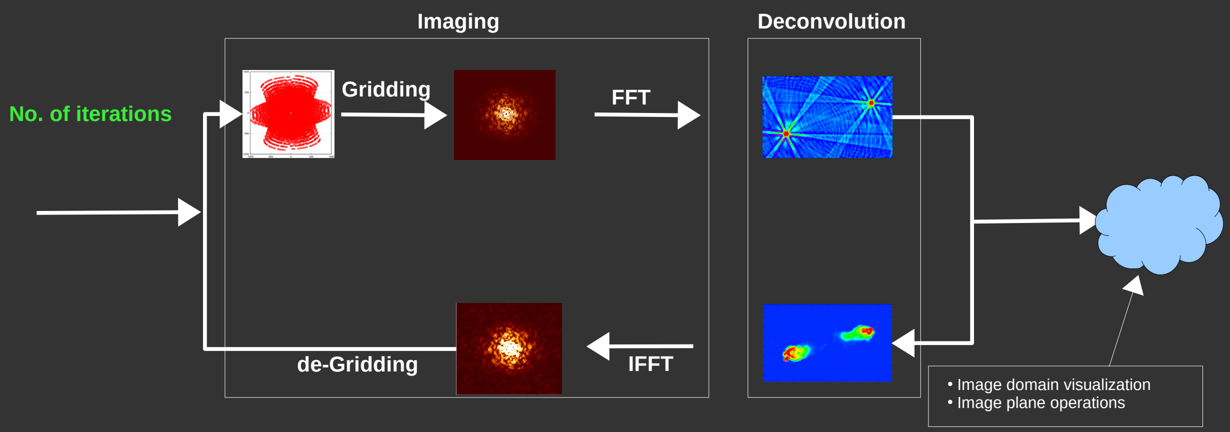


System level description

- Typical data processing workflow + Size of Computing (SofC)

  Imaging: N_{vis} x $O(10^{3\text{-}4})$ FLOPs (Complex, SP + DP)

  Image-plane deconvolution of the PSF : N_{iter} x $O((N_{pix})$ FLOPs (Real-valued, SP)

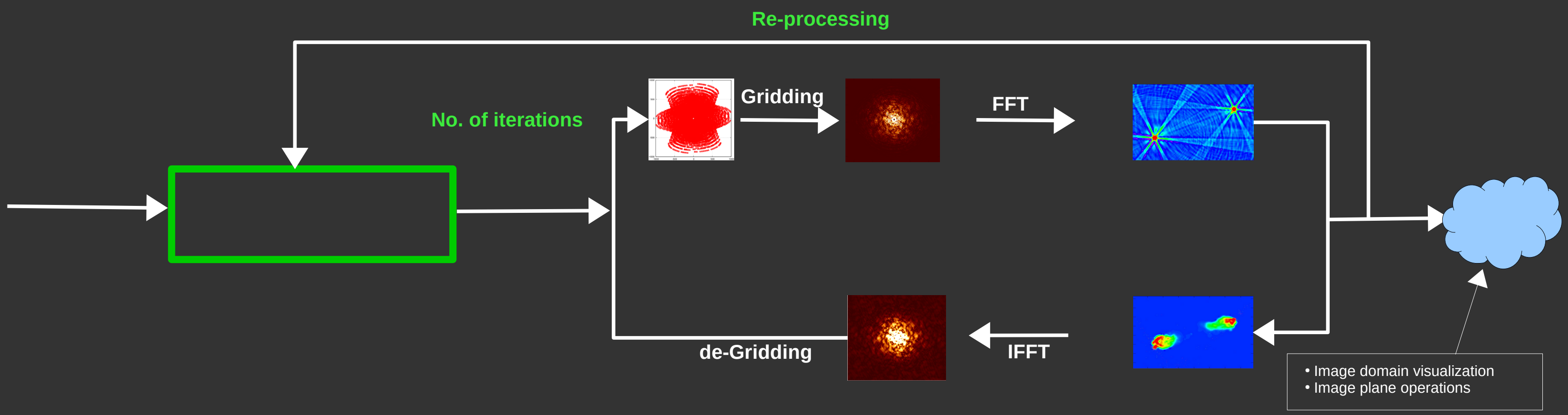


System level description

- Typical data processing workflow + Size of Computing (SofC)

  Imaging: N_{vis} x $O(10^{3\text{-}4})$ FLOPs (Complex, SP + DP)

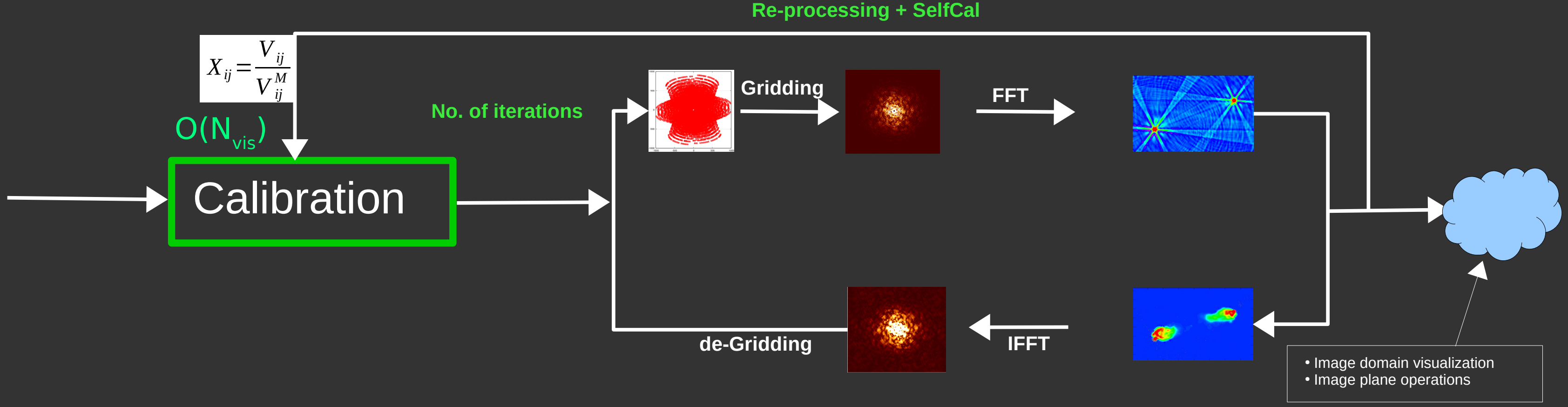
  Image-plane deconvolution of the PSF : N_{iter} x $O((N_{pix})$ FLOPs (Real-valued, SP)



System level description

- Typical data processing workflow + Size of Computing (SofC)

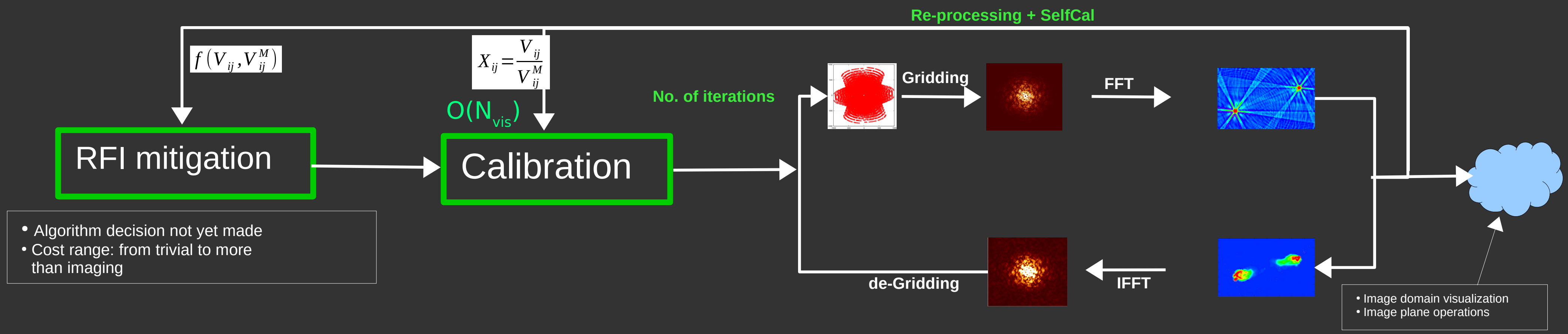
Imaging:	N_{vis} x $O(10^{3\text{-}4})$	FLOPs (Complex, SP + DP)
Image-plane deconvolution of the PSF :	N_{iter} x $O((N_{pix})$	FLOPs (Real-valued, SP)
Calibration:	$O(N_{vis})$	FLOPs (Complex, SP)
	N_{vis} x $O(10^{3\text{-}4})$	FLOPs (Complex, SP + DP)



System level description

- Typical data processing workflow + Size of Computing (SofC)

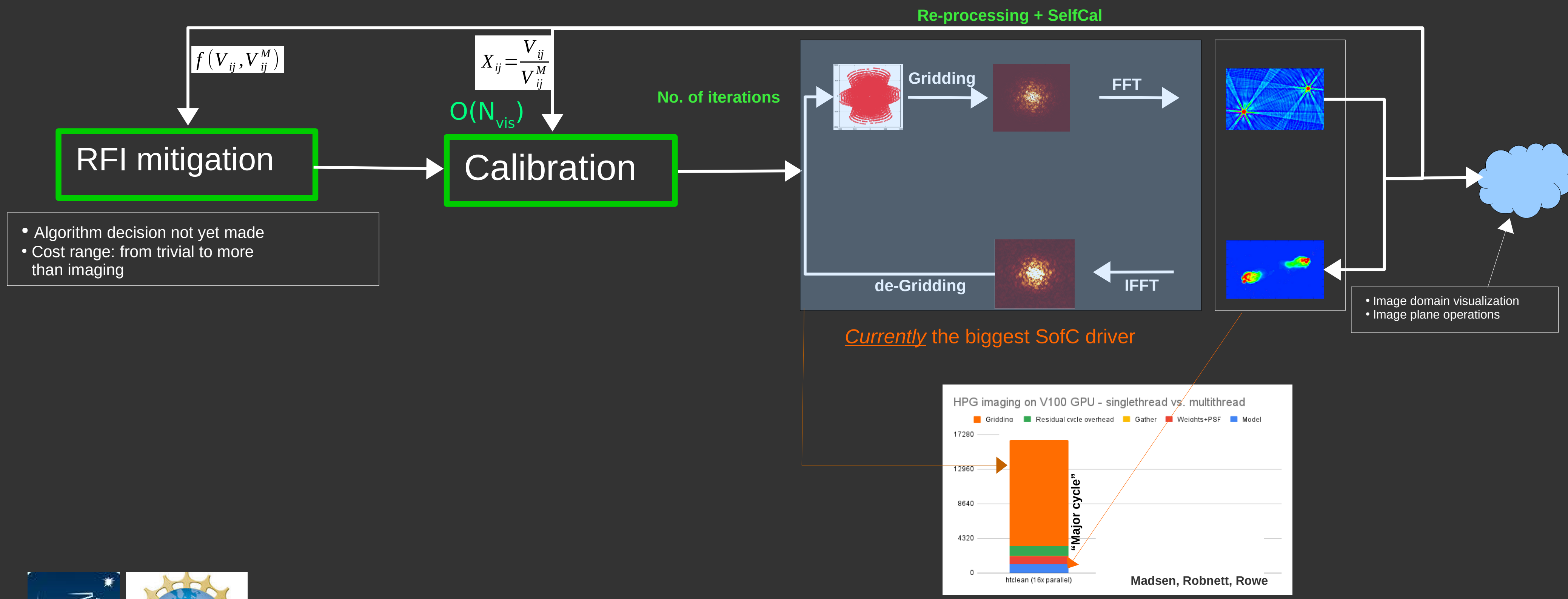
Imaging:	N_{vis} x $O(10^{3\text{-}4})$ FLOPs (Complex, SP + DP)
Image-plane deconvolution of the PSF :	N_{iter} x $O((N_{pix})$ FLOPs (Real-valued, SP)
Calibration:	$O(N_{vis})$ FLOPs (Complex, SP)
	N_{vis} x $O(10^{3\text{-}4})$ FLOPs (Complex, SP + DP)
Flagging:	Trivial → dominant!



System level description

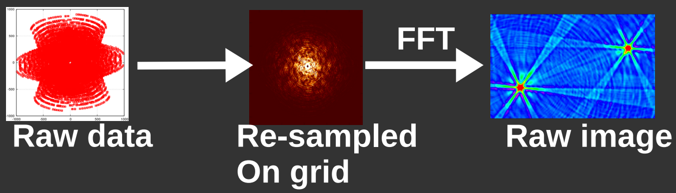
- Typical data processing workflow + Size of Computing (SofC)

Imaging:	N_{vis} x $O(10^{3\text{-}4})$ FLOPs (Complex, SP + DP)
Image-plane deconvolution of the PSF :	N_{iter} x $O((N_{pix})$ FLOPs (Real-valued, SP)
Calibration:	$O(N_{vis})$ FLOPs (Complex, SP)
	N_{vis} x $O(10^{3\text{-}4})$ FLOPs (Complex, SP + DP)
Flagging:	Trivial → dominant!

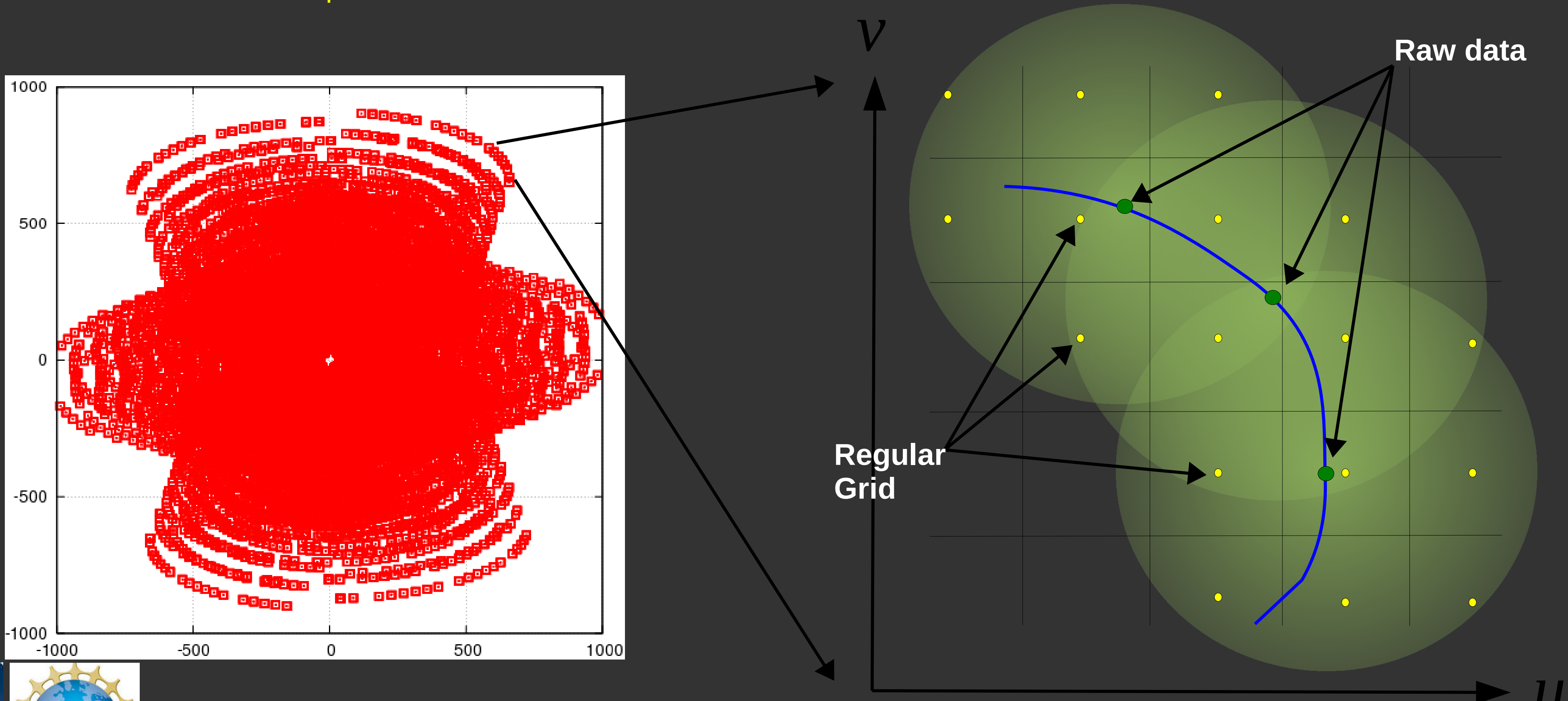


The Computing Problem: Why Gridding?

- Raw data is not on a regular grid
  - FFT require re-sampling on a regular grid

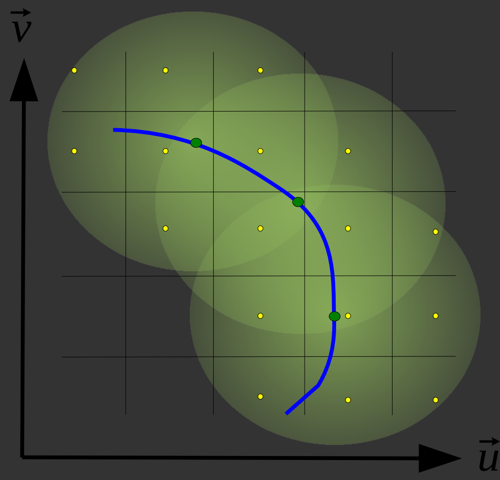


NU-FFT with the physics/optics of the telescope
encoded in specialized kernels



Requirements: HPC + Big Data

- Estimated Size of Computing
  - ngVLA: O(10^{13-14}) x (10x10) x ...= ~50 PFLOP/s



- Large scale parallelization to process large data volume
  - Not a simulation!
  - PFLOPS to keep-up with the data rates
  - 100s of Tera Bytes for a typical observing session

- Computing needs to be efficient and 24x7
  - Not a one-shot experiment on a homogeneous super-computer

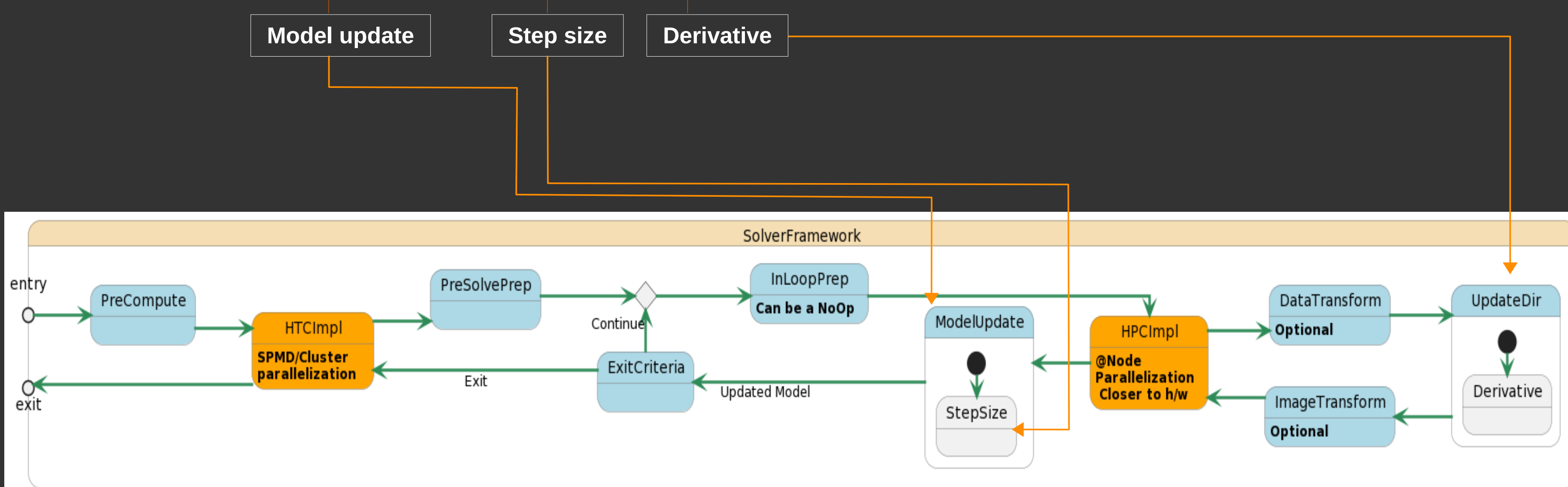
- Requirement: Seamless computing 24x7 on a heterogeneous cluster

Algorithm Architecture

- Stable, Scalable Architecture
  - Cast RA algorithms in standard terminology: Derivative, Hessian, Update,…
  - Decompose into functionally separable components which can scale individually and together

$$V^{obs} = \mathbf{G}^M S\, \mathbf{F}\, B^M I^M + noise \qquad \chi^2 = \sum_i \left| Data_i - Model_i(P) \right|^2$$

$$P_i^{k+1} = P_i^k + [H_{ij}]^{-1} \, f\left(\frac{\partial \chi^2}{\partial P_i^k}\right) \quad ; \qquad [H_{ij}] = \frac{\partial^2 \chi^2}{\partial P_i^k \, \partial P_j^k}$$



Algorithm Architecture: Components view

$$P_j^{k+1} = P_j^k + [H]^{-1} f\left(\frac{\partial \chi^2}{\partial P_j^k}\right)$$

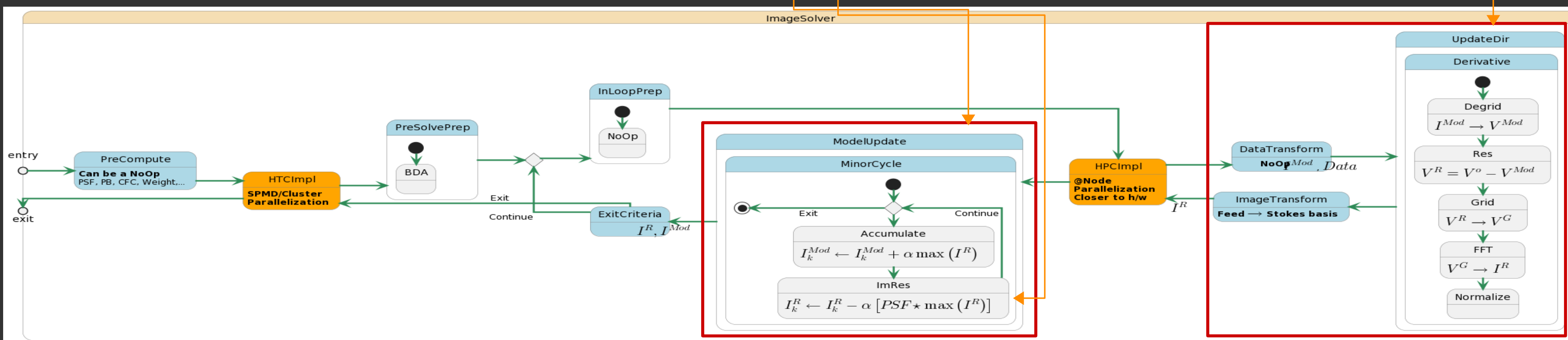
a.k.a. the "Minor cycle"

Update Model Image

a.k.a. the "Loop gain"

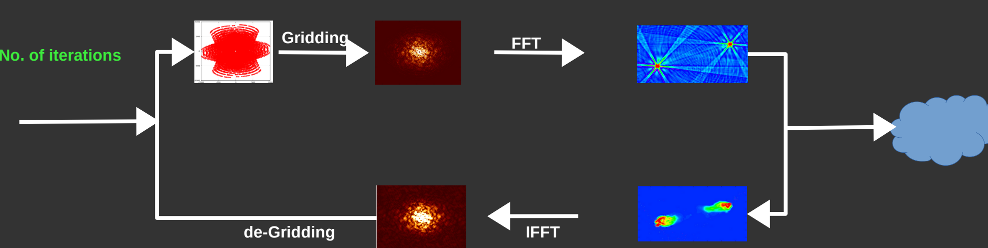
a.k.a. the "Major Cycle"

Compute Residual Image

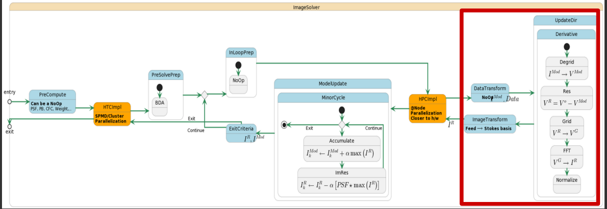


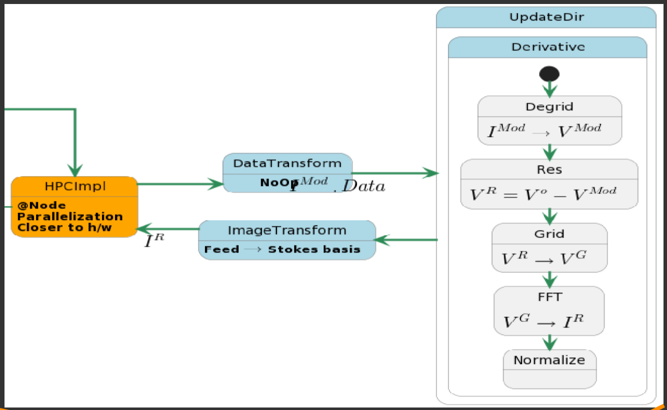
Serial

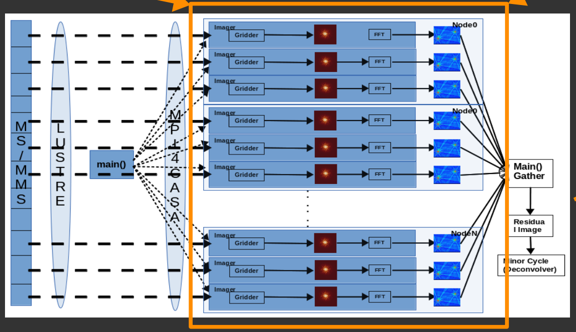
Multi-process parallel

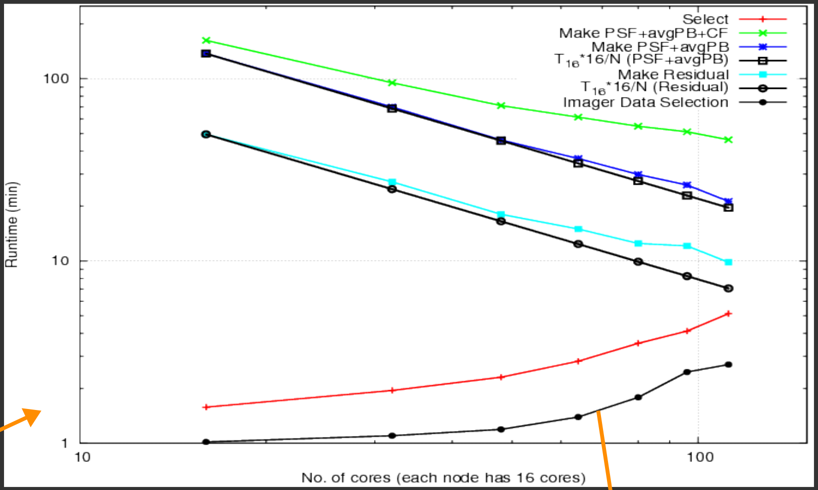


Scaling: On multi-CPU/cores hardware





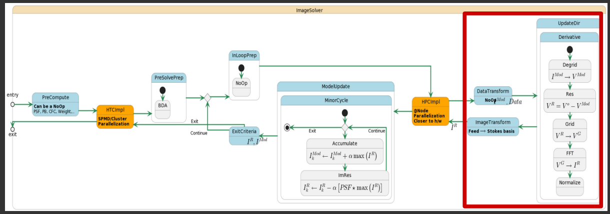


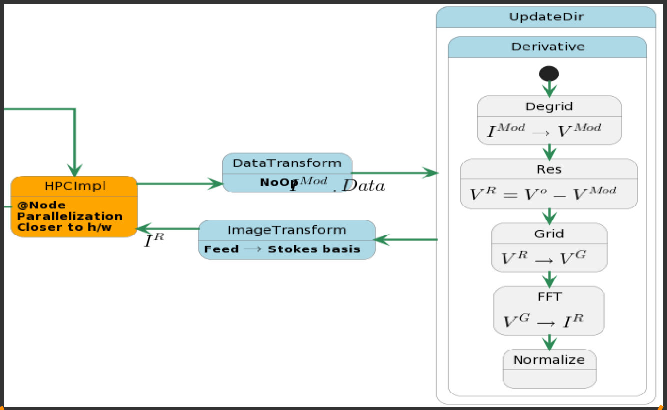


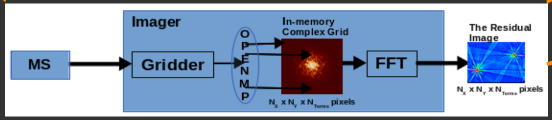
Data scatter overheads

ngVLA would need O(Million)-way parallelization!

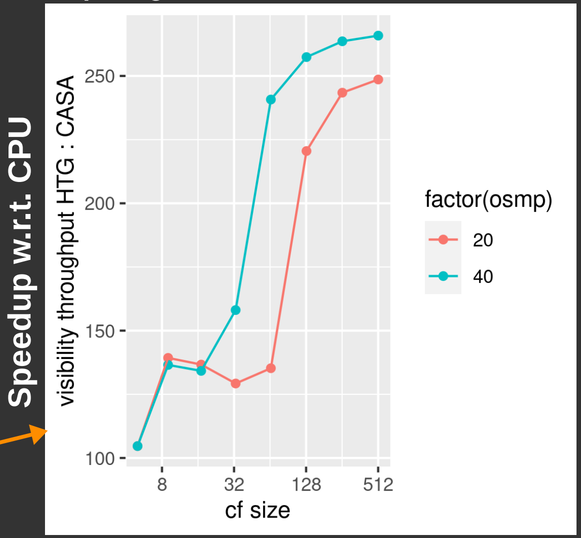
Scaling on GPU: Using Kokkos

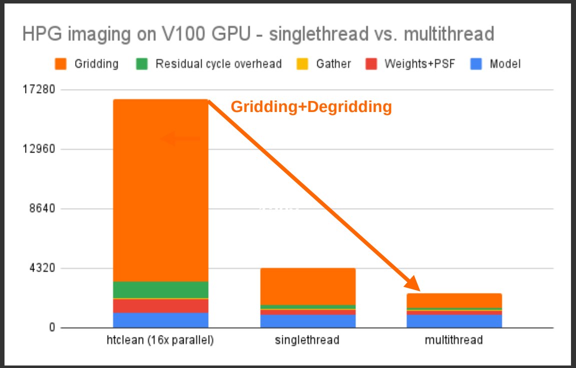






ngVLA Computing Memo #5, #7





Multi-core CPU

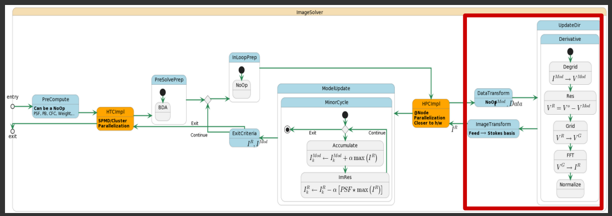
Single CPU-core +GPU

Multi CPU-cores +GPU

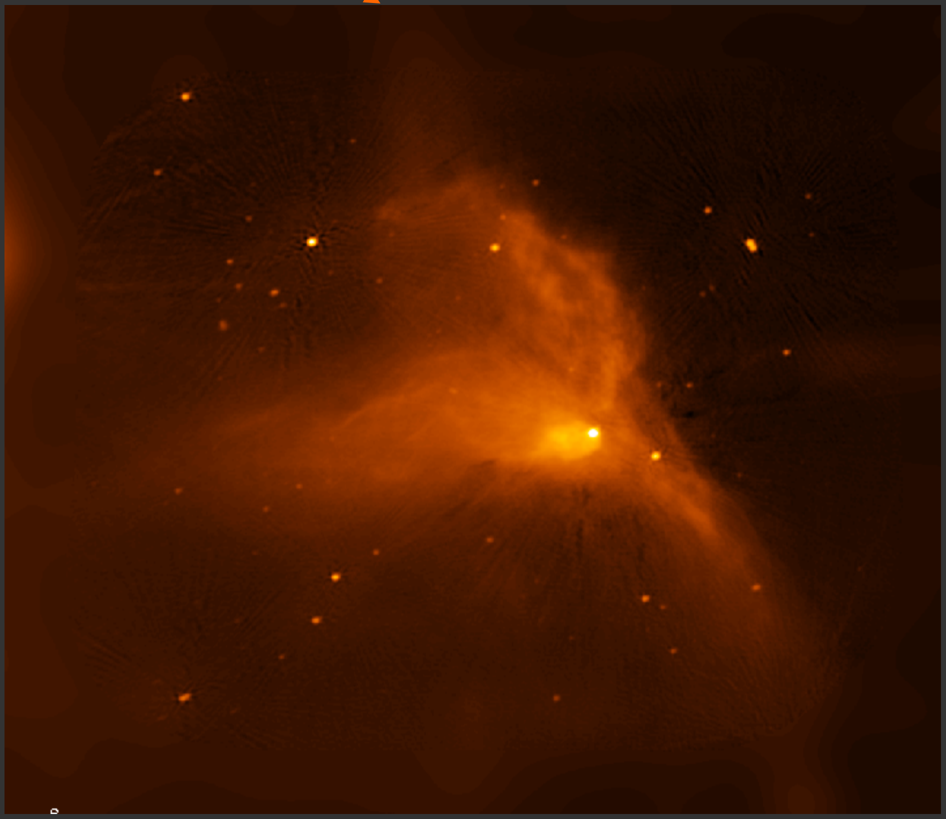
ngVLA would need O(10^3)-way parallelization!

Scaling in real-life

- What does it mean in real-life application?
  - 200-pointing wide-band mosaic: 7-10 days vs 2.5hr



Current telescope



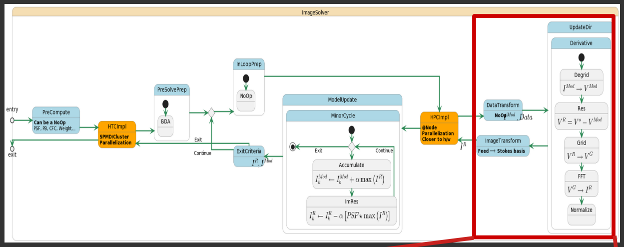
- Many data sets in the telescope archive remain UNPROCESSED due to computing limitations

- Brings down ngVLA need to **O(10^3)-way** parallelization!

ngVLA would need O(10^3)-way parallelization!

Throughput measurements

- Deployed on a cluster of GPUs (100) on the PATh facility in collaboration with
  - CHTC (UW-M), NRP, SDSC (UCSD)
  - + Multiple university computer centers across the US



Enabling-tech for many unprocessed projects in the current archive

- Throughput: O(1 TB/hr)
- 10 iterations in ~24 hr
  - Previous attempts:
    ~14 days per iteration!

- This is still a faction of the required throughput!

https://science.nrao.edu/enews/17.3/index.shtml#deepimaging

Deepest Image in the radio band of the Hubble Ultra-Deep Field

- Deployed on a cluster of GPUs (100) on the PATh facility
  - https://science.nrao.edu/enews/17.3/index.shtml#deepimaging

Data volume:	2 TB
Effective data I/O:	20 TB
Throughput:	~1 TB/hr

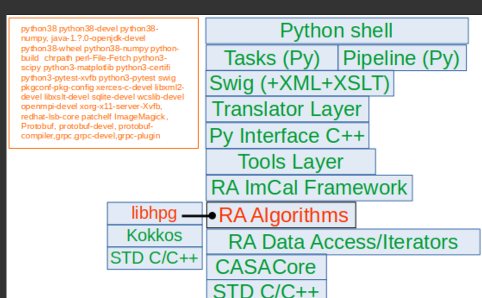
The LibRA Project: Library of RA algorithms

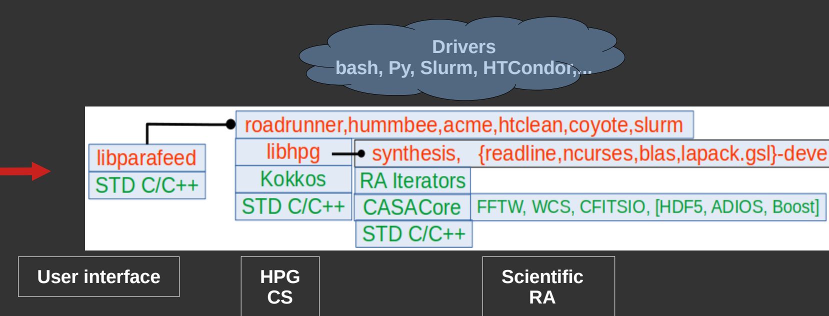
- The LibRA project was used for all RA domain functionality
- Project goals: Open-source library of RA algorithms, code re-use, relocatable s/w, ease of use
  - Derived from CASA Scientific. Now an independent code base + build system
  - Enables collaborations with RA groups and end-users + with other domains: HPC, HTC, Medical imaging,...

The LibRA Project: Library of RA algorithms

- The LibRA project was used for all RA domain functionality
- Project goals: Open-source library of RA algorithms, code re-use, relocatable s/w, ease of use
  - Derived from CASA Scientific. Now an independent code base + build system
  - Enables collaborations with RA groups and end-users   + with other domains: HPC, HTC, Medical imaging,...
- Directly use the scientific layer via standalone applications
  - Deployable on external heterogeneous cluster of CPUs + GPUs
- Automate chores: cmake build system, containerized deployment, Py binding,...
- Interfaces: Interactive, commandline, Py, C++

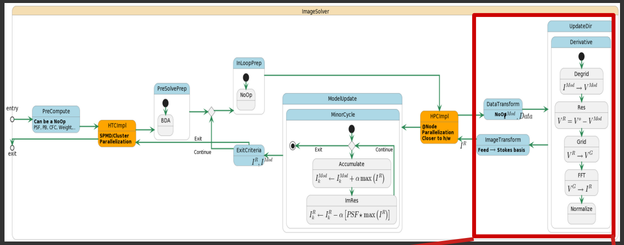
https://github.com/ARDG-NRAO/LibRA



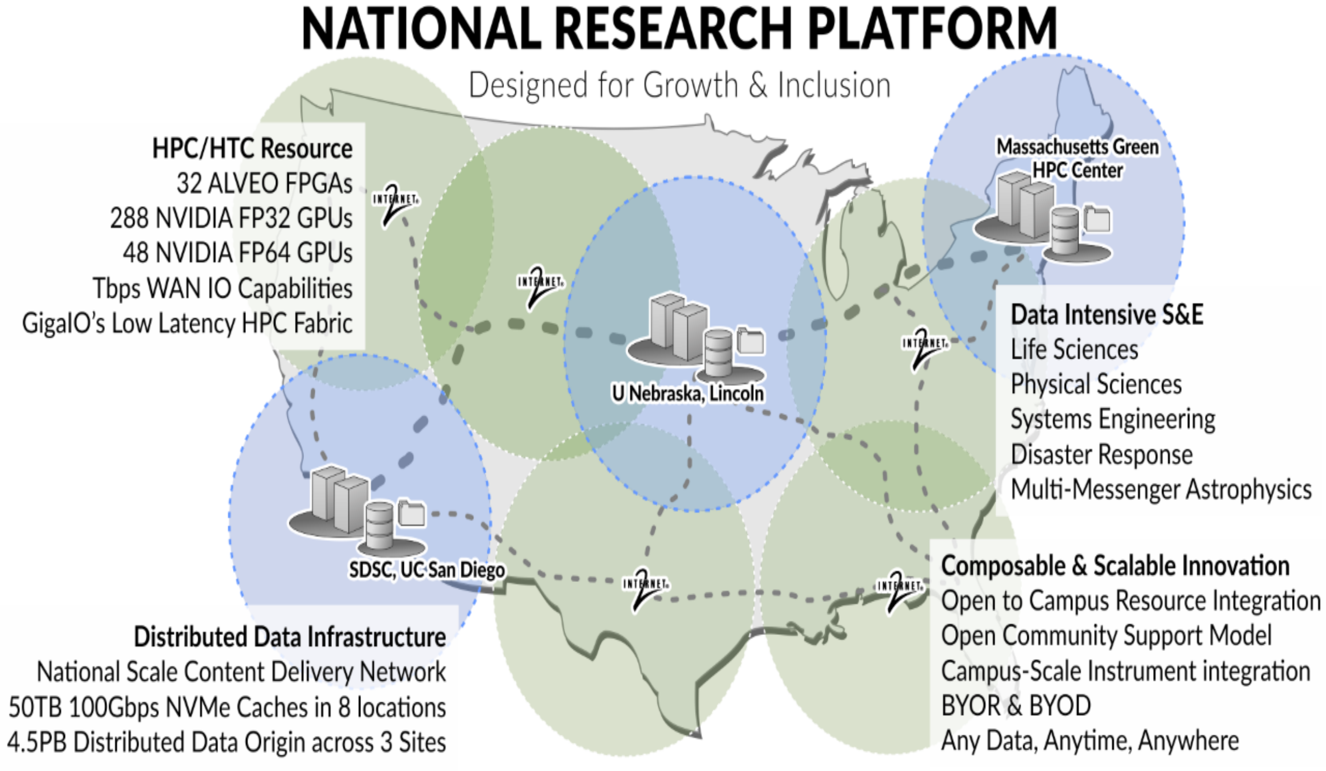


Throughput measurements

- Deployed on a cluster of GPUs (100) on the PATh facility in collaboration with
  - CHTC (UW-M), NRP, SDSC (UCSD)
  - + Multiple university computer centers across the US



Enabling-tech for many unprocessed projects in the current archive



- Throughput: O(1 TB/hr)
- 10 iterations in ~24 hr
  - Previous attempts:
    ~14 days per iteration!

- This is still a faction of the required throughput!
- Can distributed network of GPUs deliver?

Current/near future needs

- KSG data volumes: ~50 TB per hour
  Throughput: O(50) TB per hour
  Effective I/O: few x O(50) TB
  - Iterative algorithms need caching to keep data closer to the compute

- Computing Infra development
  - Inter-leaved processing of multiple projects
  - Interfacing with telescope storage/archive, pipeline processing systems
    - OSDF, Pelican?

- Edge Caching, Data re-use (iterative algorithms)

- Parallel processing
  - On connected cluster of GPUs (in-house?)
  - On distributed resources: NRP, PATh, Super computer centers,…

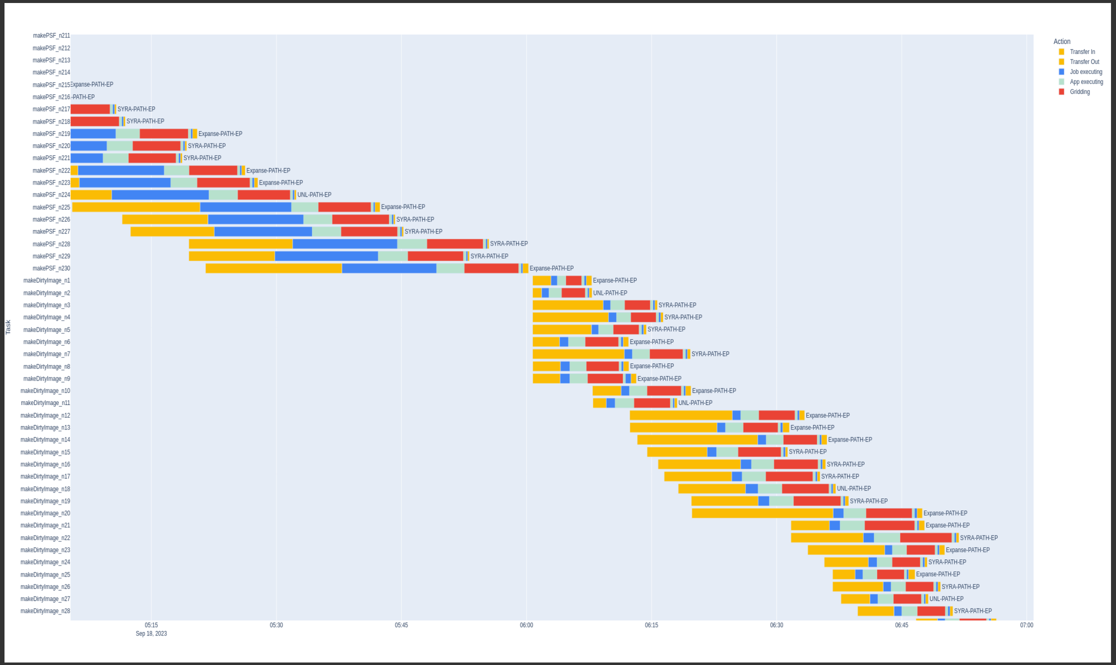
Current/near future needs

- Collaboration to access existing human and computing resources
  - Expect runtime of order days (not hours!)
  - Data volume of order 10s of TeraBytes; effective I/O 5 -- 10x larger
  - Mitigate I/O overheads due to iterative re-use (caching)

- Non-hero run: More use by the wider RA community
  - Unprocessed data from projects in the telescope archive
  - Computing power to get imaging quality compatible with telescope capabilities

- Use for algorithms R&D
  - Development, debugging
  - I/O, Cache friendly algo

- Async gather
  - Currently a Barrier



Conclusions

- Parallelization at multiple scales necessary for RA imaging
- Use of GPUs is also necessary
- Effective data I/O: Cache friendly applications and infra necessary
- New algorithms which can scale on large clusters (distributed or connected)
- Use of performance engineering tools: E.g. Kokkos
  - Seamless deployment on heterogeneous cluster of GPUs
- Collaborations between observatories and HTC/HPC groups, computer centers, and industry partners for infra development is going to be more critical than in the past

Thank you all!

- **CHTC/PATh** : Brian Bockelman, Miron Livny, Christina Koch, Brian Lin, Greg Thain
  Derek Weitzel (UNL)
  Mats Rynge (USC)
- **SDSC (UCSD)** : Frank Wuerthwein, Cynthia Dillon
- **NRP/SDSC** : John Graham, Igor Sfiligoi, Dima Mishin, Mahidhar Tatineni, Dmitry Mishin

- **Kokkos (SNL)** : Christian Trott, Lebrun-Grandie,...
- **NVIDIA** : Tom Gibbs, Eliot Eshelman, Adam Thompson, Mike O'Keeffe

- **NRAO CASA Team** : International team of radio astronomer/scientists
- **CASACore** : International team for the RA infrastructure library (US, EU, JP, TW, AU, SA)

- **NRAO Algorithms R&D Group (ARDG)**: Felipe Madsen, Mingue "Genie" Hsieh, Preshanth Jagannathan, K. Scott Rowe, Martin Pokorny (now @CalTech)

  LibRA - A library of RA algorithms