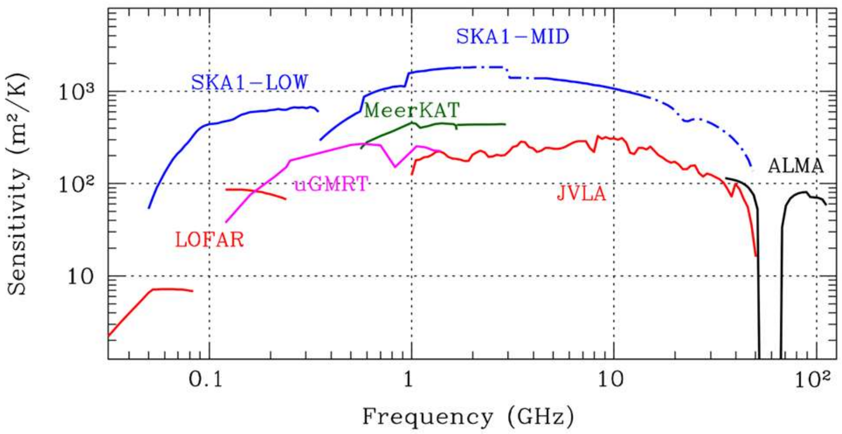
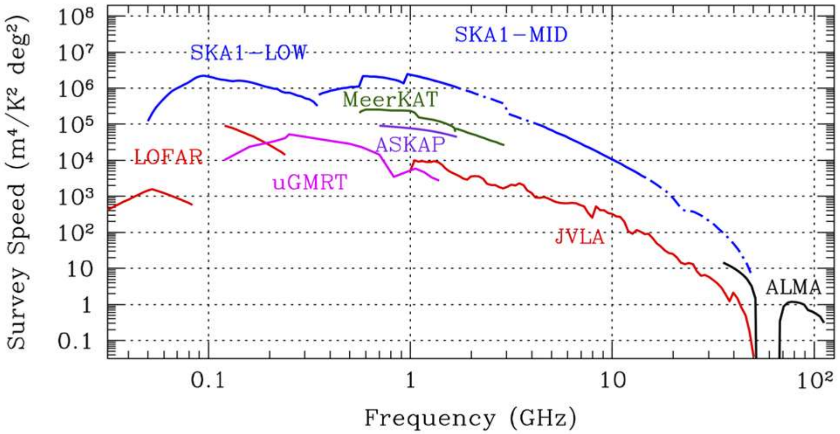
Decentralizing radio-interferometric image reconstruction by spatial frequency

Sunrise Wang

OBSERVATOIRE DE LA CÔTE D'AZUR
UNIVERSITÉ CÔTE D'AZUR
RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité
LAGRANGE
cnrs

Introduction





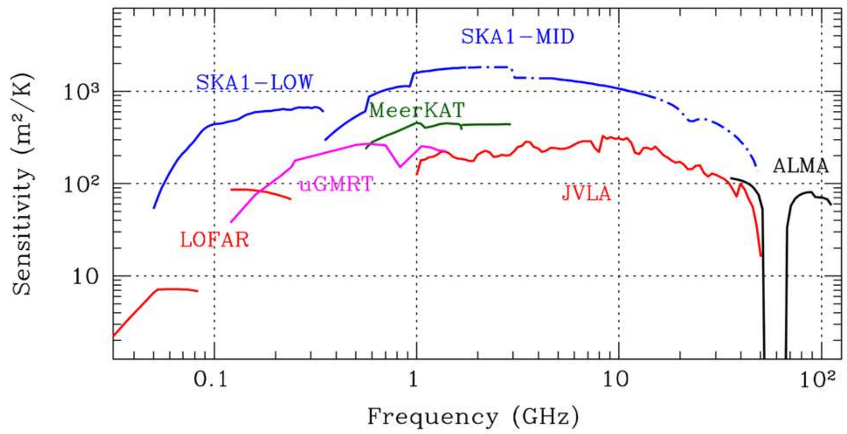
SKA-Low

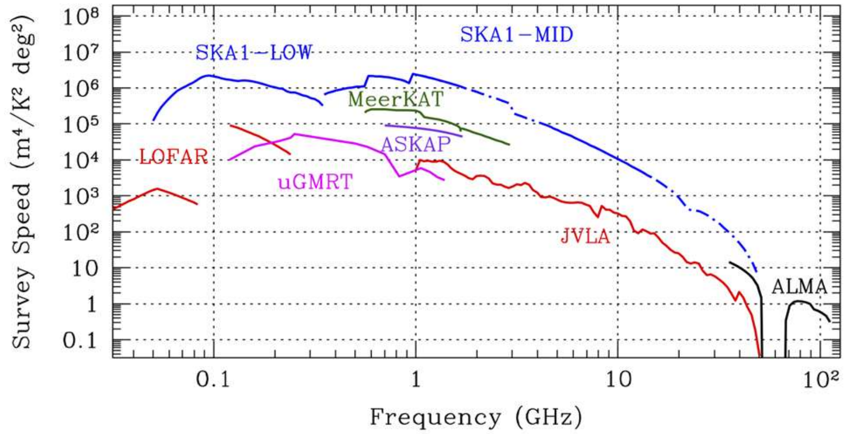
SKA-Mid

image sources: [3]
data source: [1, 2]

[1] Labate, Maria G., et al. "Highlights of the square kilometre array low frequency (SKA-LOW) telescope." Journal of Astronomical Telescopes, Instruments, and Systems 8.1 (2022): 011024-011024.
[2] Swart, Gerhard P., Peter E. Dewdney, and Andrea Cremonini. "Highlights of the SKA1-Mid telescope architecture." Journal of Astronomical Telescopes, Instruments, and Systems 8.1 (2022): 011021-011021.
[3] https://www.skao.int

Introduction





SKA-Low

SKA-Mid

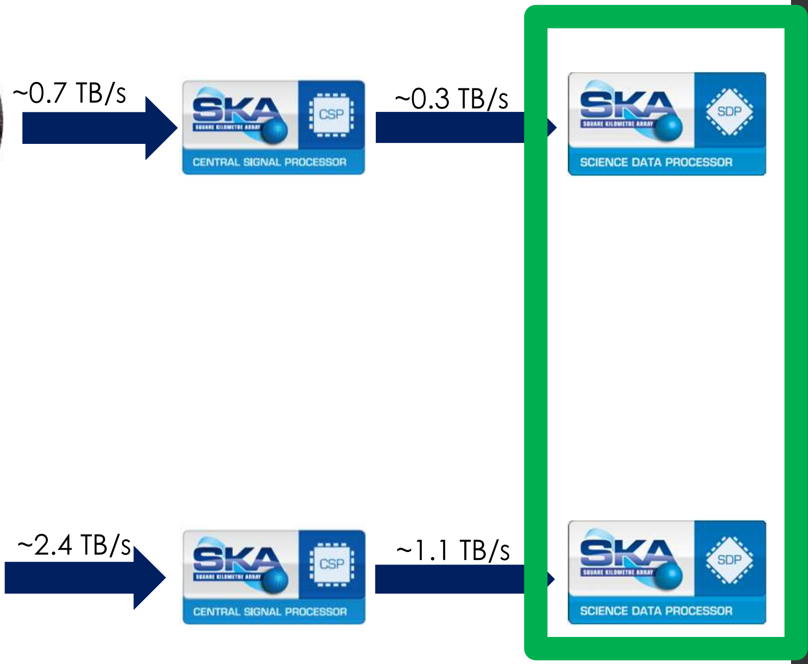
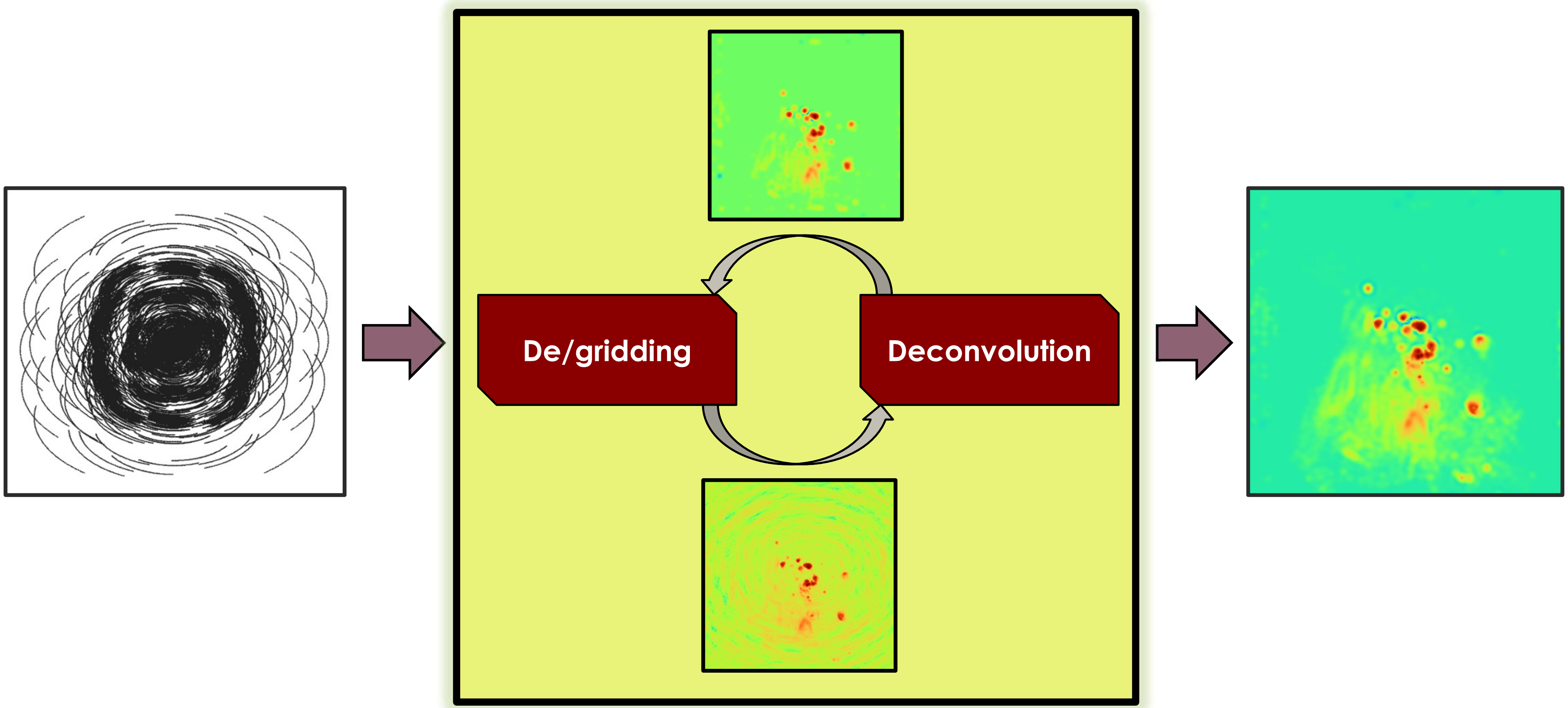


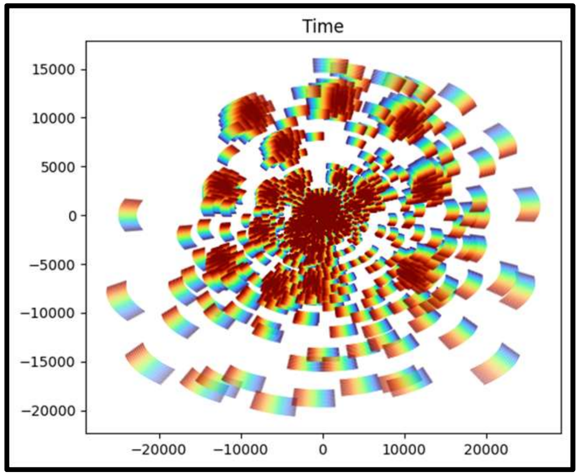
image sources: [3]
data source: [1, 2]

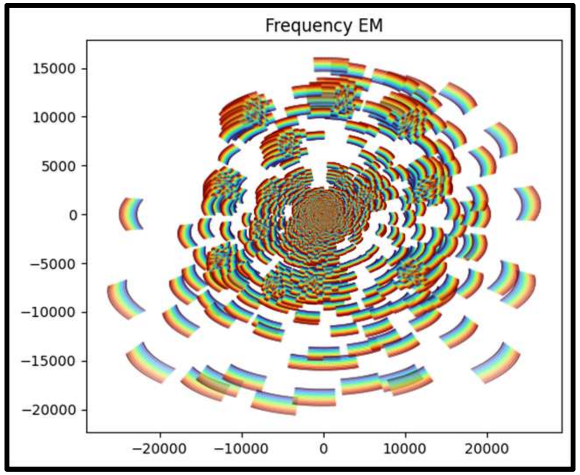
[1]Labate, Maria G., et al. "Highlights of the square kilometre array low frequency (SKA-LOW) telescope." Journal of Astronomical Telescopes, Instruments, and Systems 8.1 (2022): 011024-011024.
[2] Swart, Gerhard P., Peter E. Dewdney, and Andrea Cremonini. "Highlights of the SKA1-Mid telescope architecture." Journal of Astronomical Telescopes, Instruments, and Systems 8.1 (2022): 011021-011021.
[3] https://www.skao.int

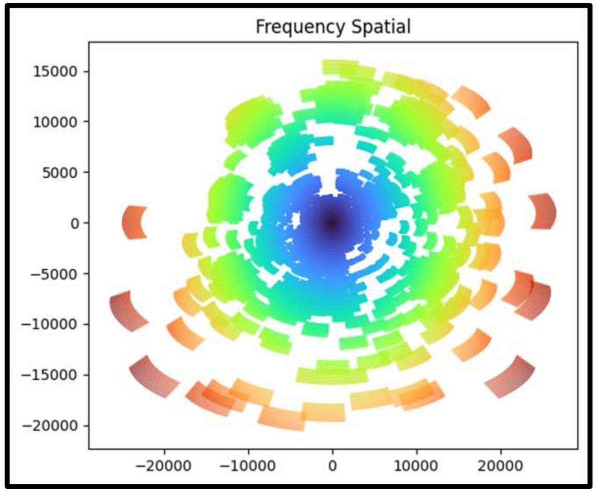
Major-Minor Loop Reconstruction



Introduction

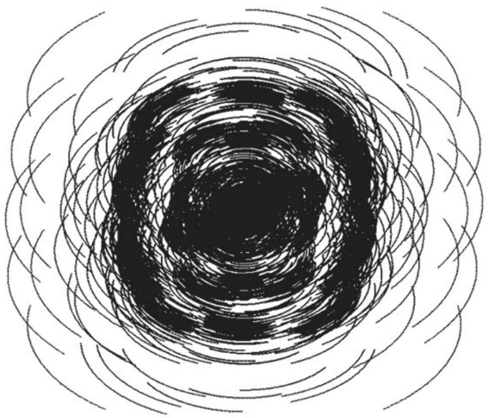






- Focus on scaling radio-interferometric imaging pipeline, with a view of the upcoming SKA telescopes
- Parallelize processing of visibilities (de/gridding)
- Traditional "simpler methods include parallelizing by time and EM-frequency domains
- We focus on framework for spatial frequency parallelization

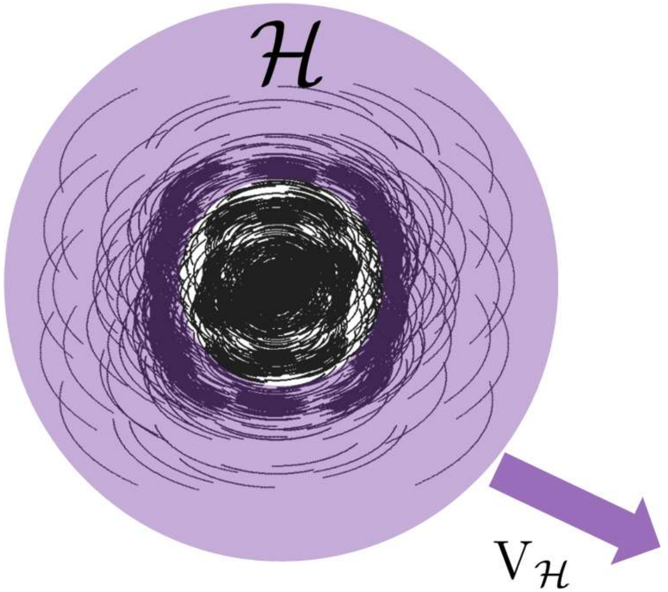
Decentralized Framework for 2 partitions

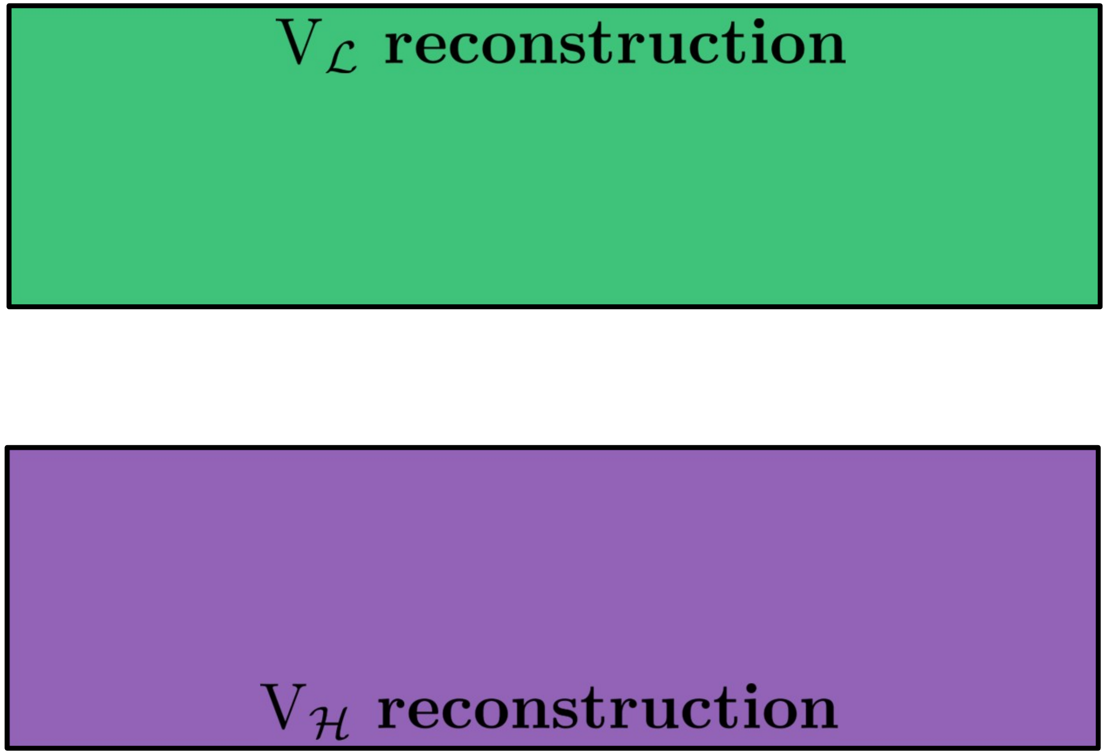


$V_{\mathcal{L}}$ **reconstruction**

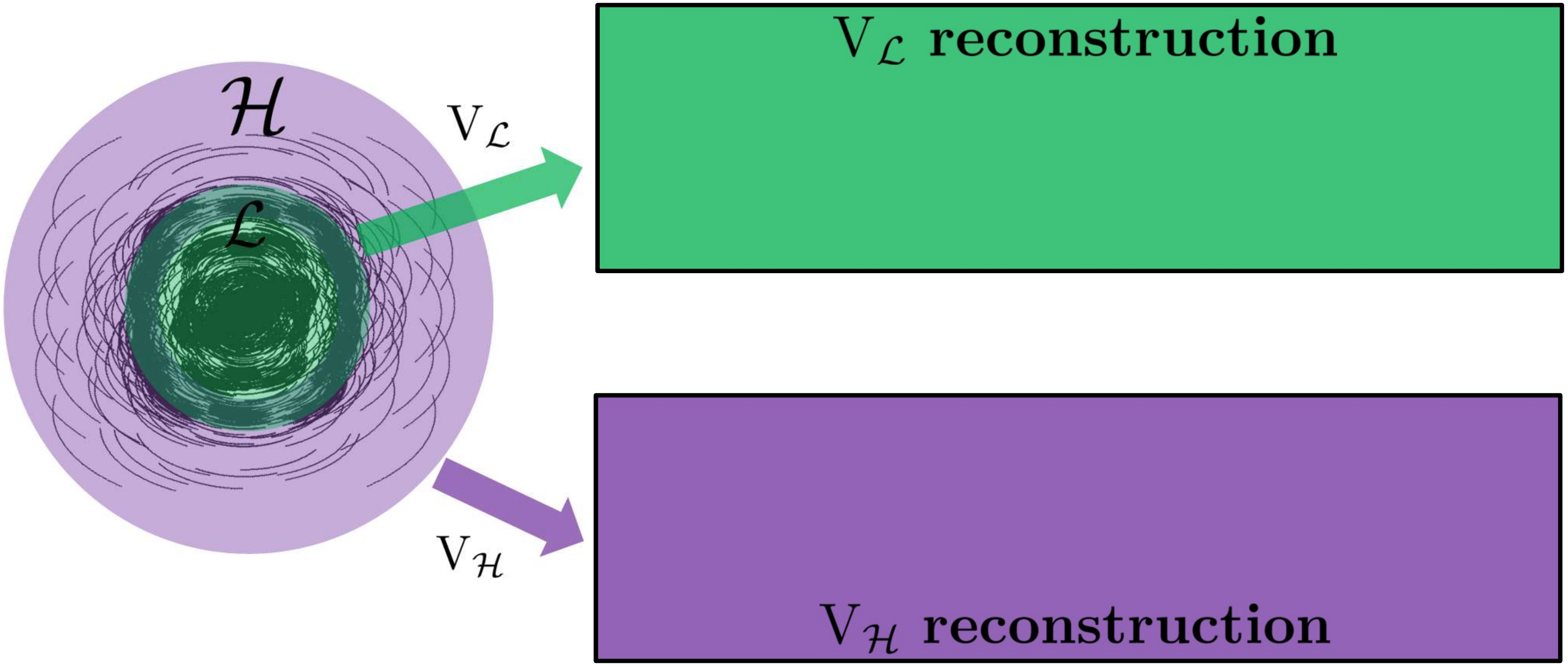
$V_{\mathcal{H}}$ **reconstruction**

Decentralized Framework for 2 partitions

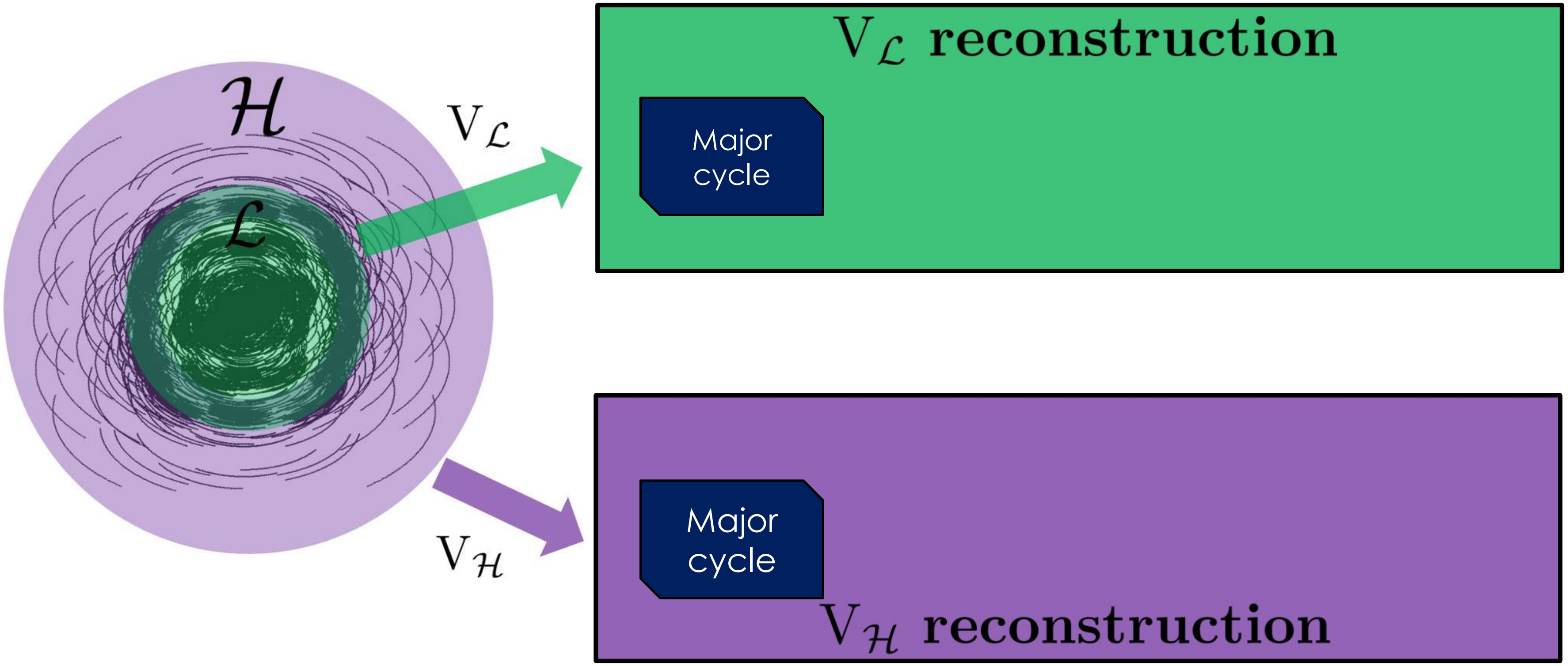




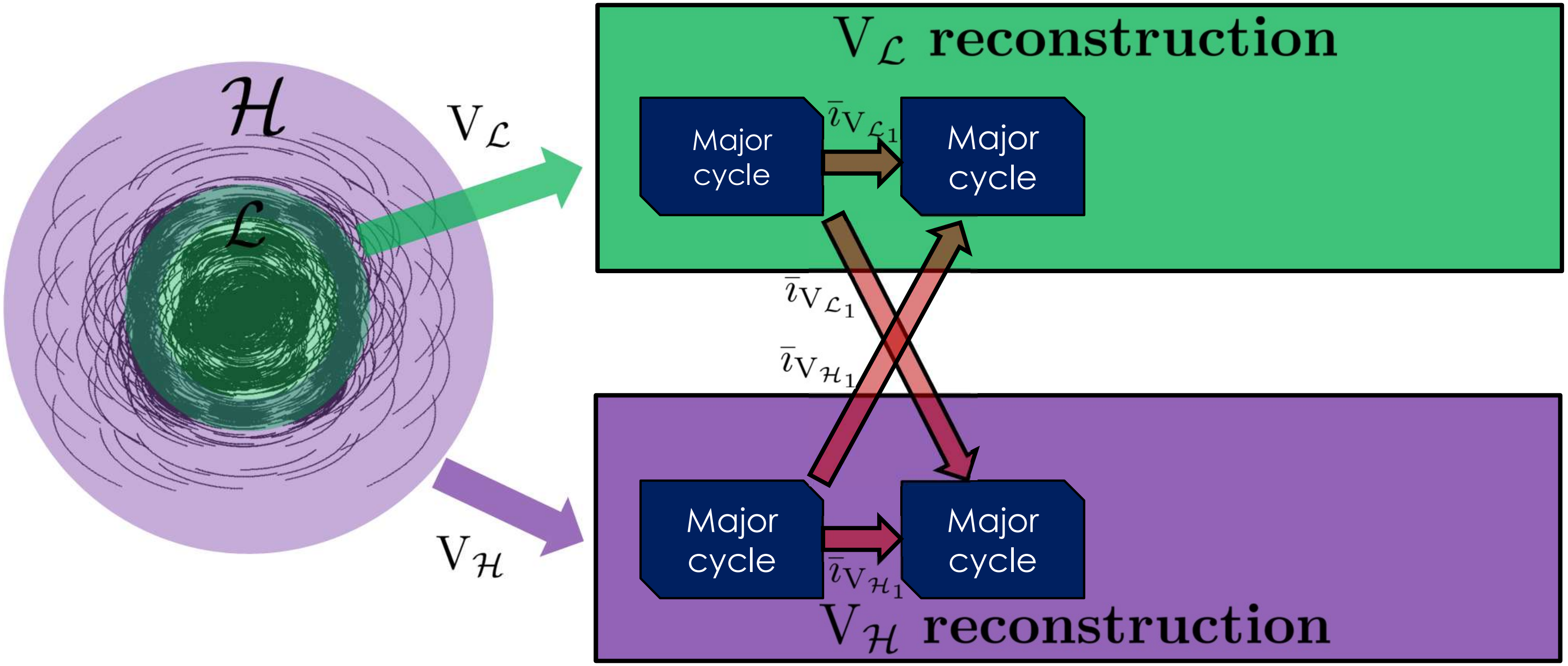
Decentralized Framework for 2 partitions



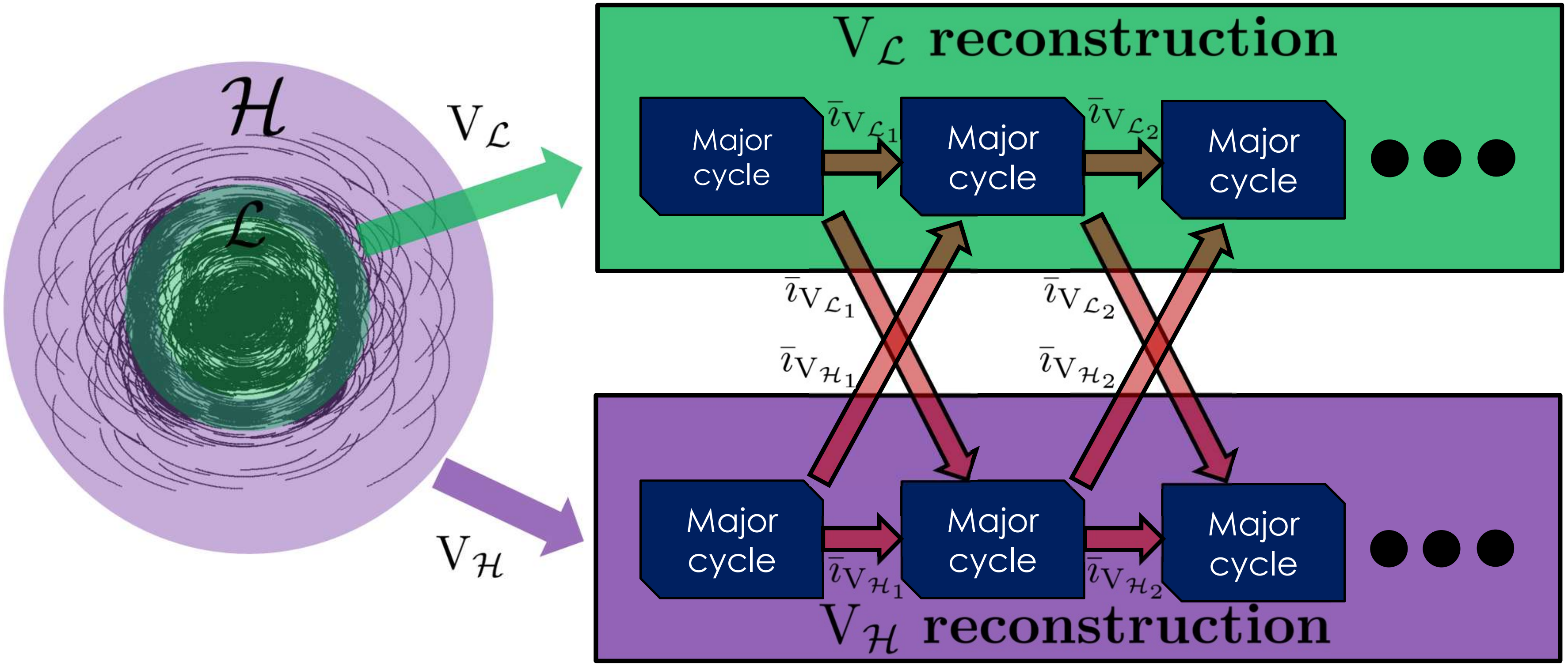
Decentralized Framework for 2 partitions



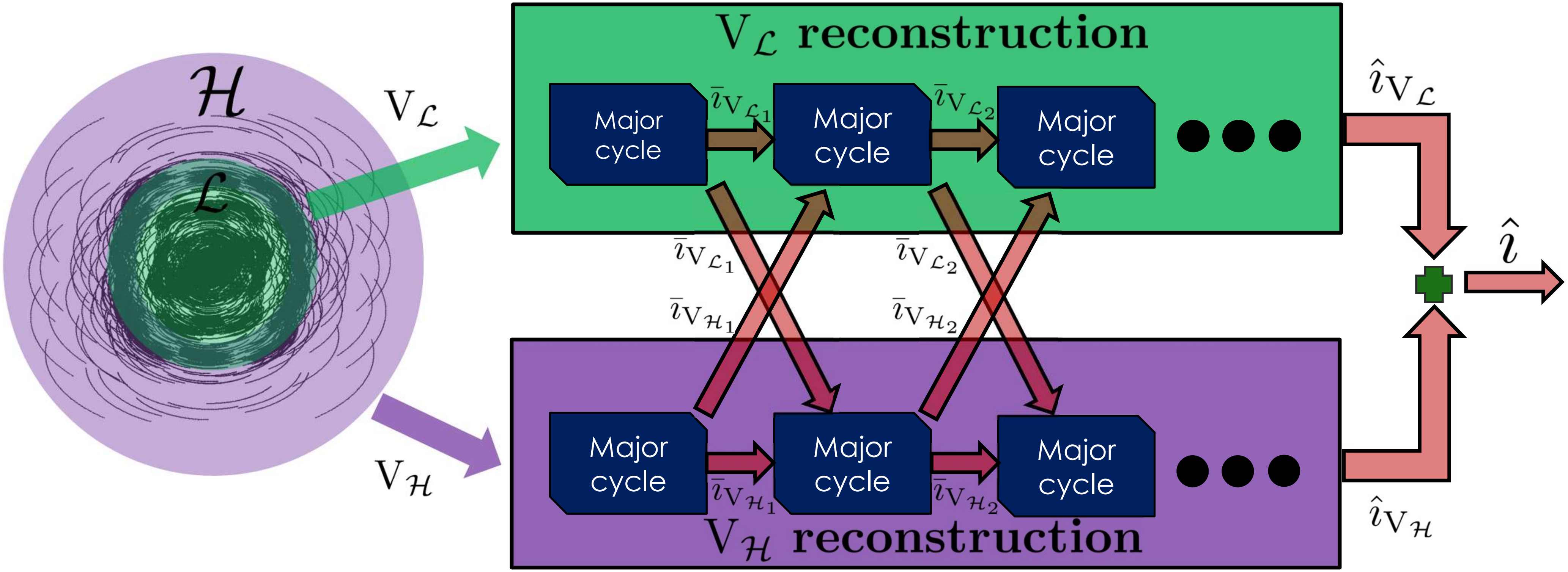
Decentralized Framework for 2 partitions



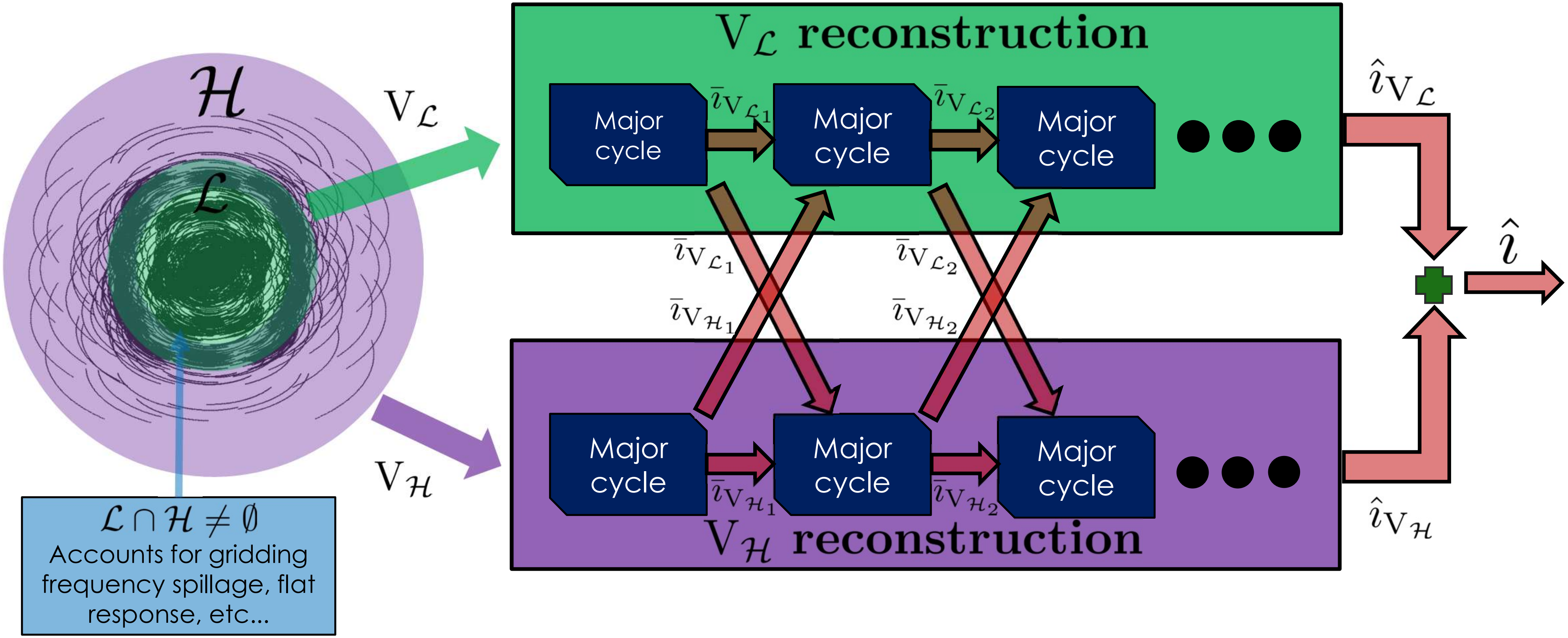
Decentralized Framework for 2 partitions



Decentralized Framework for 2 partitions



Decentralized Framework for 2 partitions



Example 1: Parallelized L1 reconstruction

Convolution by PSF operator

Deconvolution framework for every major cycle n, similar to [1, 2]

$$\alpha_n = \arg\min_{\alpha} \|\tilde{\imath}_n - HW\alpha\|_2^2 + \lambda_n \|\alpha\|_1$$

$$\bar{\imath}_n = W\alpha_n$$

Wavelet transform operator (Daubechies 1-8)

Used over $\|v - G^{\dagger}FW\alpha\|_2^2$ for efficiency
Assumes $Hi \approx F^{\dagger}GG^{\dagger}Fi$
Errors corrected in major cycle

[1] Wiaux, Yves, et al. "Compressed sensing imaging techniques for radio interferometry." Monthly Notices of the Royal Astronomical Society 395.3 (2009): 1733-1742.
[2] Garsden, Hugh, et al. "LOFAR sparse image reconstruction." Astronomy & astrophysics 575 (2015): A90.

Example 1: Parallelized L1 reconstruction

Deconvolution framework for every major cycle n, similar to [1, 2]

$$\alpha_n = \arg\min_{\alpha} \|\tilde{\imath}_n - HW\alpha\|_2^2 + \lambda_n \|\alpha\|_1$$

$$\bar{\imath}_n = W\alpha_n$$

Decentralized L1 deconvolution for each node j

$$\alpha^n_{\mathrm{V}_j} = \arg\min_{\alpha} \|\Gamma_{\mathcal{L}}(\tilde{\imath}^n_j - H_j W\alpha)\|_2^2 + \lambda^n_{\mathrm{V}_j}\|\alpha\|_1 + \gamma_n \sum_{k=0, k\neq j}^{K} \|\rho_k^{n-1} - \Gamma_k W\alpha\|_2^2,$$

$$\bar{\imath}^n_{\mathrm{V}_j} = W\alpha^n_{\mathrm{V}_j}, \rho_k^{n-1} = \sum_{i=1}^{n-1} \Gamma_k \bar{\imath}^i_{\mathrm{V}_k} - \Gamma_k \sum_{i=1}^{n-1} \bar{\imath}^i_{\mathrm{V}_j},$$

$$\gamma_n = 0 \text{ if } n = 1 \text{ and } \gamma_n = 1 \text{ otherwise}$$

Data fidelity term from local visibilities

Filters for ensuring spatial frequency locality, inverse variance weighting, and

Additional data fidelity term for received images. Acts as surrogate for missing visibilities.

[1] Wiaux, Yves, et al. "Compressed sensing imaging techniques for radio interferometry." Monthly Notices of the Royal Astronomical Society 395.3 (2009): 1733-1742.
[2] Garsden, Hugh, et al. "LOFAR sparse image reconstruction." Astronomy & astrophysics 575 (2015): A90.

Example 2: Parallelized MS-CLEAN reconstruction

CLEAN iteratively removes the brightest source at the most relevant scale convolved by the PSF from the residual[1]. We can denote this as:

$$\bar{\imath}^n = \text{ms-CLEAN}(\tilde{\imath}^n, H, \mathcal{S}, K)$$

Dirty

PSF

Scales

Iterations

[1] [2]Cornwell, Tim J. "Multiscale CLEAN deconvolution of radio synthesis images." *IEEE Journal of selected topics in signal processing* 2.5 (2008): 793-801.

Example 2: Parallelized MS-CLEAN reconstruction

CLEAN iteratively removes the brightest source at the most relevant scale convolved by the PSF from the residual[1]. We can denote this as:

$$\bar{\imath}^n = \text{ms-CLEAN}(\tilde{\imath}^n, H, \mathcal{S}, K)$$

Decentralized ms-CLEAN deconvolution for each node *j*

$$\bar{\imath}^n_{V_j} = \text{ms-CLEAN}(\tilde{\imath}^n_{j\cup}, H_{j\cup}, S^j, K),$$

$$\tilde{\imath}^n_{j\cup} = \mu_j \Gamma_j \tilde{\imath}^n_j + \sum_{k=0,k\neq j}^{K} \mu_k H_k \rho_k^{n-1},$$

$$H_{j\cup} = \mu_j \Gamma_j H_j + \sum_{k=0,k\neq j}^{K} \mu_k \Gamma_k H_k$$

$$\rho_k^{n-1} = \sum_{j=1}^{n-1} \Gamma_{\mathcal{H}} \bar{\imath}^j_{V_{\mathcal{H}}} - \Gamma_{\mathcal{H}} \sum_{j=1}^{n-1} \bar{\imath}^j_{V_{\mathcal{L}}}$$

Pseudo full-resolution dirty

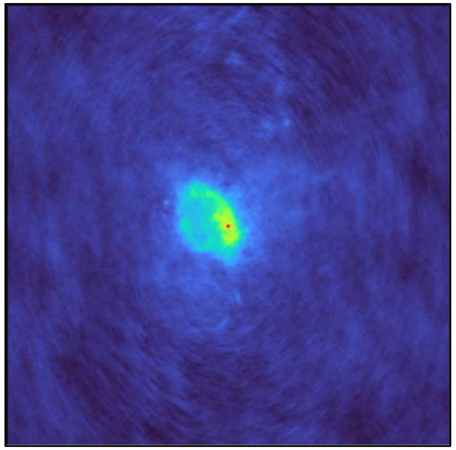
Pseudo full-resolution PSF

[1] [2]Cornwell, Tim J. "Multiscale CLEAN deconvolution of radio synthesis images." *IEEE Journal of selected topics in signal processing* 2.5 (2008): 793-801.

Experiment Datasets

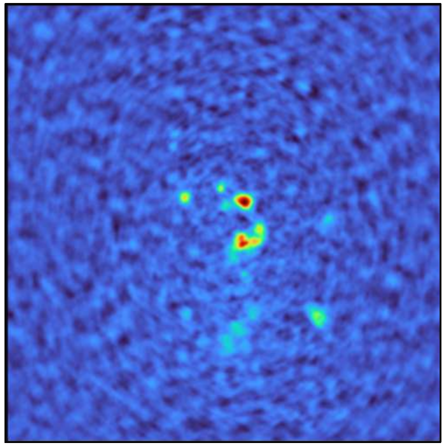
Sgr A

- Telescope Config: SKA-MID AA4
- Observation time: HA=[-0.5,0.5]
- Integration time per vis: 5s
- EM frequency: 1GHz – 1.00064GHz
- Frequency channels: 64
- Channel bandwidth: 10KHz
- Total vis (with autocorr): 898698240
- Noise: 0.05 sigma of signal
- Pixel resolution: 512 x 512



Simulation Process:

- Initial images tapered and cutout from 1.28GHz MeerKAT mosaic of the galactic plane[1]
- Visibility positions generated from observation parameters
- Images degridded using RASCIL (with wgridder) to visibilities to obtain values
- Noise added to visibilities
- Pseudo RA-DEC coordinates for phase-centers



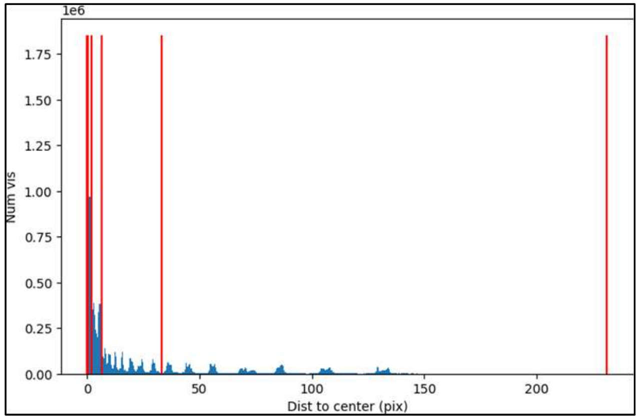
Sgr B2

- Telescope Config: SKA-LOW AA4
- Observation time: HA=[-0.25,0.25]
- Integration time per vis: 5s
- EM frequency: 200MHz – 200.2MHz
- Frequency channels: 20
- Channel bandwidth: 10KHz
- Total vis (with autocorr): 945561600
- Noise: 0.05 sigma of signal
- Pixel resolution: 512 x 512

Partitioning and Baseline dependent averaging

Partitioning initial dataset

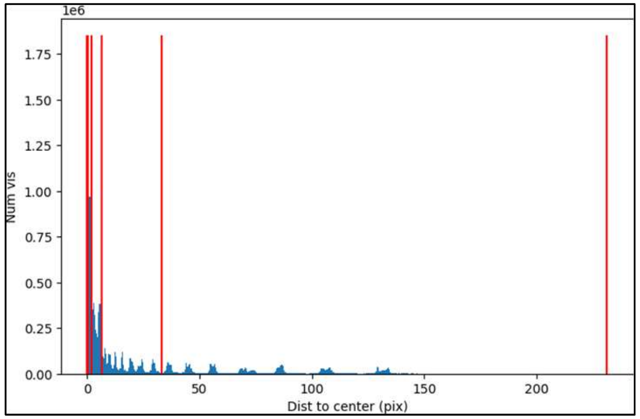
- Visibilities partitioned by sampling inverse CDF at fixed intervals determined by number of partitions
  - Does not account for overlap regions
  - Not optimal but maybe good enough
- Datasets difficult to partition
  - High density short baseline visibilities
  - Can have lots of overlapping visibility partitions
  - Uneven amounts of spatial frequency information



Partitioning and Baseline dependent averaging

Partitioning initial dataset

- Visibilities partitioned by sampling inverse CDF at fixed intervals determined by number of partitions
  - Does not account for overlap regions
  - Not optimal but maybe good enough
- Datasets difficult to partition
  - High density short baseline visibilities
  - Can have lots of overlapping visibility partitions
  - Uneven amounts of spatial frequency information



Baseline dependent averaging

- Averages based on some decorrelation threshold. For this we use the product of the time and frequency decorrelations:

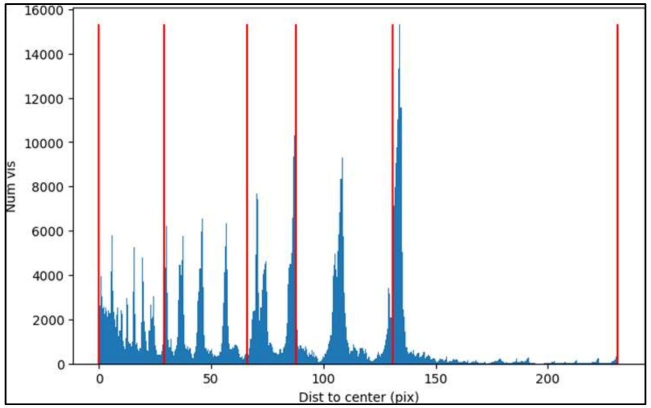
$$\rho = \rho_f \times \rho_t$$

$$\rho_f = \mathrm{sinc}\left(\frac{\pi \nu_\Delta \tau_g}{2}\right)$$

$$\rho_t = \mathrm{sinc}\left(\pi T\left(\frac{\mathrm{d}u}{\mathrm{d}t}l + \frac{\mathrm{d}v}{\mathrm{d}t}m + \frac{\mathrm{d}w}{\mathrm{d}t}(n-1)\right)\right)$$

$$\approx 1 - \frac{\pi^2 T^2}{6}\left(\frac{\mathrm{d}u}{\mathrm{d}t}l + \frac{\mathrm{d}v}{\mathrm{d}t}m + \frac{\mathrm{d}w}{\mathrm{d}t}(n-1)\right)^2$$

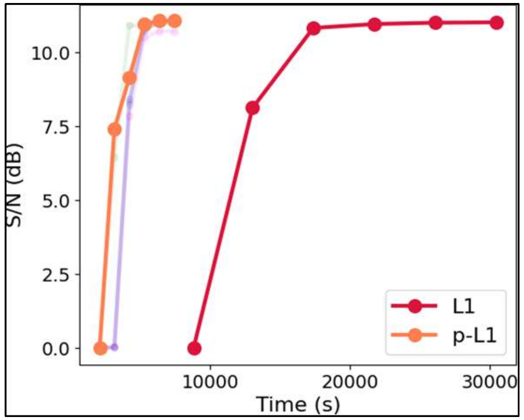
- Averaging done on the time domain (in power 2 levels) so Taylor approximation is used here for invertibility.
- Flattens visibility density distribution, allowing for much better partitioning

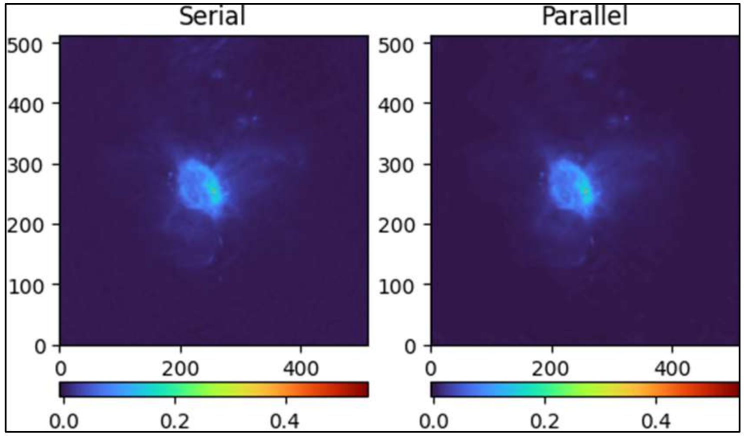


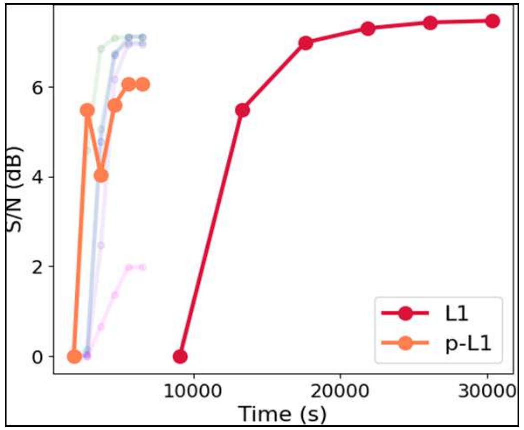
Preliminary results – Time and Accuracy for L1 reconstruction

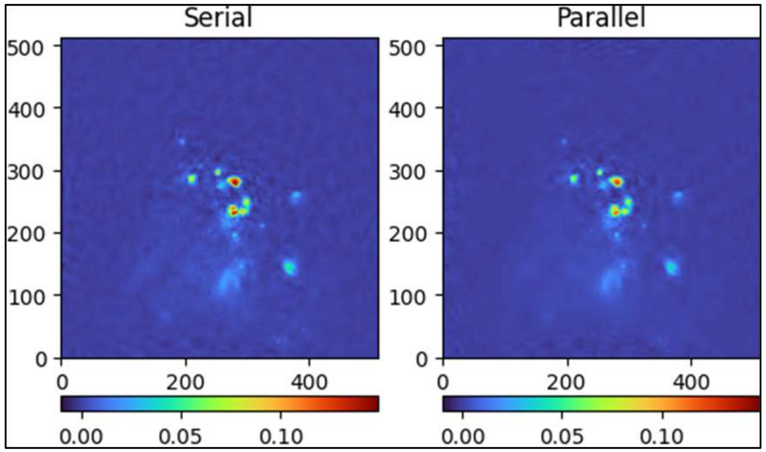
Summary

- Results run on a cpu cluster (Jean Zay cpu_p1)
- Works well for Sgr A dataset, less well for Sgr B2 dataset due to one node performing a reconstruction that lacks flux
  - Not sure yet of the cause, may be due to the S/N of the initial visiblities but need to investigate more
- Promising acceleration from parallelization
  - 4.12x from Sgr A dataset
  - 4.65x from Sgr B2 dataset



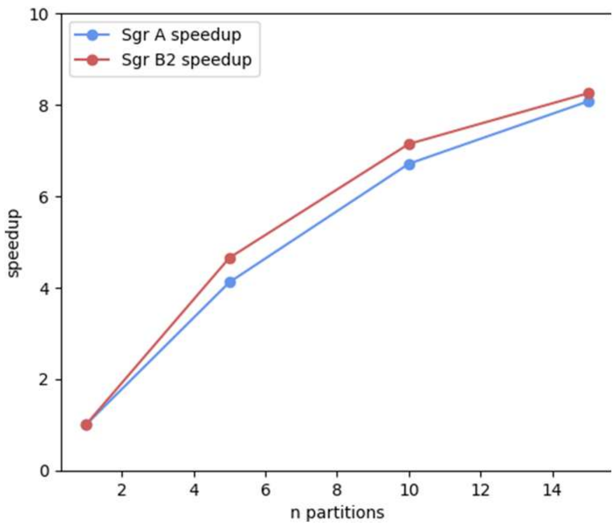






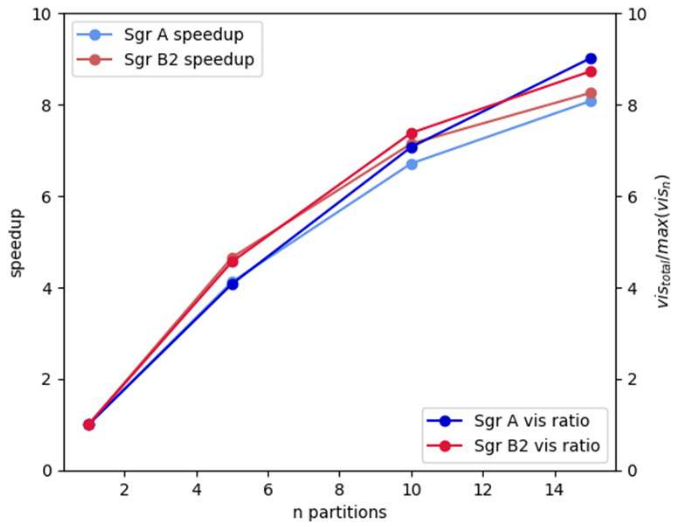
Preliminary results – Scaling when increasing parallelization

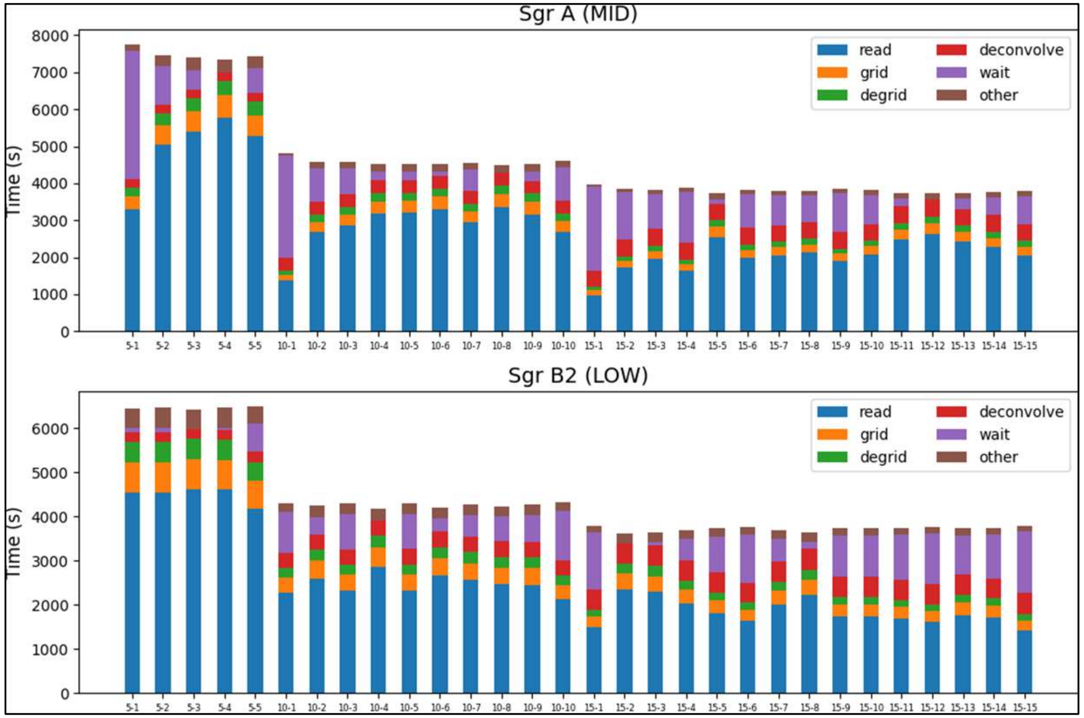
- Scaling becomes worse with larger number of nodes, getting a speedup of a little over 8x for 15 partitions
- Largely due to non-ideal load balancing and visibility duplication from transition regions
- Can improve with better partitioning



Preliminary results – Scaling when increasing parallelization

- Scaling becomes worse with larger number of nodes, getting a speedup of a little over 8x for 15 partitions
- Largely due to non-ideal load balancing and visibility duplication from transition regions
- Can improve with better partitioning





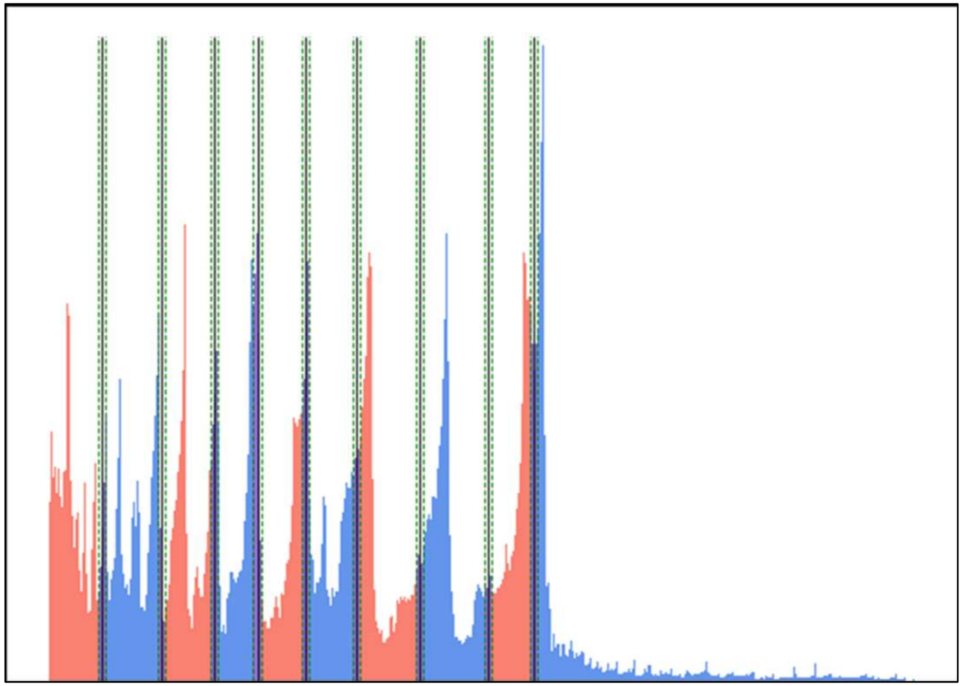
Improving load balancing

Ideal partitioning

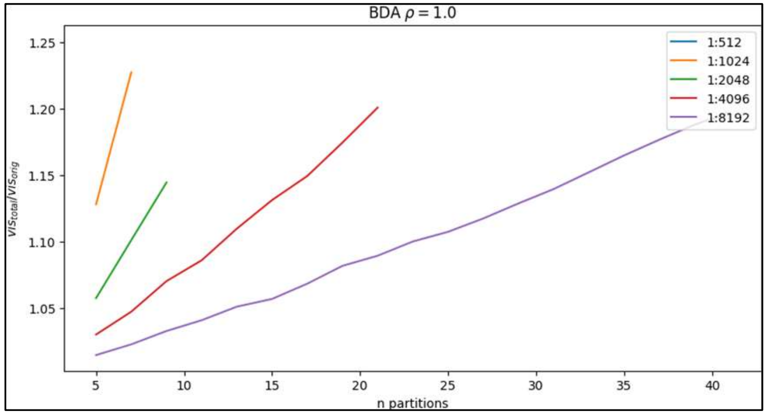
- Recently developed a better method for load balancing
- Finds perfectly load balanced configurations as long as a solution exists and numerical precision allows
- In the ideal case, the partitions satisfies the following:

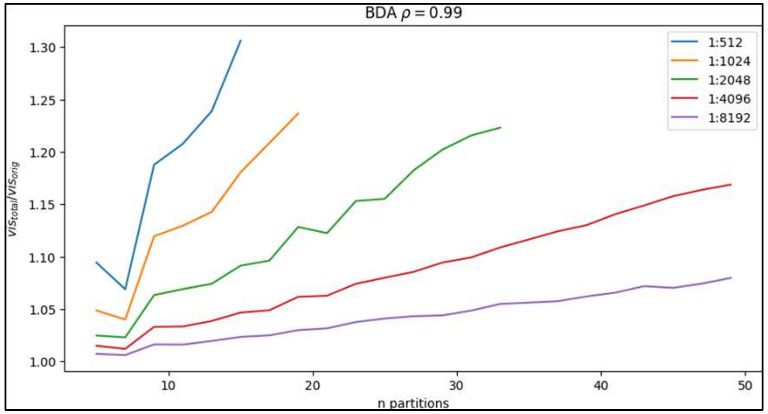
$$\mathcal{C}(\ell_1 + \delta) = \alpha$$
$$\mathcal{C}(\ell_n + \delta) - \mathcal{C}(\ell_{n-1} - \delta) = \alpha, n \in \{2, 3, ..., N-1\}$$
$$1 - \mathcal{C}(\ell_{N-1} - \delta) = \alpha$$

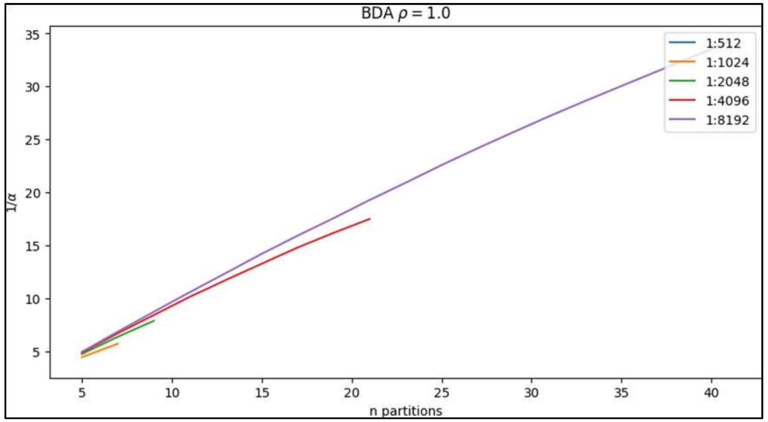
- Can compute the other parameters if we have α, so only need to find the root of the last equation
- There isn't always a solution

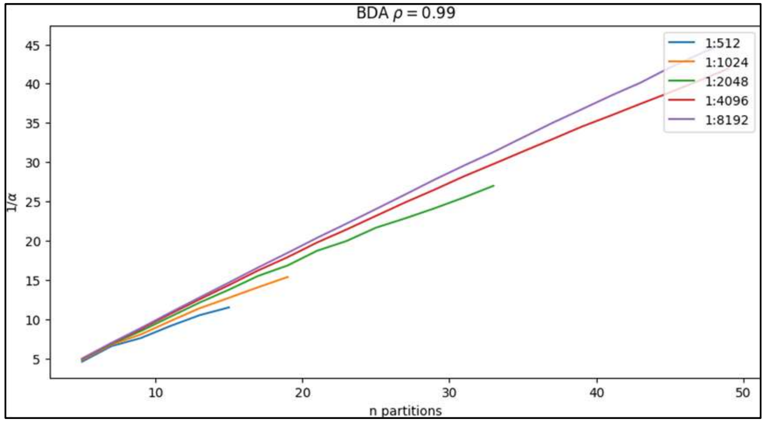


Theoretical scaling









Summary

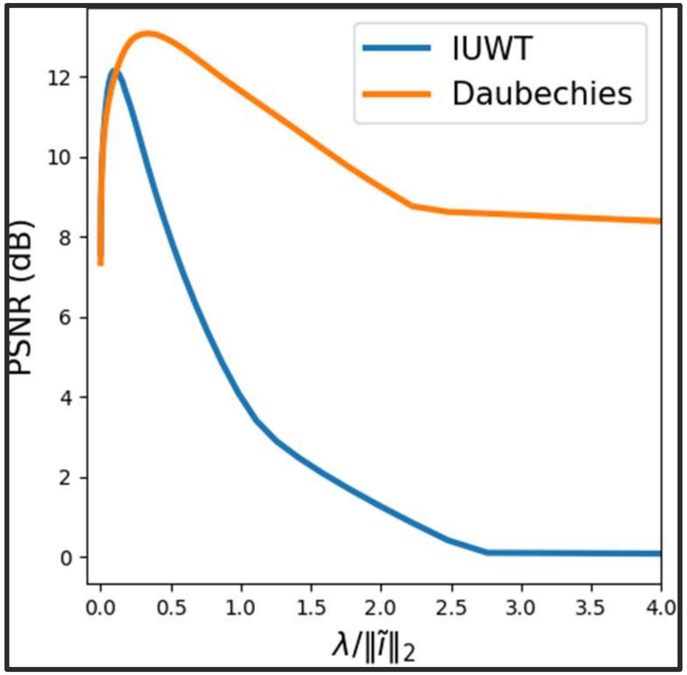
- Can use the improved load balancing method to find theoretical speedups
- Dependent on a variety of factors
  - Transition region relative to pixel resolution
  - BDA level (to a certain extent)
  - Telescope
- Visibility duplication maxes out at roughly 25% before no solution is found

Future work

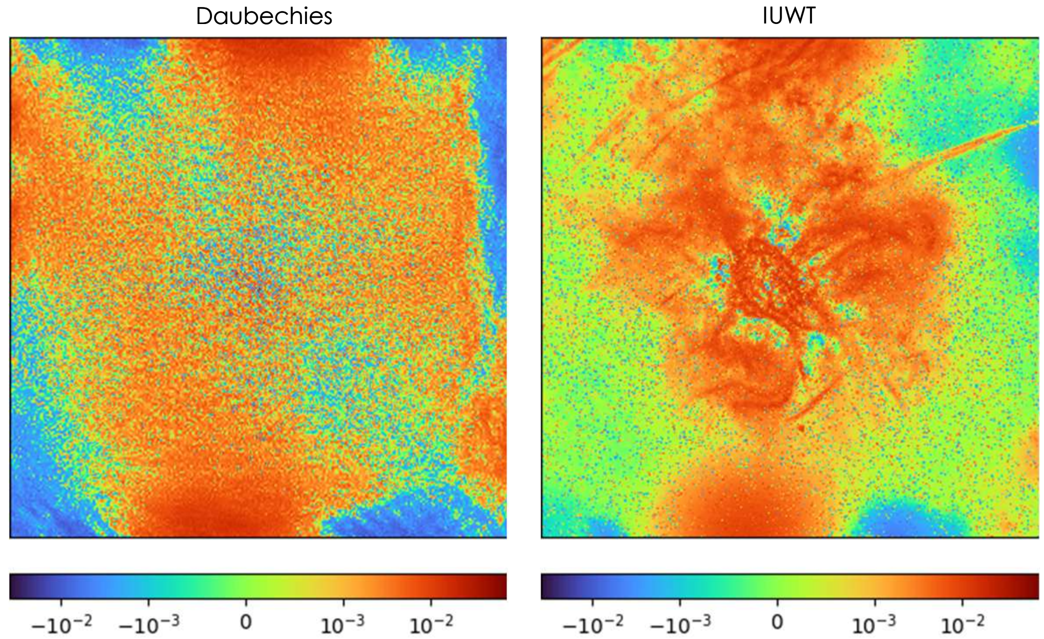
- Obtain full experimental results with improved load balancing
- More realistic datasets
  - Longer observation times and more channels. Need for efficient de/gridding and I/O to achieve this as number of visibilities can easily balloon to trillions if not more.
  - More realistic image sizes. Pixel resolutions for single pointings for SKA-Mid and SKA-Low are estimated at around 20k x 20k and 4k x 4k, respectively. Can be larger if mosaicing.
    - Needs efficient and scalable de/gridding and deconvolution algorithms.
- Evaluation metrics for convergence
  - S/N is not really a good measurement for real datasets as there is no ground truth
  - Statistical tests are doable but expensive
  - Science dependent
  - Possible use-case for in-situ tools

Appendices

IUWT vs Daubechies

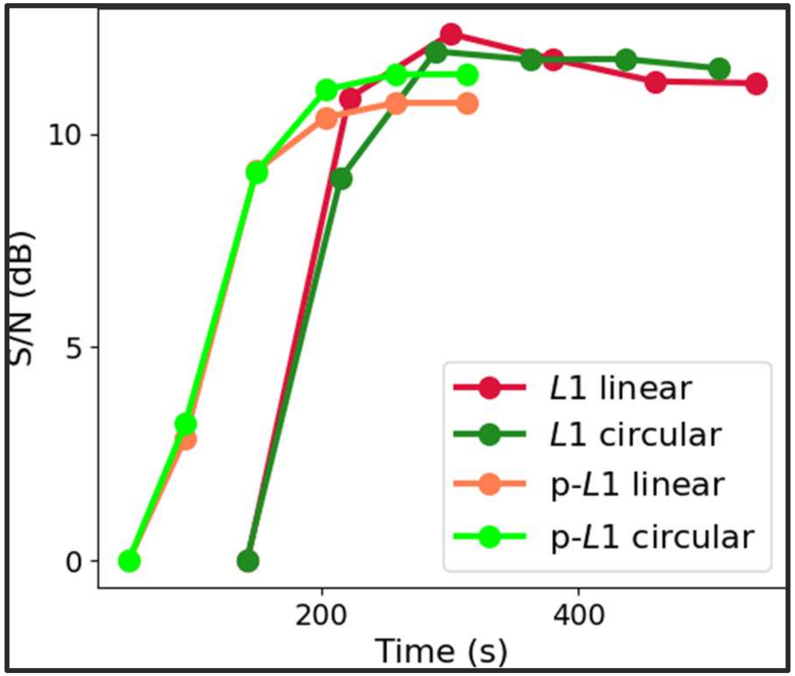


First major-cycle residuals for Sgr A



IUWT seems worse at reconstructing large-scale anisotropic extended emissions.

Linear vs circular convolution



- Using linear convolution instead of circular can be desired so that bright extended sources don't wrap around.
- More complicated to find the step size as operator does not diagonalize with Fourier transform, have to rely on something like power iteration.
- Results don't necessarily seem better as shown on the left, could be due to sources, will need more testing.
- More expensive to compute although it doesn't seem to make much difference in the grand scheme.

Selection of λ

Inspired by the work of [1], for the problem:

$$\|G_j(\tilde{\imath} - HW\alpha)\|^2 + \sum_{\substack{i=1 \\ i \neq j}} \|G_iC_i - W\alpha)\|^2 + \lambda\|W\alpha\|_1$$

We use:

$$\lambda_n = \eta_n \lambda_{max_n}$$

$$\eta_n = \alpha + (1-\alpha)\frac{e^{\beta t_n} - 1}{e^{\beta} - 1}$$

$$t_n = \frac{n}{N-1}$$

$$\lambda_{max_n} = 2\|W^\dagger(H_j^\dagger G_j^\dagger G_j \tilde{\imath}_{n_j} + \sum_{i=1, i\neq j} G_i^\dagger G_i C_{n_i})\|_\infty$$

Which is the upper-bound for the regularization parameter for when the solution is non-zero.

Filters

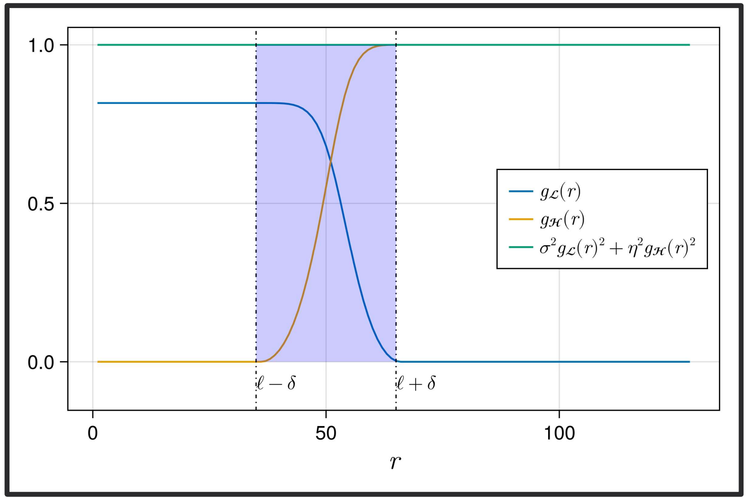
$$r > \ell + \delta : \quad |g_{\mathcal{H}}(r)|^2 = 1/\sigma^2, \; g_{\mathcal{L}}(u) = 0$$

$$r < \ell - \delta : \quad g_{\mathcal{H}}(r) = 0, \; |g_{\mathcal{L}}(r)|^2 = 1/\eta^2$$

$$\ell - \delta < r < \ell + \delta : \quad \sigma^2 |g_{\mathcal{H}}(r)|^2 + \eta^2 |g_{\mathcal{L}}(r)|^2 = 1$$

$$g_{\mathcal{L}}(r) = \alpha(r) \left(1 - \sin\left(\frac{\pi}{2\delta}(r - \ell)\right)\right)$$

$$g_{\mathcal{H}}(r) = \alpha(r) \left(1 + \sin\left(\frac{\pi}{2\delta}(r - \ell)\right)\right)$$



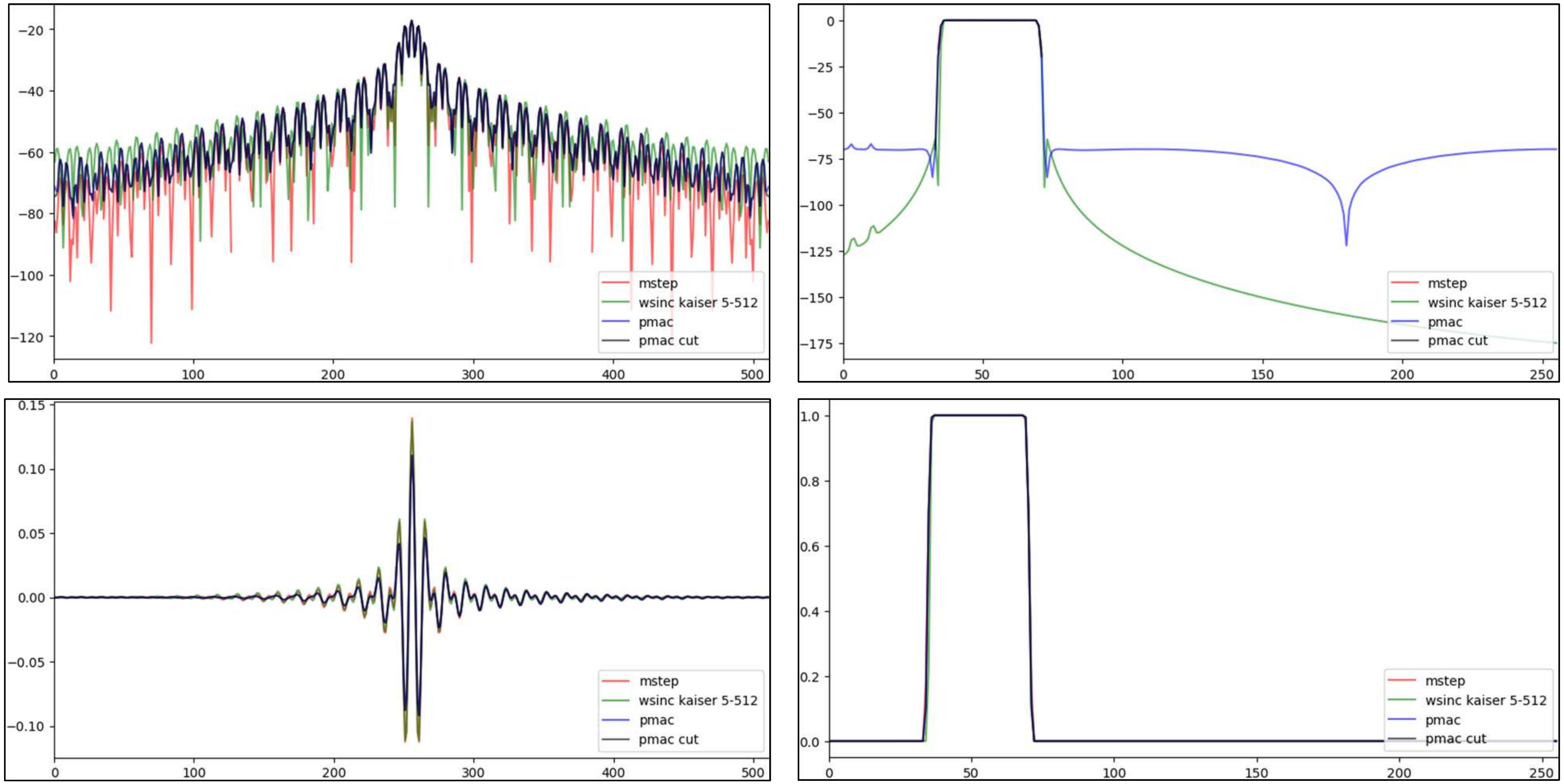
- 1-D filters used as distance in 2-D, resulting in an annulus
- Compared against more traditional methods such as windowed sinc and Parks-McClellan, not a large difference in image quality.
- Better for us due to not operating on a discrete grid

Filter comparison

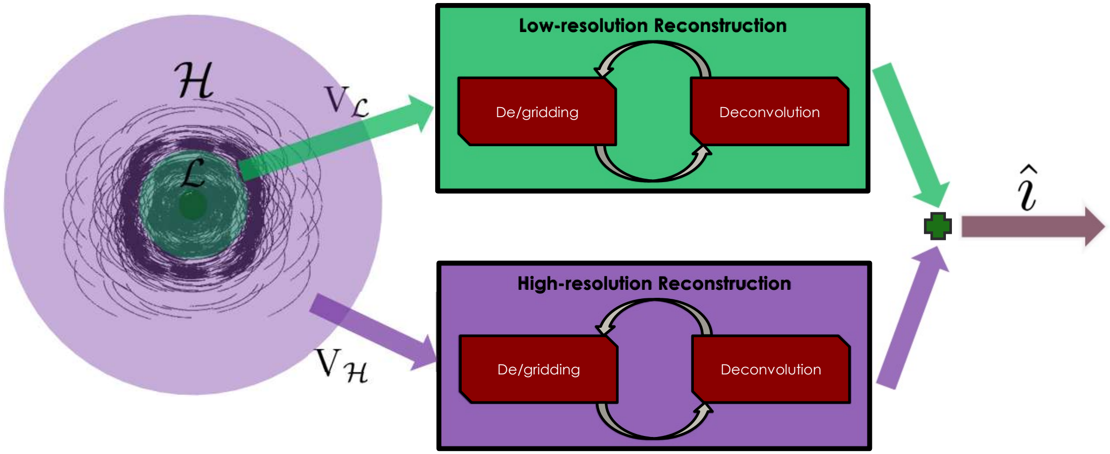
Impulse response | Frequency response

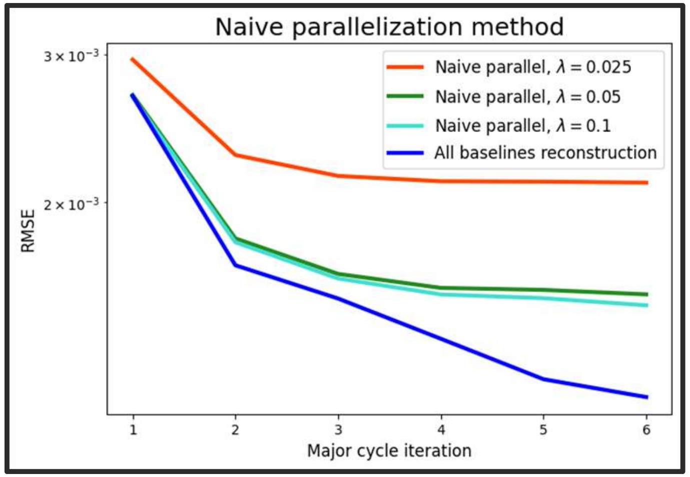
dB

power



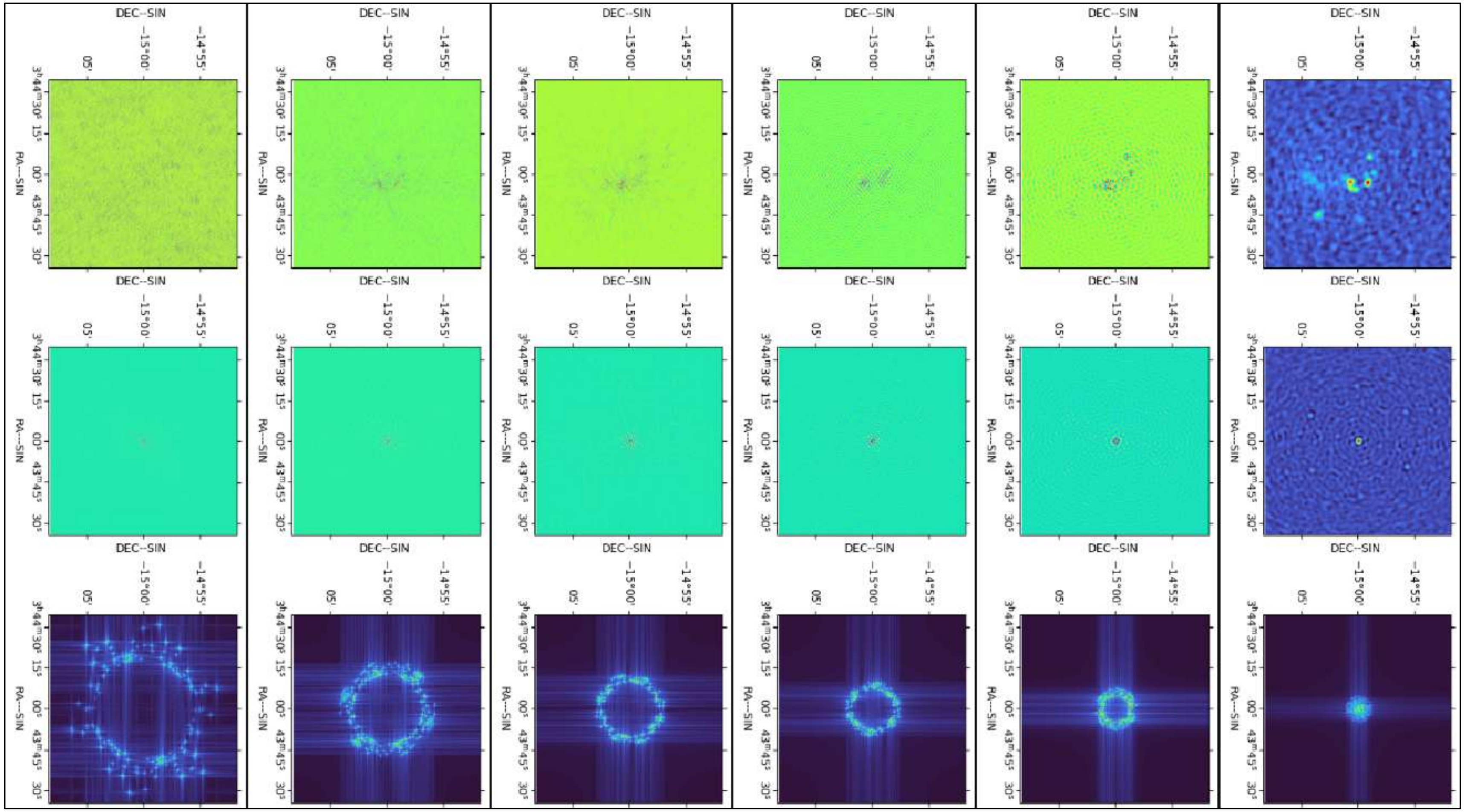
Naively adding separately deconvolved images





- Naïve parallel reconstructions seem always worse.
- Possibly due to terms not regularized together, which introduces some assumptions on sparsity.
- Can probably tune lambda so that the same result is obtained, but unclear how.

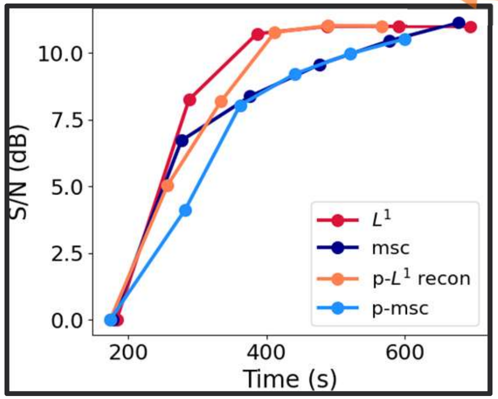
Partition visualization



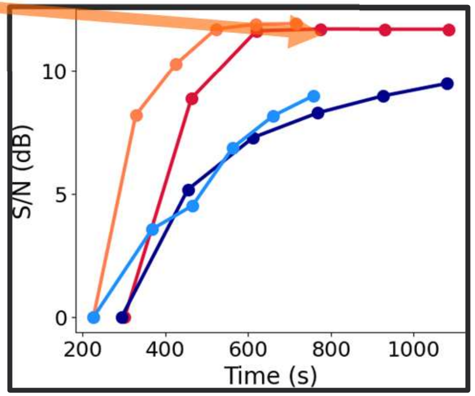
2 Partition results - Simulated

S/N obtained with comparison to ground truth

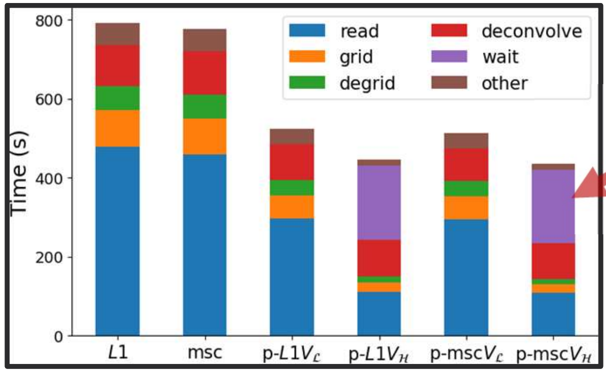
Sgr B2



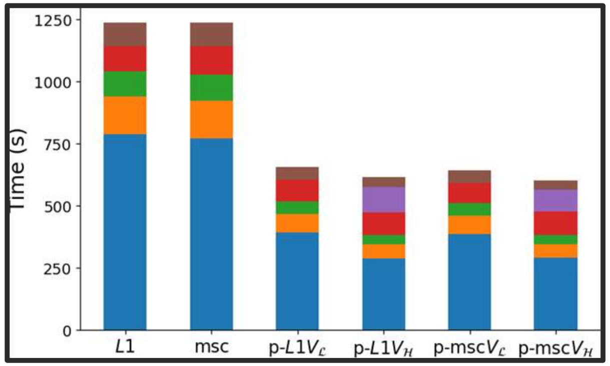
Sgr C



Poor parallelization for Sgr B2, much better for Sgr C. Mainly due to uneven partitioning.



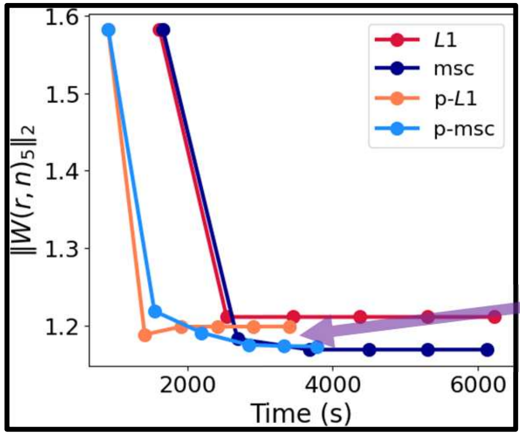
Poor partitioning causes a lot of idle time for node with less visibilities.



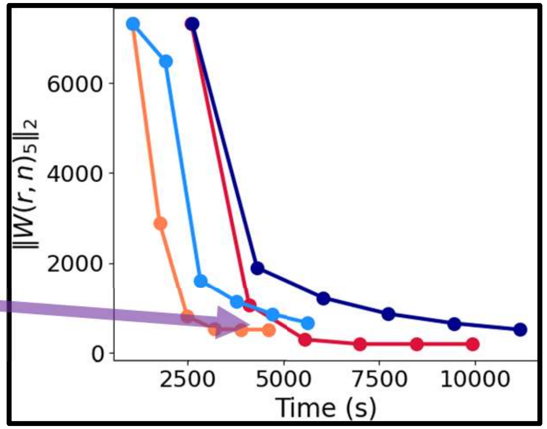
2 Partition results - Real

Wasserstein-1 distance between residual and ideal residual obtained via visibility negation. Computed per pixel (windowed), with L2 being plotted.

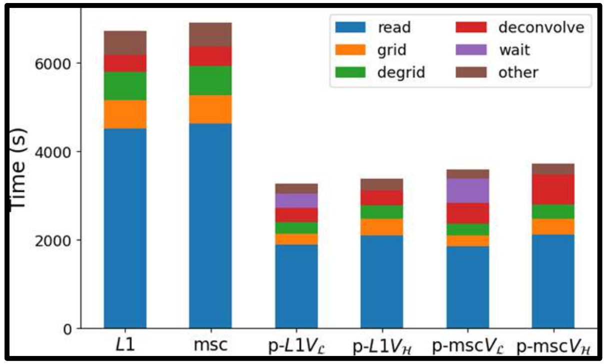
HL Tau



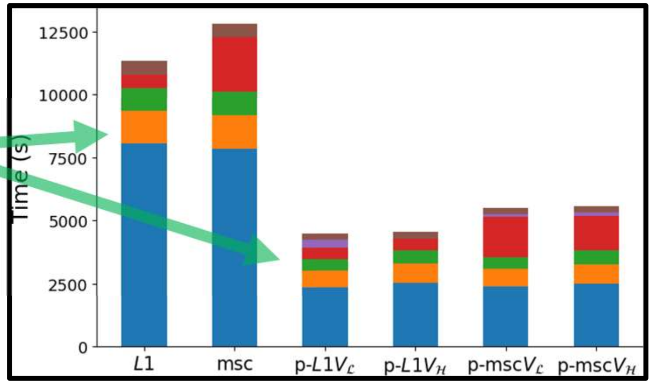
Cygnus A



Speedup much better for real datasets, close to optimal 2x



Over 2x speedup can be due to RASCIL overheads.



Results – Scaling to larger image sizes

Average processing times for Cygnus A dataset per major cycle

Alg.	Node	Pix. res.	Deconv.	Degrid	Grid	Disk I/O	Transf.	Other
p-msc	$V_{\mathcal{L}}$	1728 × 1728	354.70s	92.27s	117.21s	342.12s	0.01s	13.51s
	$V_{\mathcal{H}}$	1728 × 1728	288.83s	109.84s	129.62s	353.67s	0.01s	11.88s
	$V_{\mathcal{L}}$	10k × 10k	17778.44s (× 50)	1450.58s (× 16)	2519.88s (× 21)	358.31s (× 1)	0.45s (× 52)	21.35s (× 2)
	$V_{\mathcal{H}}$	10k × 10k	18014.80s (× 62)	2159.66s (× 20)	2533.84s (× 20)	362.50s (× 1)	0.67s (× 51)	21.11s (× 2)
p-L1	$V_{\mathcal{L}}$	1728 × 1728	91.66s	92.44s	111.52s	332.31s	0.02s	11.52s
	$V_{\mathcal{H}}$	1728 × 1728	89.33s	108.17s	126.85s	359.43s	0.02s	13.49s
	$V_{\mathcal{L}}$	10k × 10k	3595.75s (× 39)	1449.08s (× 16)	2457.80s (× 22)	365.53s (× 1)	0.59s (× 30)	20.16s (× 2)
	$V_{\mathcal{H}}$	10k × 10k	3573.73s (× 40)	2173.30s (× 20)	2555.44s (× 20)	363.62s (× 1)	0.60s (× 25)	20.22s (× 1)

Primary bottlenecks seem to be deconvolution and de/gridding (to a lesser extent).

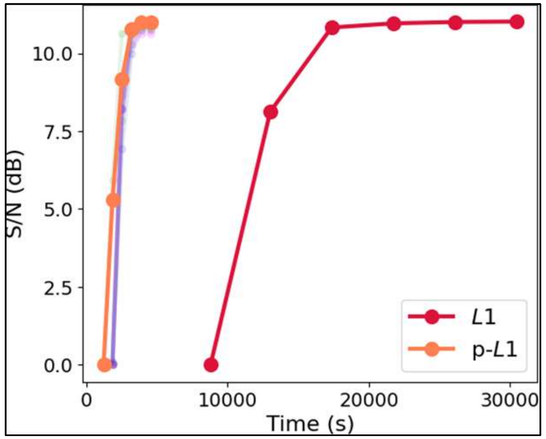
Transfer time also increases similarly to deconvolution, but cost negligible. Even for 100kx100k images, with the current cost increases, a transfer only takes ~72s which is substantially less than even the 10kx10k deconvolution.

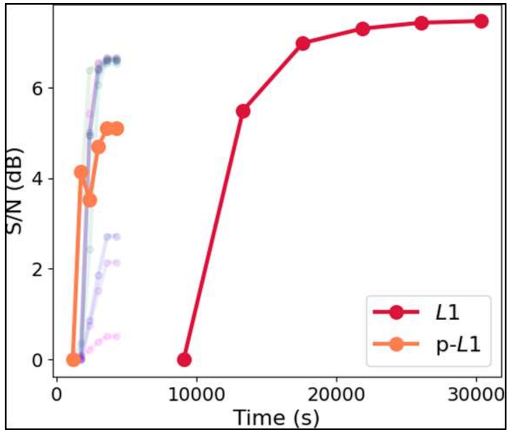
Results for 10 and 15 partitions

Sgr A

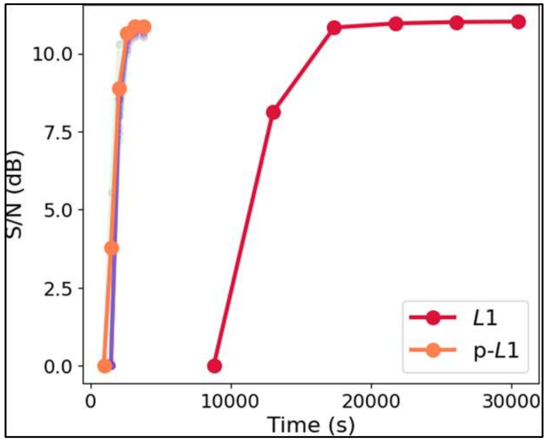
Sgr B2

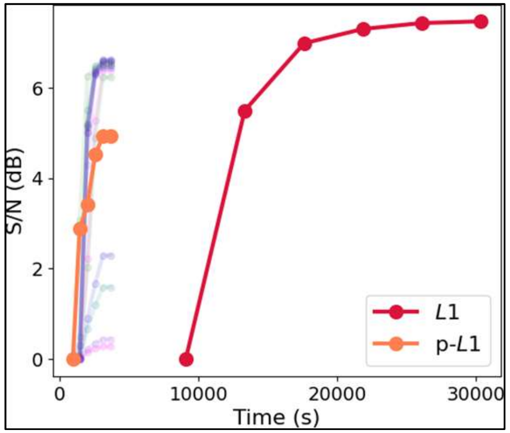
10 partitions



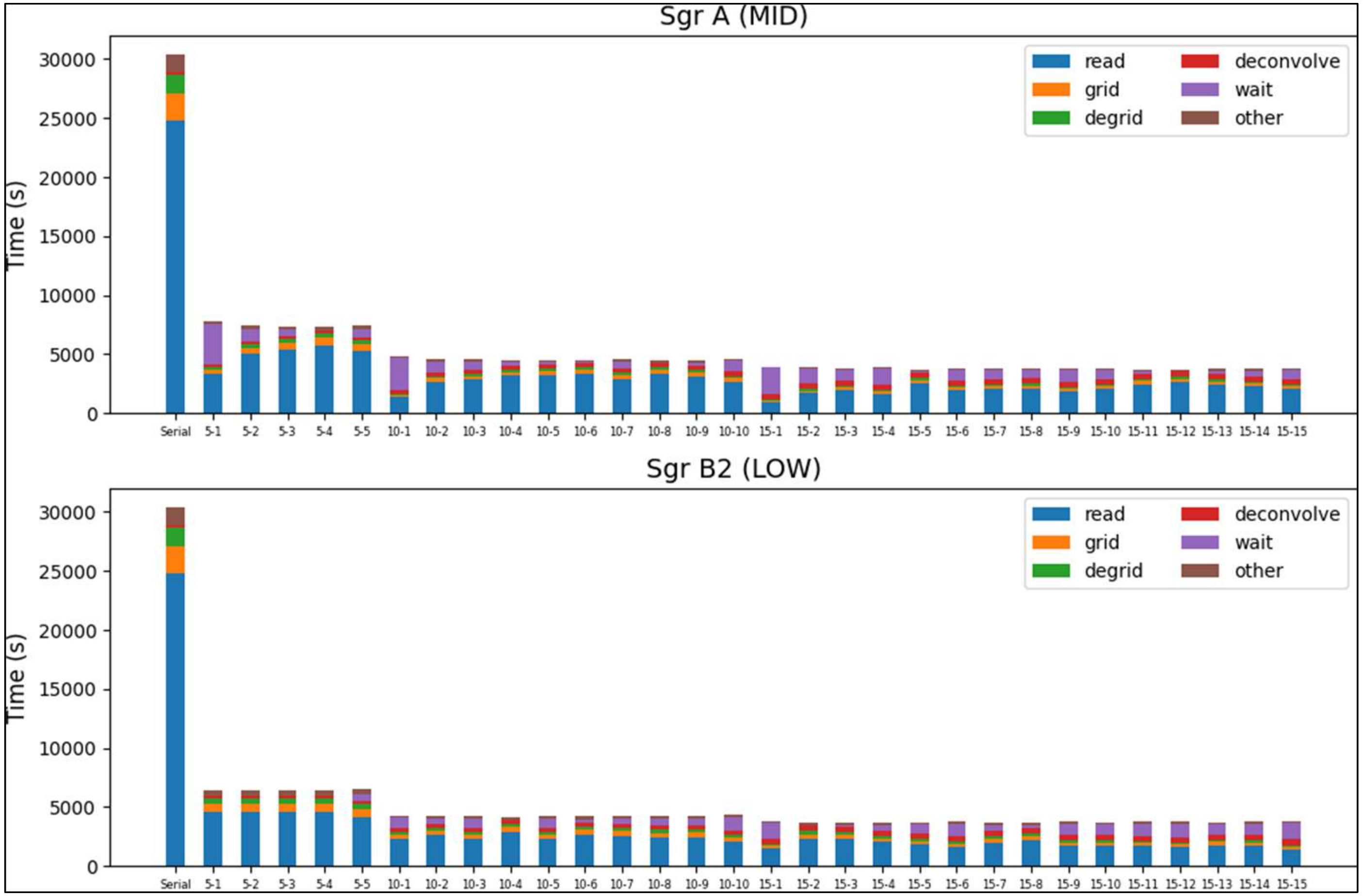


15 partitions

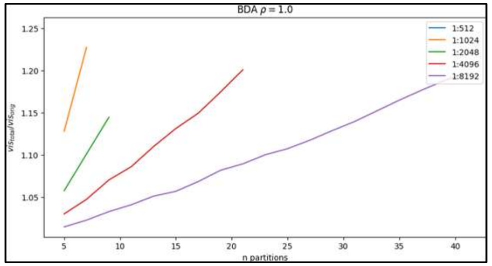


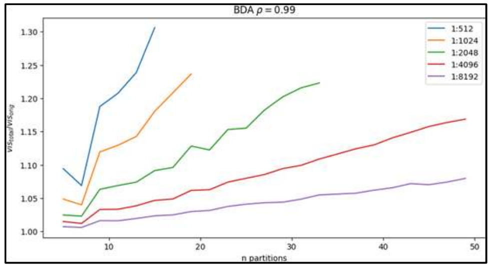


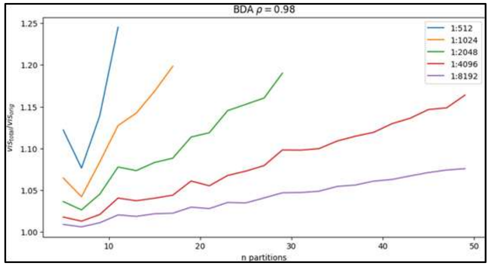
Breakdown with serial as well

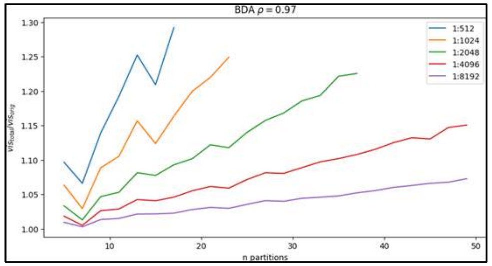


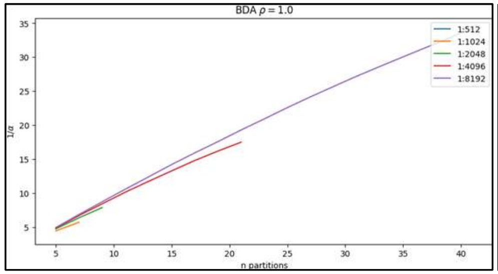
More theoretical scaling – LOW AA4

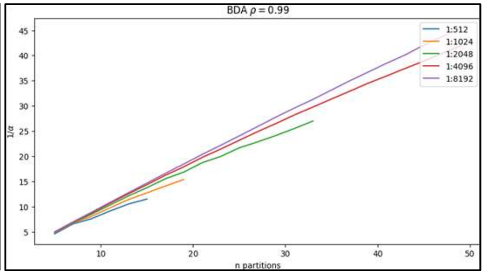


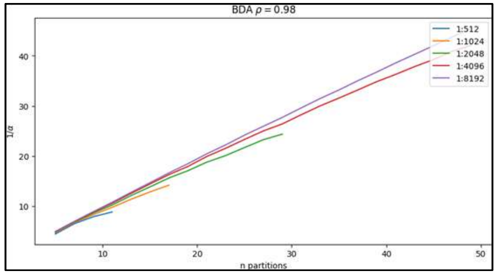


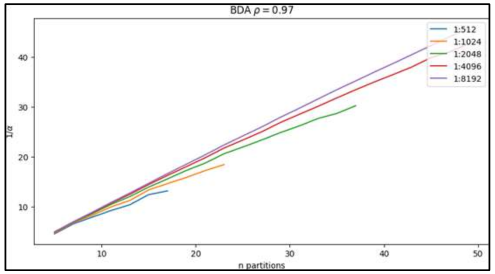












More theoretical scaling – MID AA4