GPU Computing on a New Frontier in Cosmology

Lincoln Greenhill *(Harvard / Smithsonian)*

Thanks to G. Bernardi M. Clark, R. Edgar,
D. Mitchell, S. Ord, R. Wayth

ICCS 2011

Outline

- the frontier - science drivers
- instruments - new architectures & tech. drivers
- tera-scale <u>real-time</u> signal processing w/ GPUs
  - computing resource at the instrument
  - example: MWA cal/im (Edgar et al. 2010, *CPC*)
- scaling (PETA-EXA)
  - instruments for the next decade^{+} are drivers
    - apps: x-correlation & calibration and imaging
    - data rates ↑, r/t processing increasingly required

L. Greenhill

The Cosmological Record

- The early IGM (largely H) is traced by the λ21cm transition
  - forbidden hyperfine transition: $1^2s_{1/2}$ state
  - T_{spin}
- λ21cm is a unique tracer: broad angular distribution; high-z signal
  - complements IR spectroscopy, imaging; cross-correlation

L. Greenhill

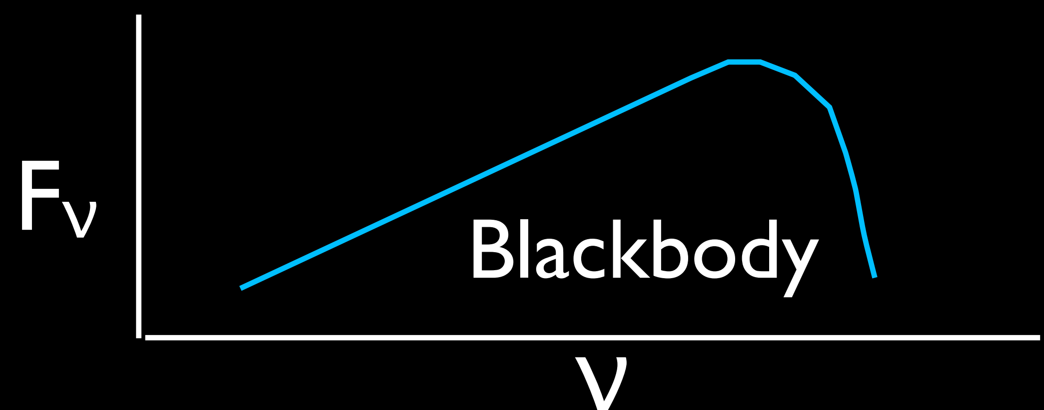
Science Goal

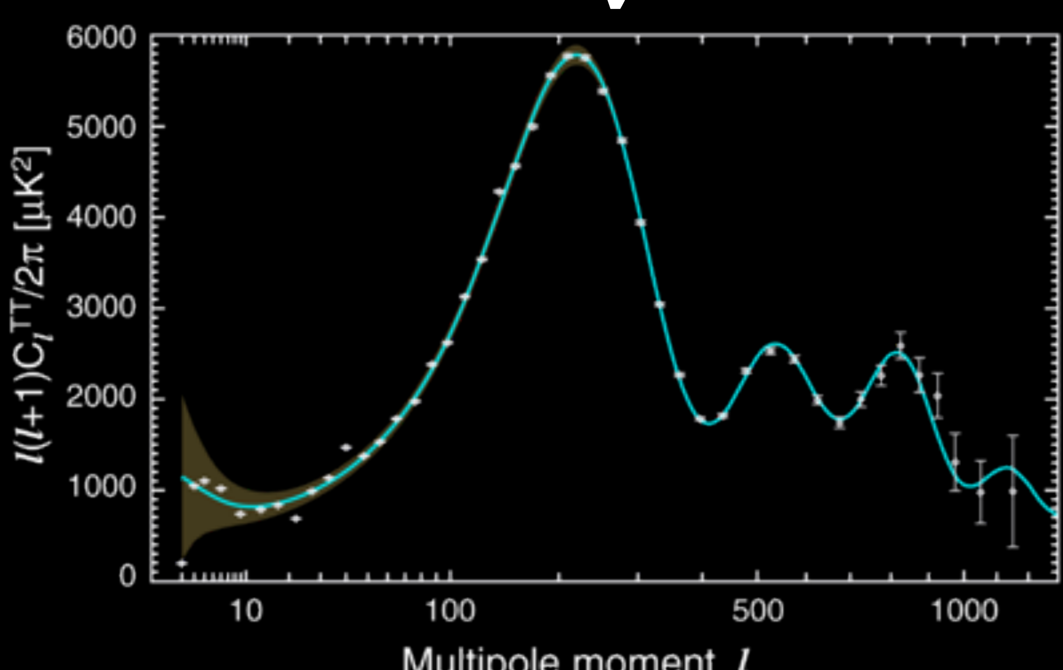
- characterization of IGM during the EOR ($6<z<30$)
  - frequency & angular power spectrum (near-term)
  - direct imaging (long-term)
- constrain evolution of early source populations, structure formation, perturbations, etc
- achieve sensitivity to unpolarized mK background
  - $O(10^{3-4})$ deg^2 in $O(10^3)$ hrs
  - difficult in view of foregrounds: 10^{5-6} x EOR signal

L. Greenhill

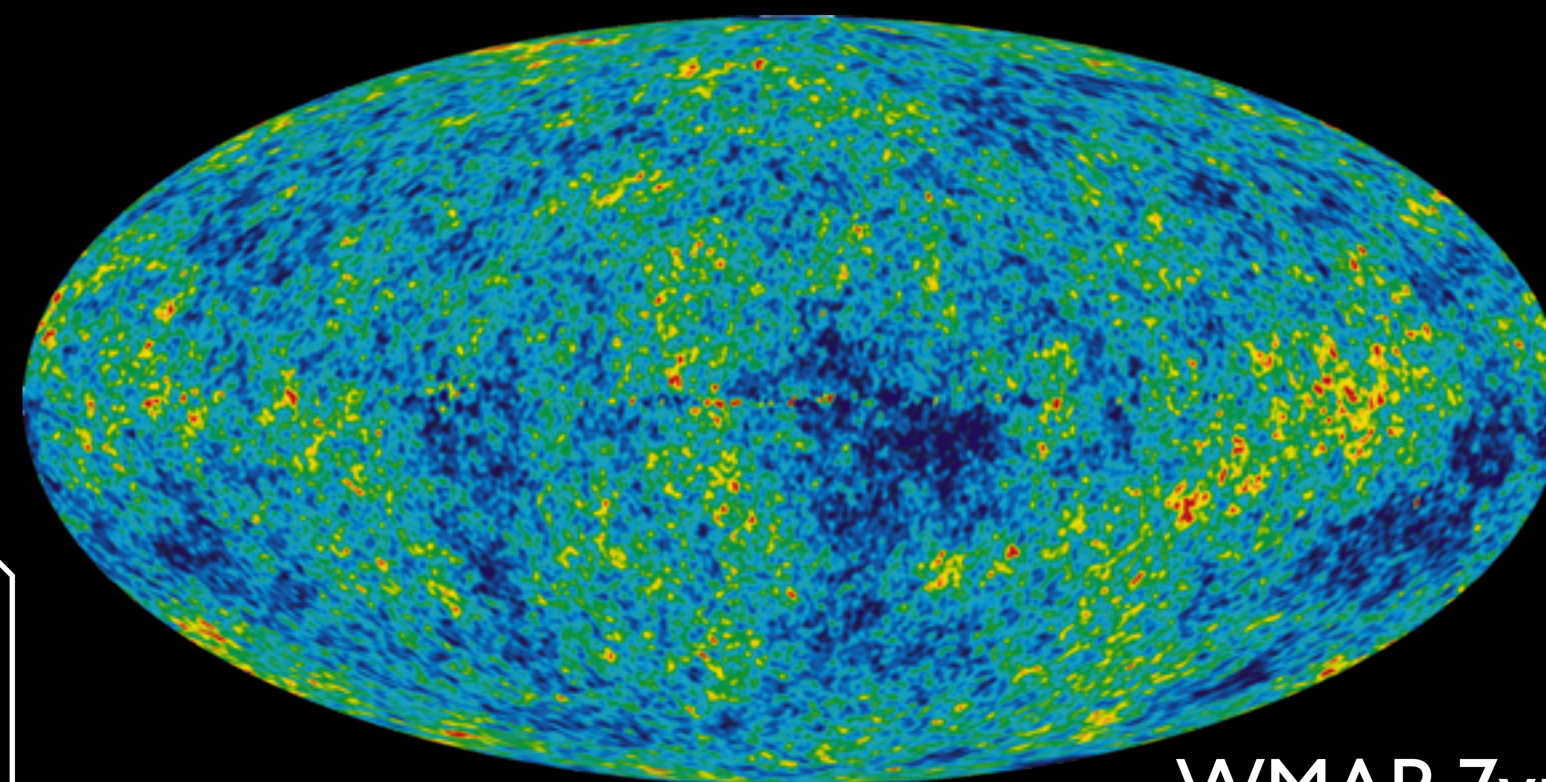
The EOR is "like" the Cosmic Microwave Background, but better...

The CMB samples just one redshift: ~1100





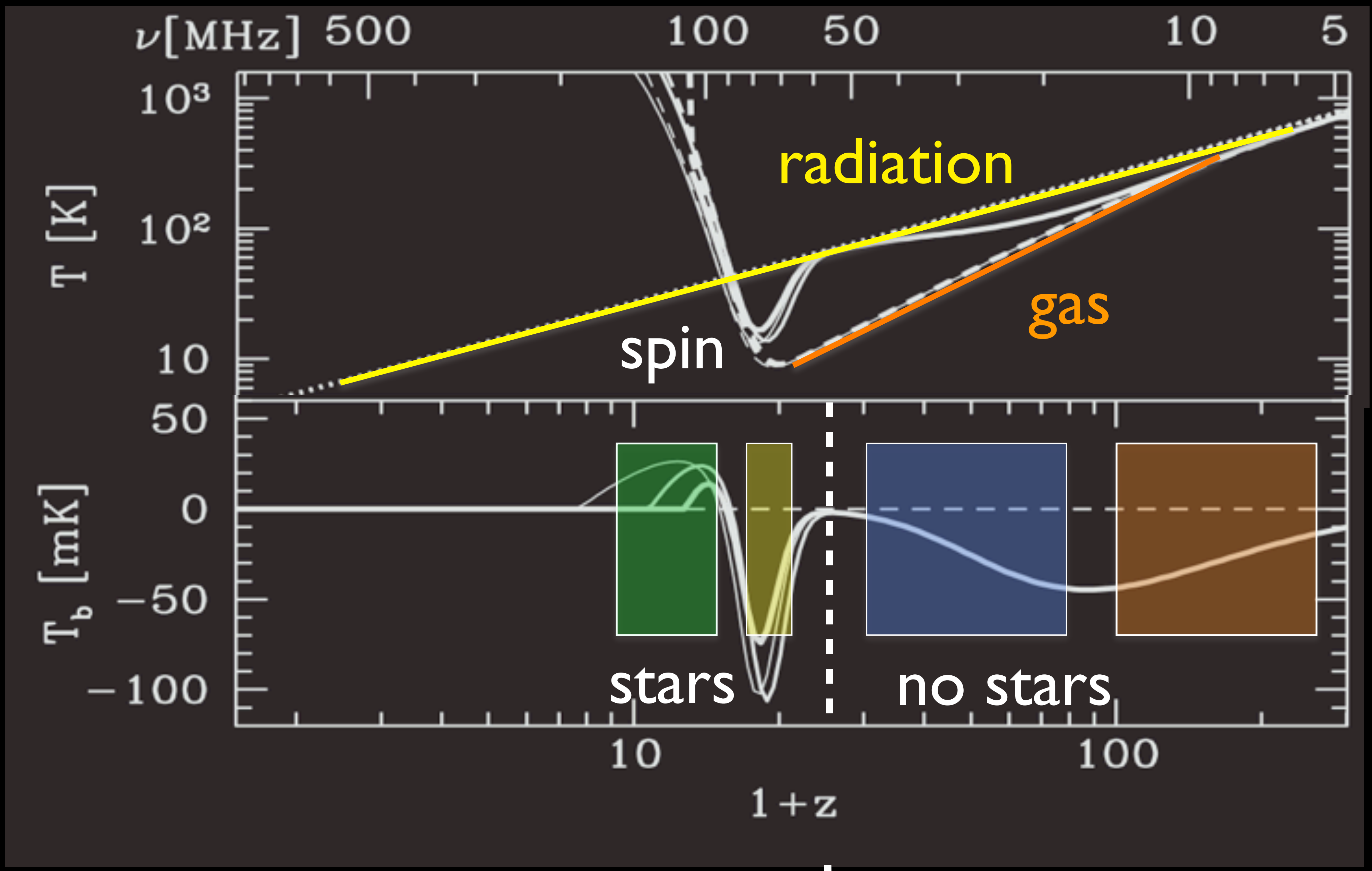
2D



WMAP 7yr
(GSFC)

L. Greenhill

λ21cm DC Signature on the Sky



Lyα couples T_{spin} to T_{gas}

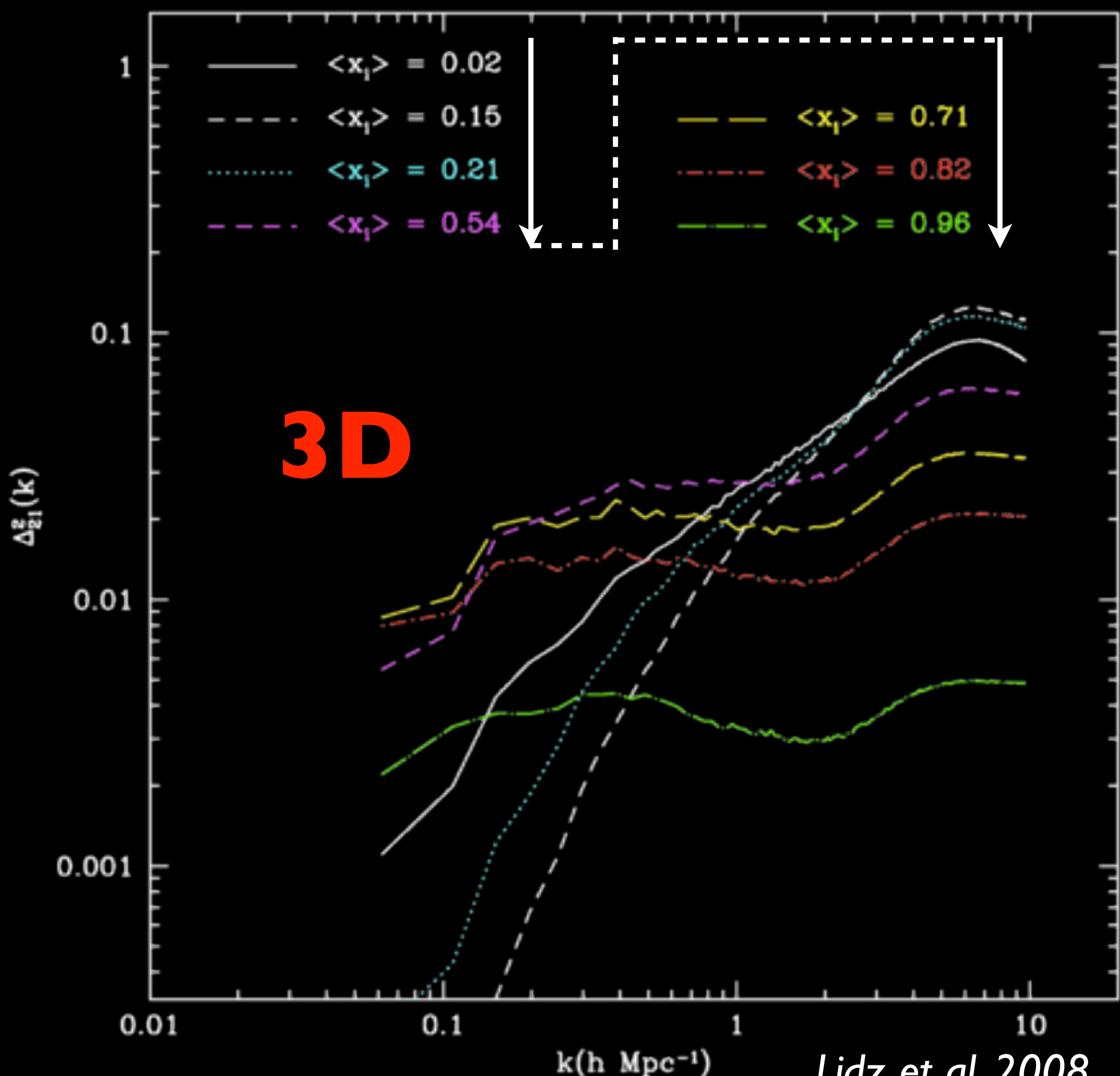
T_{gas} rises due to X-ray heating

(e.g., Furlanetto et al. 2006, references therein)

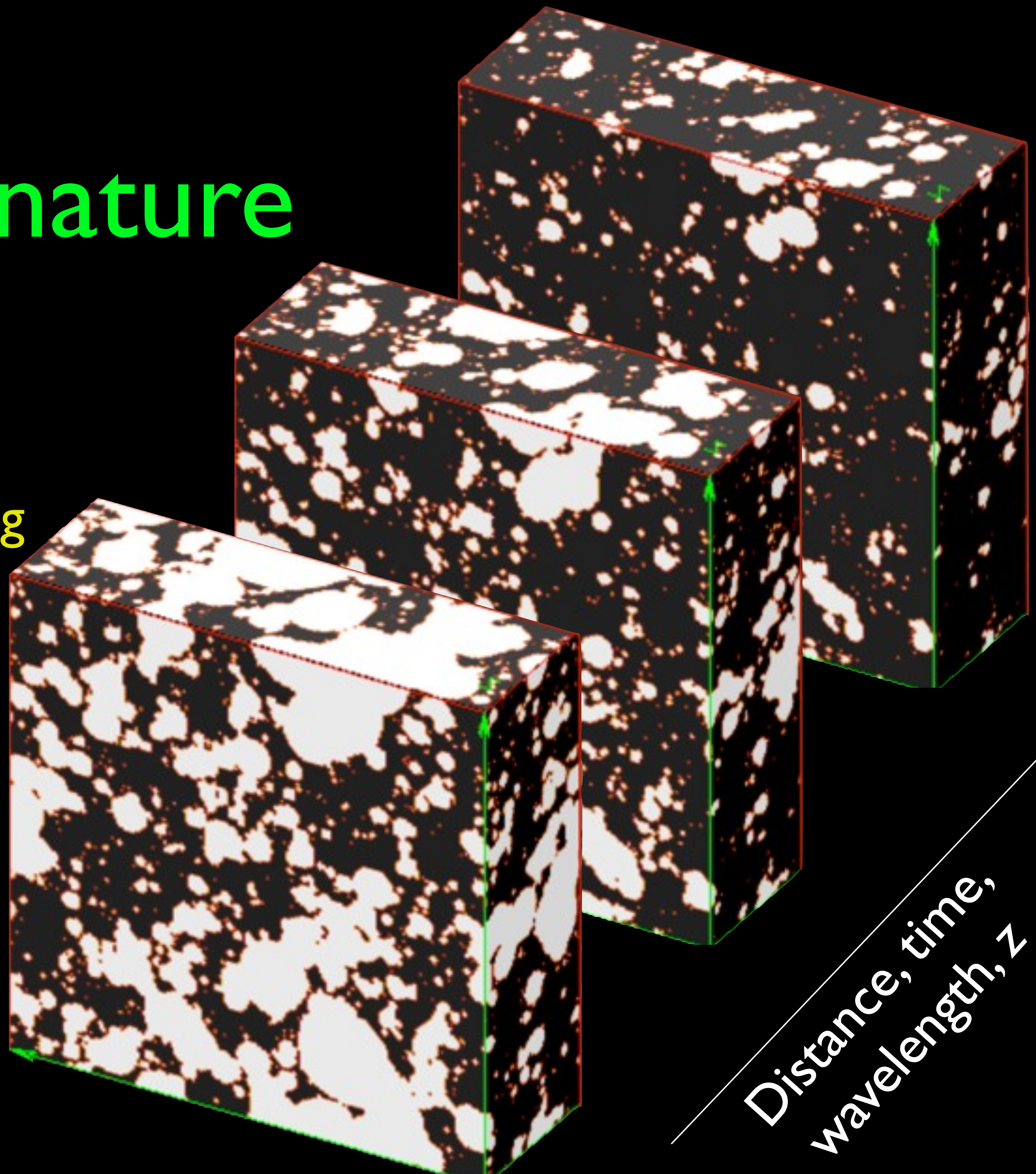
L. Greenhill

λ21 cm AC Signature

low S/N per pxl ➟ power spectra
high S/N & OOB rejection ➟ imaging



Lidz et al. 2008

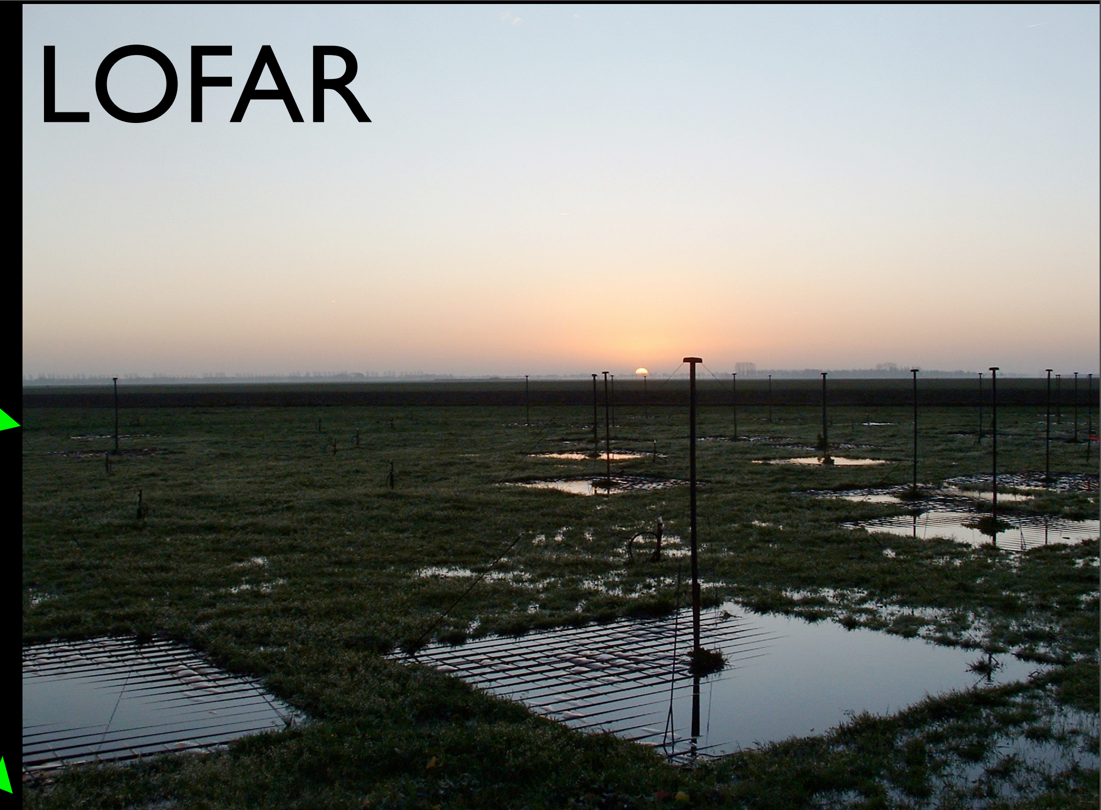


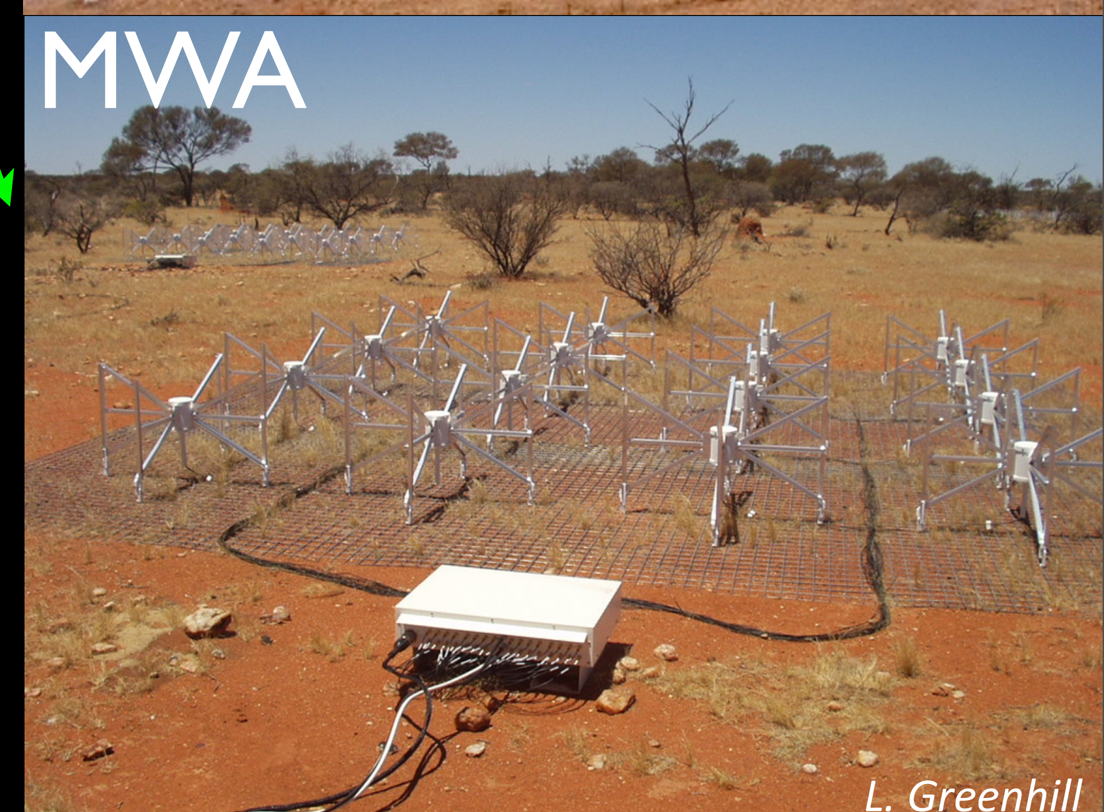
- Slicing the early universe
- More distant gas appears at longer wavelength

L. Greenhill

New Gen.

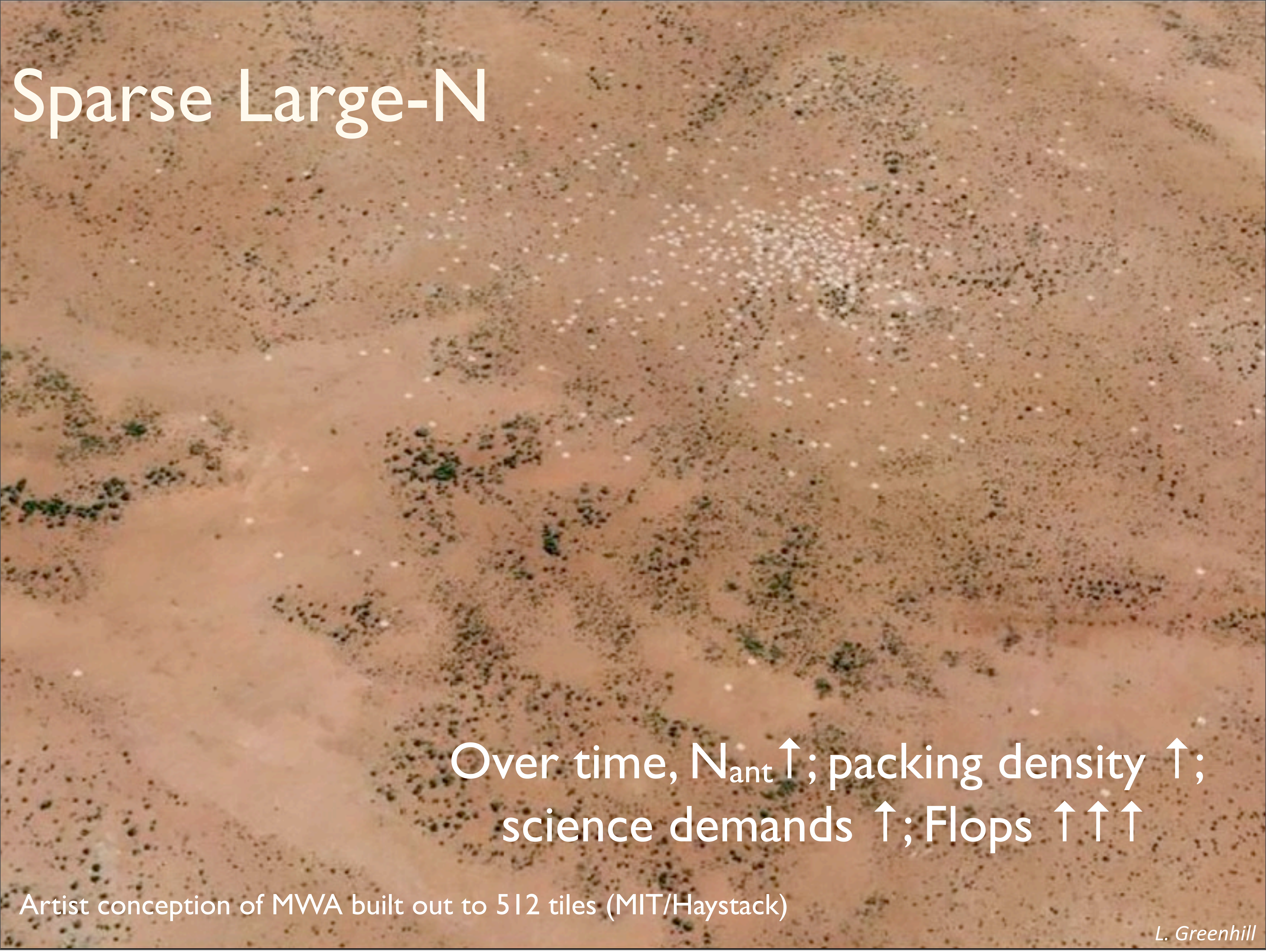
- at low-ν, collecting area is comparatively cheap w/ wide F.o.V.
- but antenna gain is low → mass deployments of antennas required
- signal processing complexity is a throttle






EVLA
LEDA / LWA
Large-N
L. Greenhill

Sparse Large-N



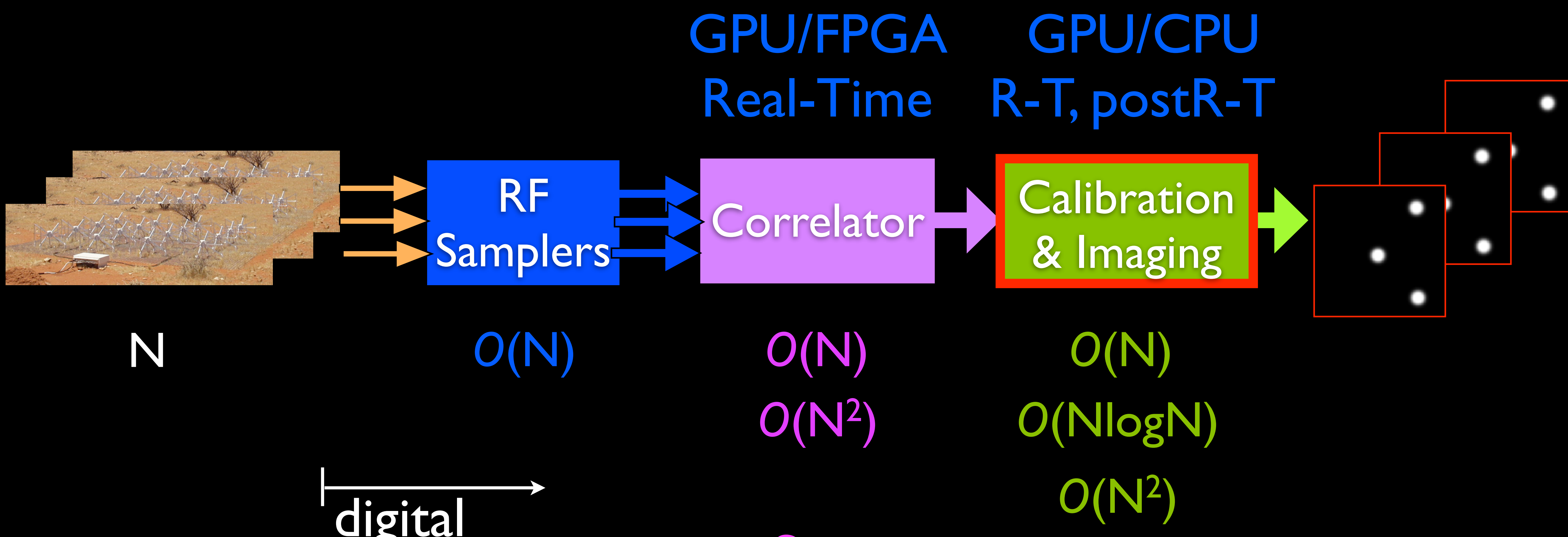
Over time, N_{ant} ↑; packing density ↑; science demands ↑; Flops ↑↑↑

Artist conception of MWA built out to 512 tiles (MIT/Haystack)

L. Greenhill

Dipole Array Signal Processing

Heterogeneous HPC



Dipole arrays break the assumptions relied upon in operation of traditional arrays

See Clark talk

This talk

Example: Murchison Widefield Array

- 512 detectors distributed over 1 km^2; 5% prototype now operational
  - 40 Gb s^{-1} output rate from correlator on 72 1-gE pipes (parallelizes by ν bands)
  - 130,816 pairs processed on *O*(1) μs time scales
  - accumulation to 2,4,8^s
  - 768 frequency channels
  - 2 polarizations per detector → 4 products correlation
- extant 5% prototype in operation
- 80-300 MHz receiver waveband (VHF/UHF)
  - 30.72 MHz instantaneous bandwidth (would prefer > 100 MHz)
- MWA calibration & imaging is real-time stream processing
  - one pass (unlike most other examples among radio arrays, but a likely future)
- notable computational science elements
  - 1 pipe from correlator = 1 pipe for calibration/imaging
  - end-to-end pipeline execution on GPUs; heterogeneous calculation; from scratch
  - broad mix of mathematical operations: FFT, convolution, matrix ops, grid, ... SP

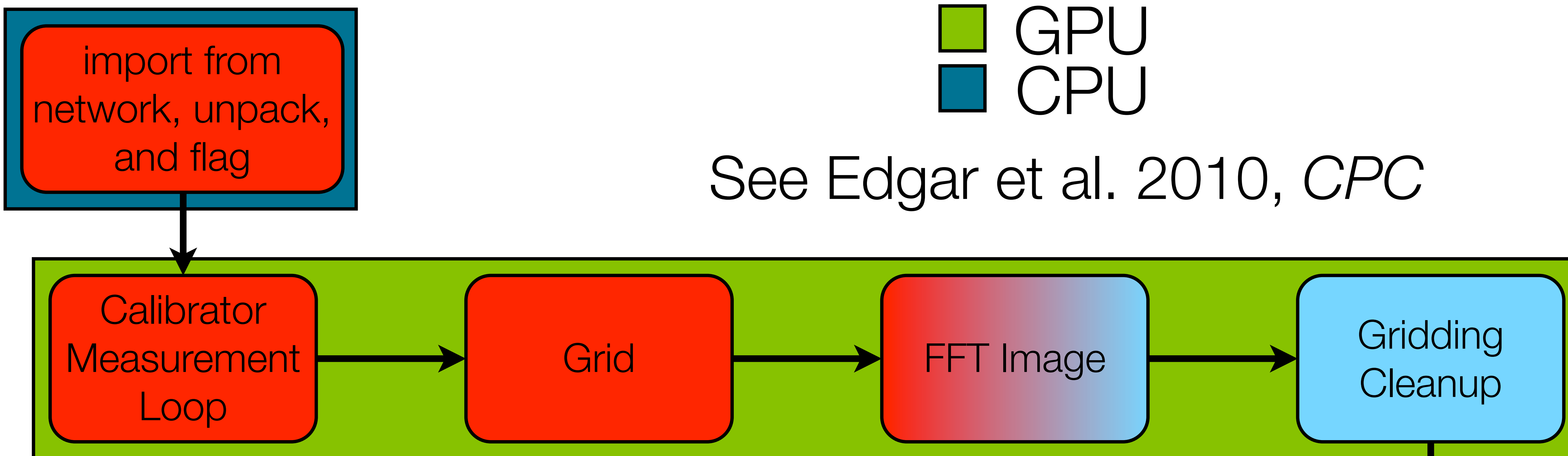
Adapted from Richard Edgar

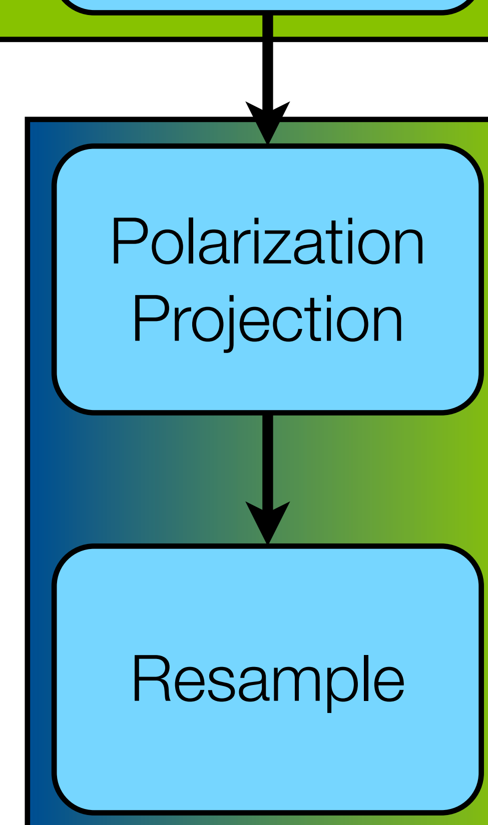
MWA motivation for GPU use

- CPUs problematic vis-a-vis power budget
    - 30 kW initial power spec. on site
    - O(20)+ TFlop problem to be completed in < 8^s
    - CPU: adopt avg. O(10) GFlop s^{-1} REAL
        - 250 multi-core processors; assume 200 W per processor + ancillary bits
        - 50 kW
  - CPU as well drive inefficient parallelization - increases communications & cost
    - natural parallelization of problem: 64-72 nodes
- Can we do the job with GPUs?
  - lab testing validates 64 GPU test configuration (C2070; now in construction)
  - meets 8^s cap
  - ~ 30 kW
  - enables natural parallelization of problem
  - vast headroom enables upgrade in algorithms (80 TFlop s^{-1} theoretical capacity)

MWA Calibration & Imaging

See Edgar et al. 2010, *CPC*

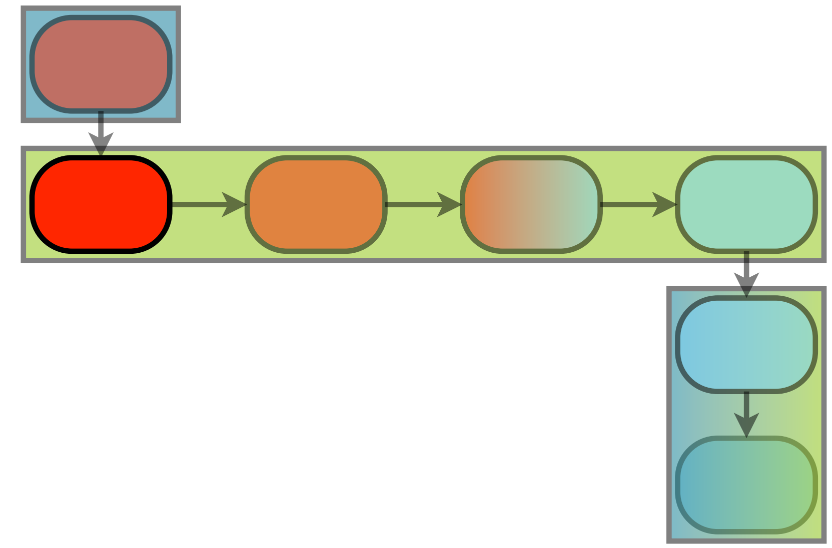




- remote location (600 km from Perth)
  power and storage are limited
- real-time processing enables data compression: 40 Gb $s^{-1} \rightarrow O(1)$ Gb s^{-1}
- multi-TFlop calculation every 8^s
- open-ended demand for Flops
  - more resources -> better calibration & science
- flavor of computational steps follow...

Adapted from Richard Edgar

Calibrator Measurement Loop

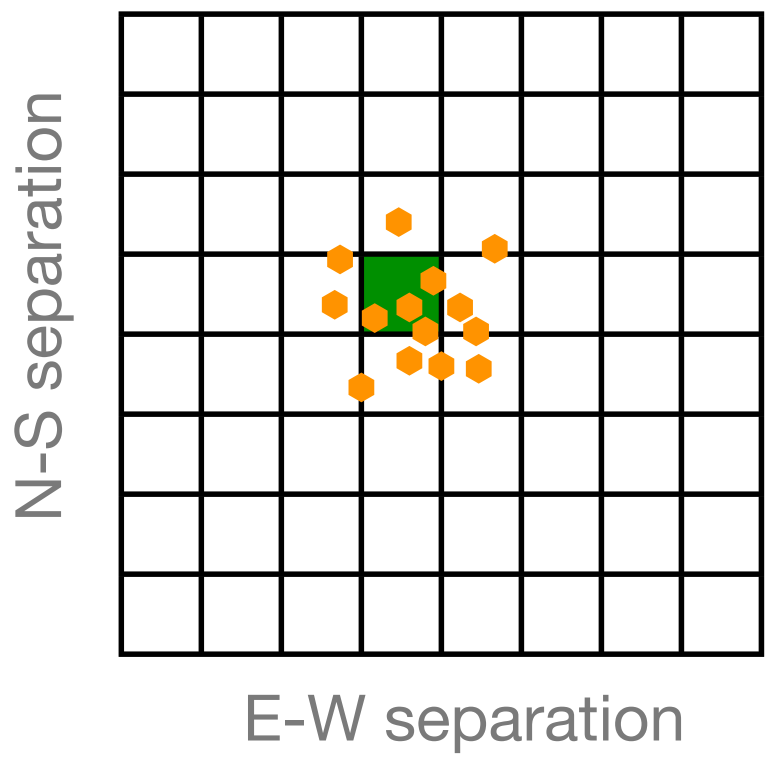


- phase array → complex-data vector & matrix transforms
  - coherent addition
  - measure source strengths, locations vs catalog
  - estimate antenna gains and ionospheric distortions on grid across the sky
  - apply calibrations to data vectors
  - peel bright sources (build and subtract data vectors for models)
- solve for gain patterns of antennas across consecutive ν channels
- solve for ionospheric rubber sheet based on offsets as fn of angle on sky
  - use known ν^2 dependence
- each node has consecutive channels
  - gross parallelization of problem over frequency
  - MPI communication on GPU cluster for antenna gain and ionospheric fits
  - only point where channels communicate

Adapted from Richard Edgar

Gridder

- Interpolate correlator output (antenna pairs) onto regular grid to enable FFT
- Must convolve each data point with a compact kernel - *O*(2%) size
- implement Gather operation to avoid race condition in || processing
  - roundabout compared to Scatter op. used on CPUs
  - parallel operation of GPU wins out if Search is efficient



parallelize by Fourier-domain pixel

- sort data (•) into bins ~ kernel size, *O*(30) pixels
- sort data by bin
- tabulate 1st and last data in each bin
- use tables to pare data searched
- pull in data applying kernel

room for improvement

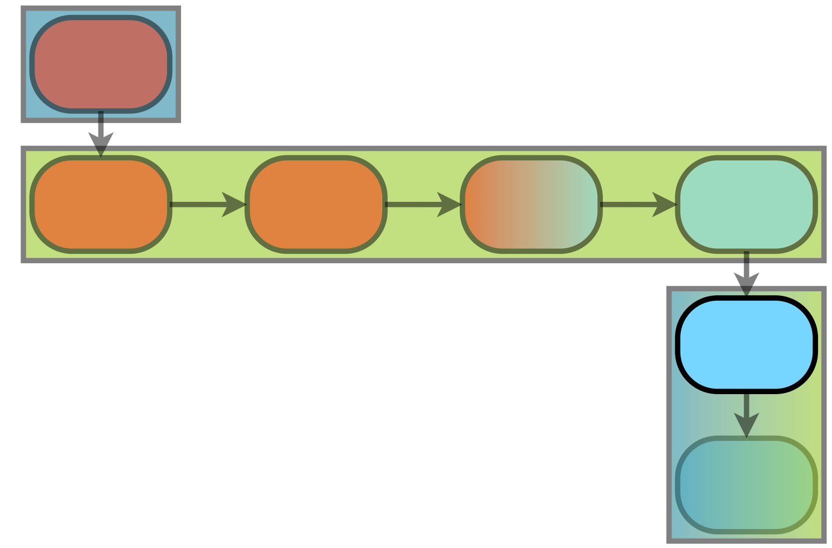
- z-ordering
- parallelize over complexity and polarization 1st

most likely axes for scaling

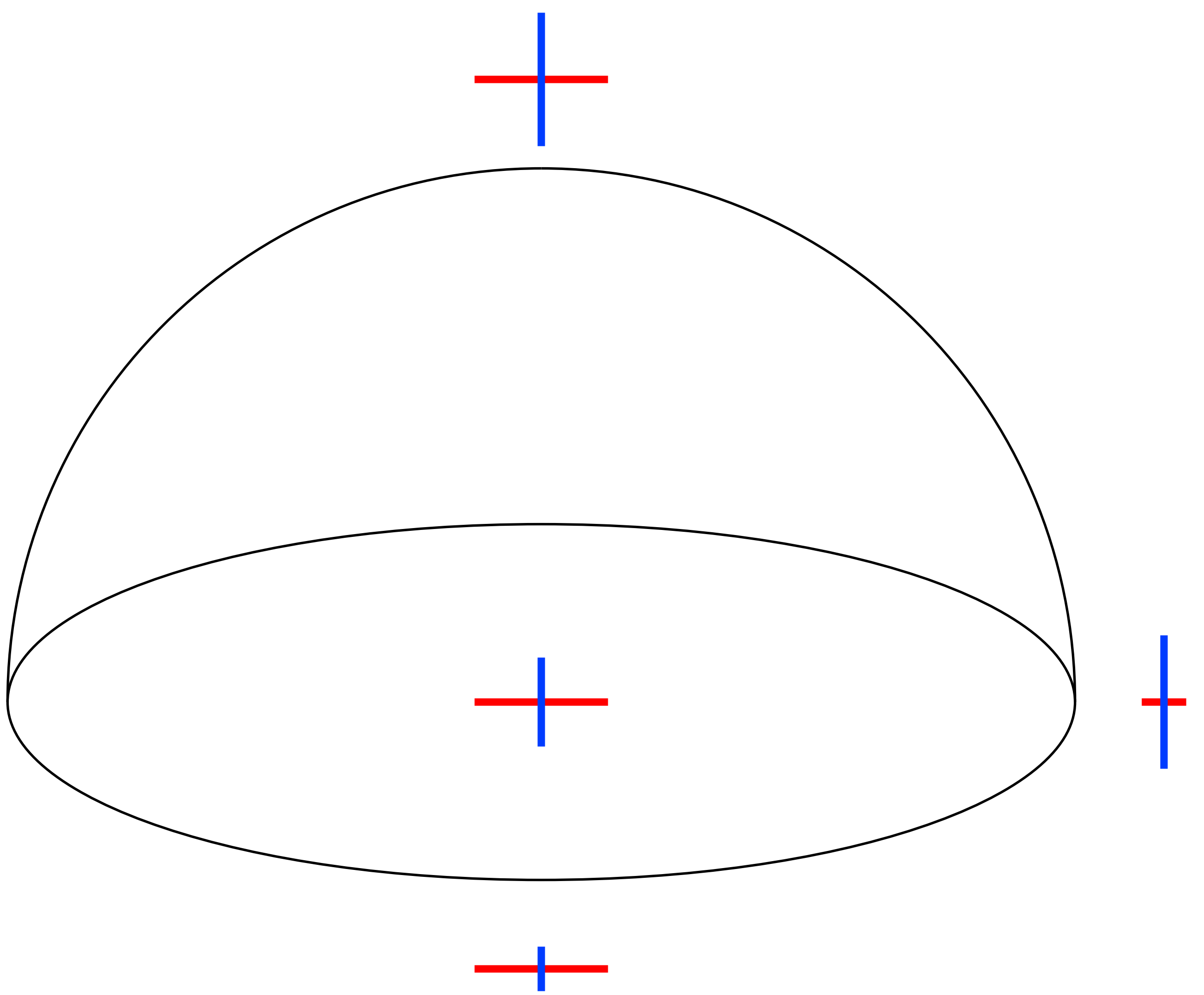
- no. of data points
- kernel size

Adapted from Richard Edgar

Polarization Projection

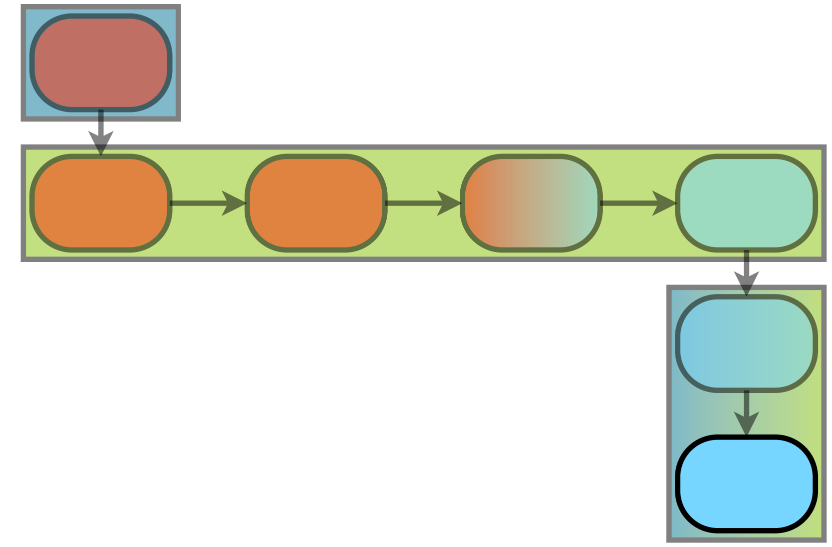


- Have four polarisations in ground frame
- Want polarisations in sky frame
- A different transform for every pixel on the sky
- Each pixel is 4 element vector
  - Multiply by 4x4 matrix
- Leverages heterogeneous model
  - projection matrices predictable
    - computed on the CPU
    - applied on GPU

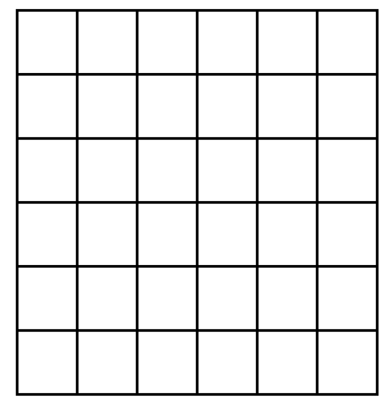


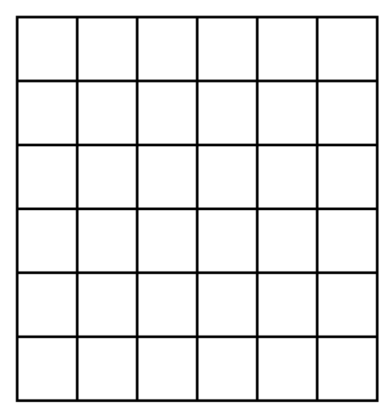
Adapted from Richard Edgar

Regridding Images



- ionospheric distortion
- distortion due to wide field of view
- sky curvature (use HEALPIX frame)
- Heterogeneous computing model
    - vertex overlaps predictable
        - computed on CPU
        - applied on GPU

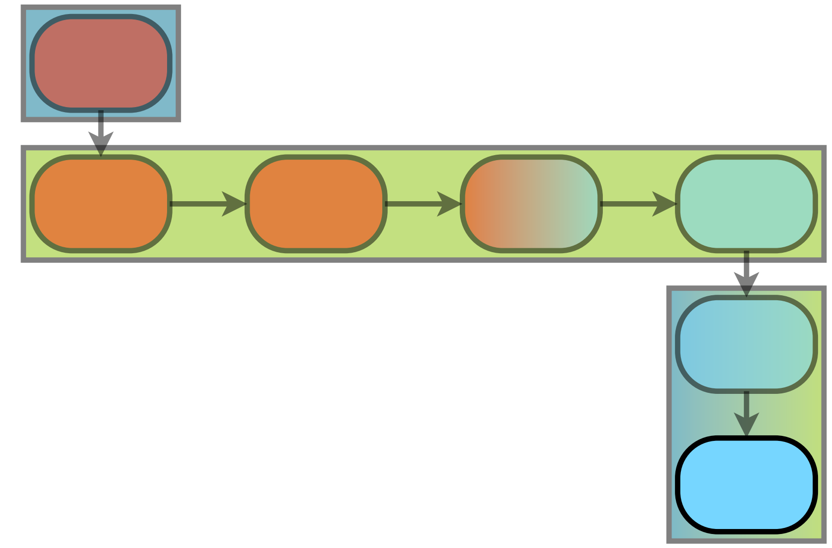




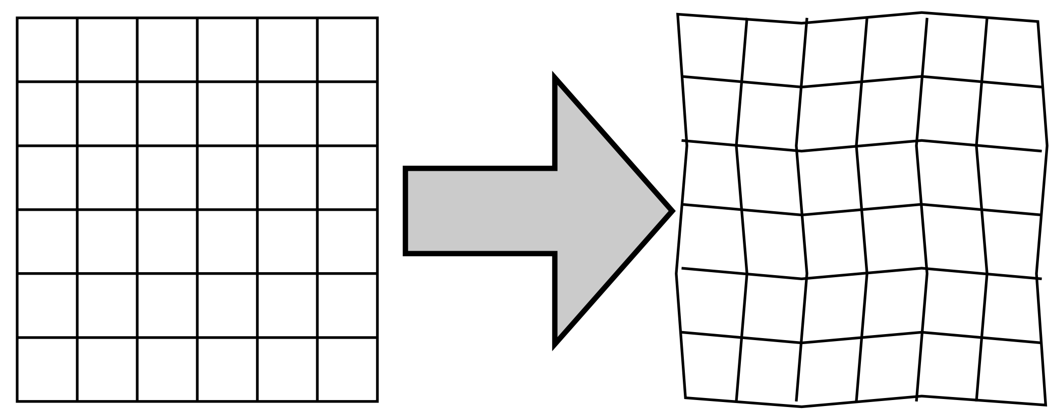
Current production implementation is simple wgt'd avg. Require flux-conserving interpolation

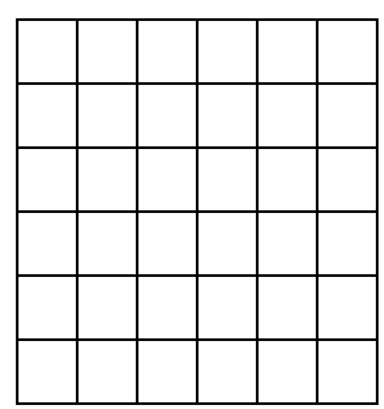
Adapted from Richard Edgar

Regridding Images



- ionospheric distortion
- distortion due to wide field of view
- sky curvature (use HEALPIX frame)
- Heterogeneous computing model
    - vertex overlaps predictable
        - computed on CPU
        - applied on GPU

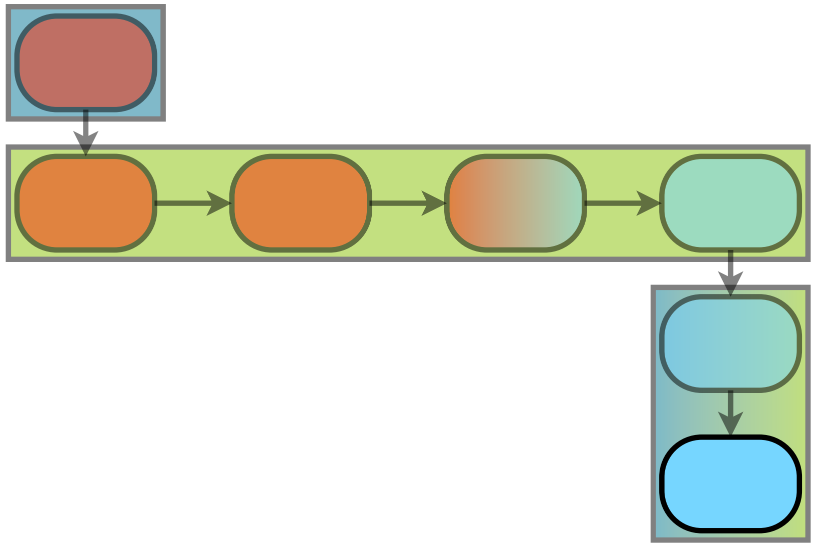




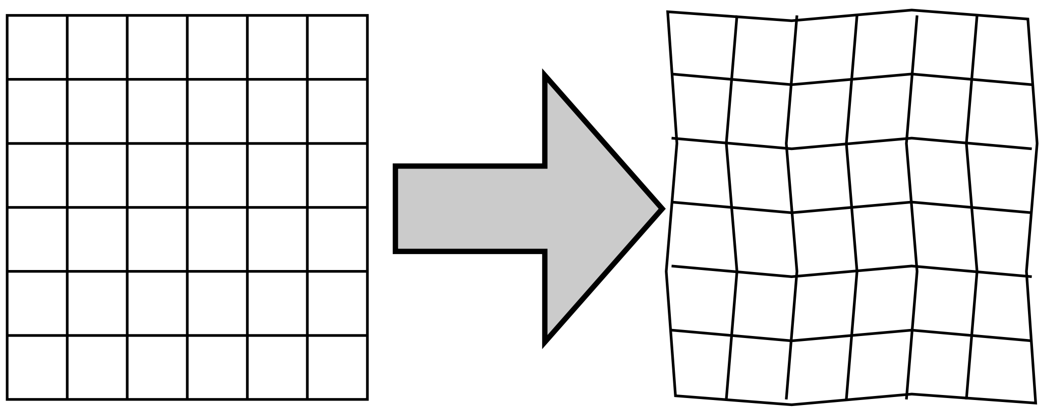
Current production implementation is simple wgt'd avg. Require flux-conserving interpolation

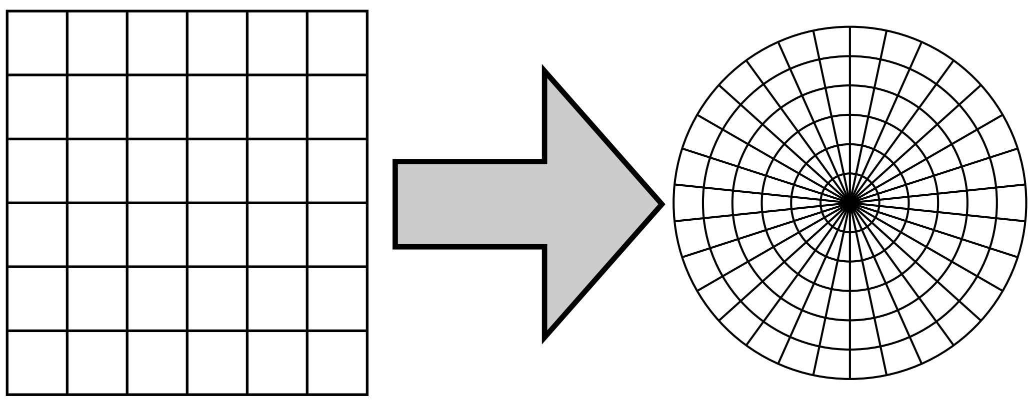
Adapted from Richard Edgar

Regridding Images



- ionospheric distortion
- distortion due to wide field of view
- sky curvature (use HEALPIX frame)
- Heterogeneous computing model
    - vertex overlaps predictable
        - computed on CPU
        - applied on GPU

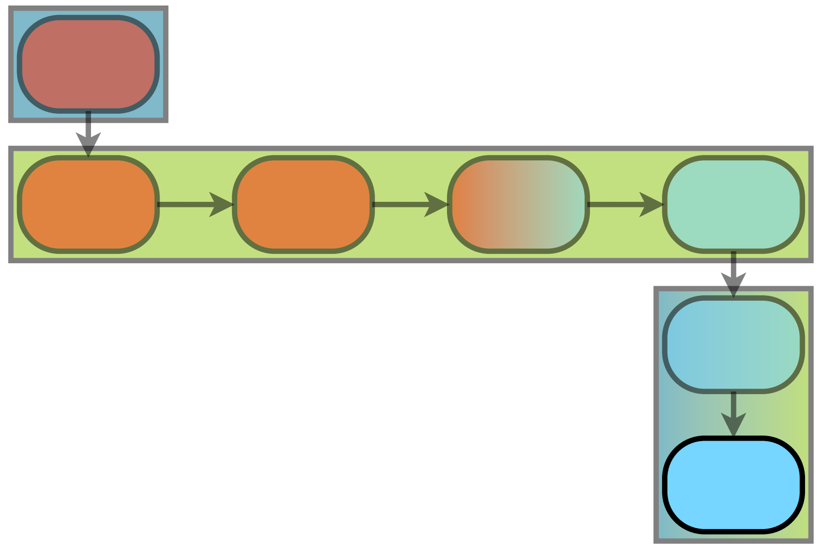




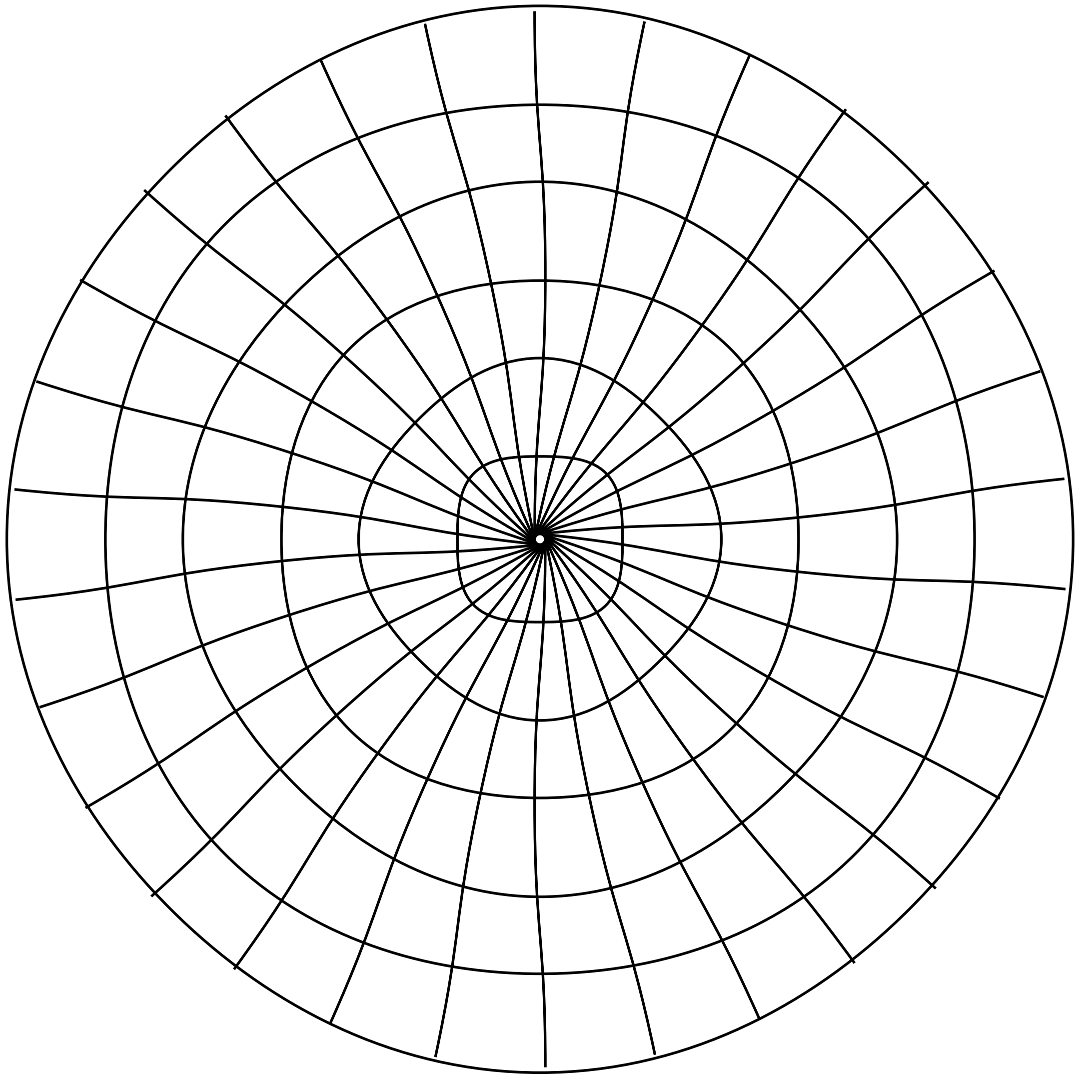
Current production implementation is simple wgt'd avg. Require flux-conserving interpolation

Adapted from Richard Edgar

Regridding Images



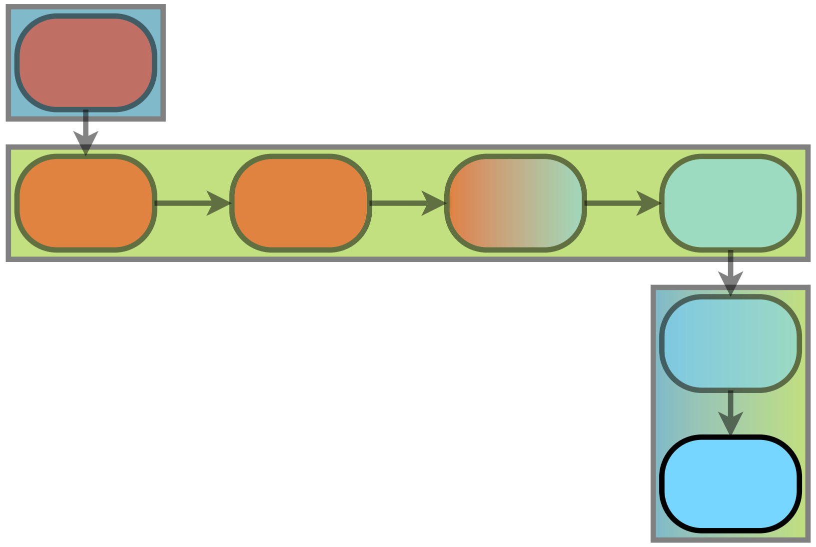
- ionospheric distortion
- distortion due to wide field of view
- sky curvature (use HEALPIX frame)
- Heterogeneous computing model
  - vertex overlaps predictable
    - computed on CPU
    - applied on GPU



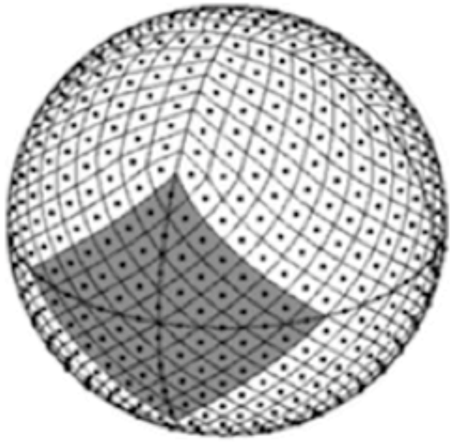
Current production implementation is simple wgt'd avg. Require flux-conserving interpolation

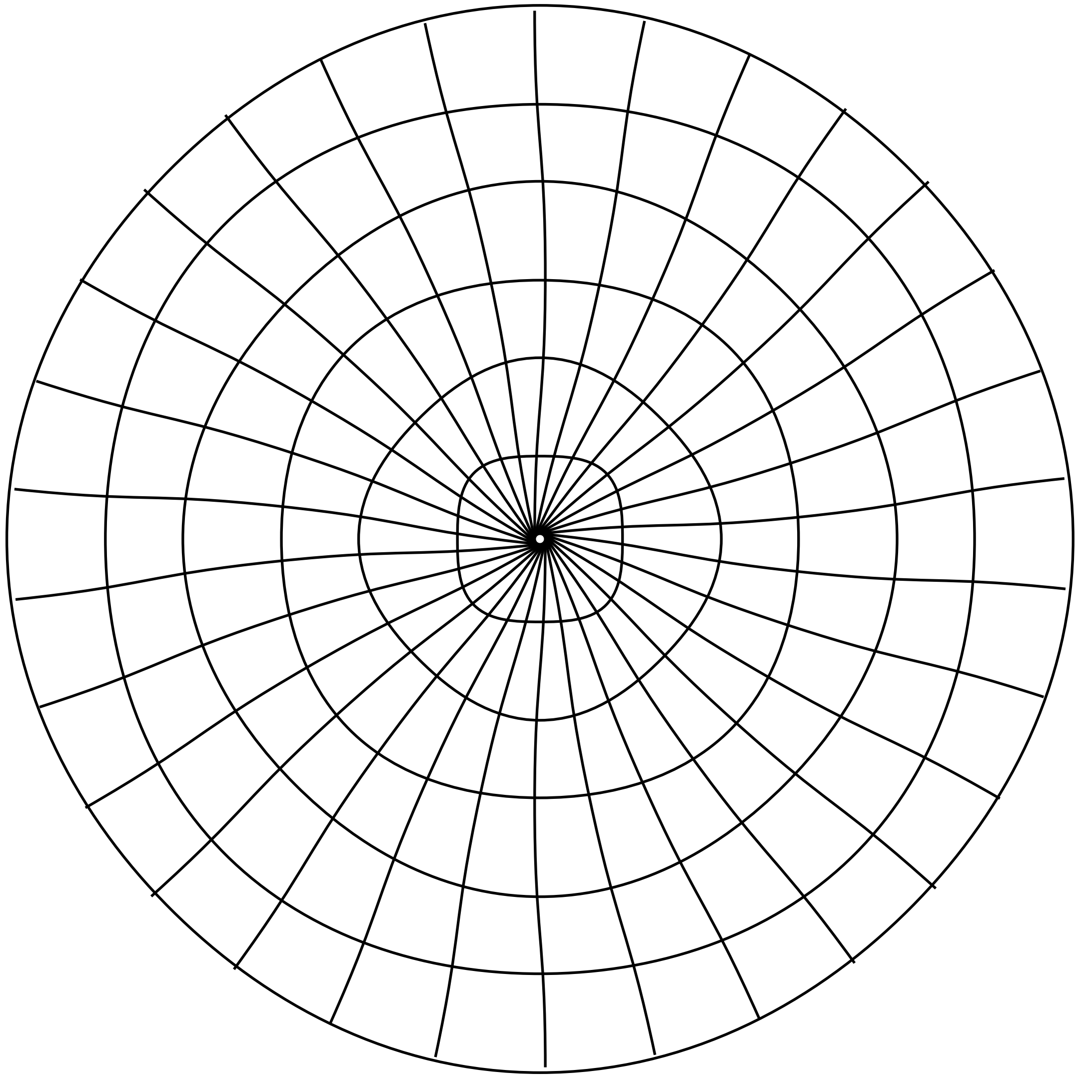
Adapted from Richard Edgar

Regridding Images



- ionospheric distortion
- distortion due to wide field of view
- sky curvature (use HEALPIX frame)
- Heterogeneous computing model
  - vertex overlaps predictable
    - computed on CPU
    - applied on GPU

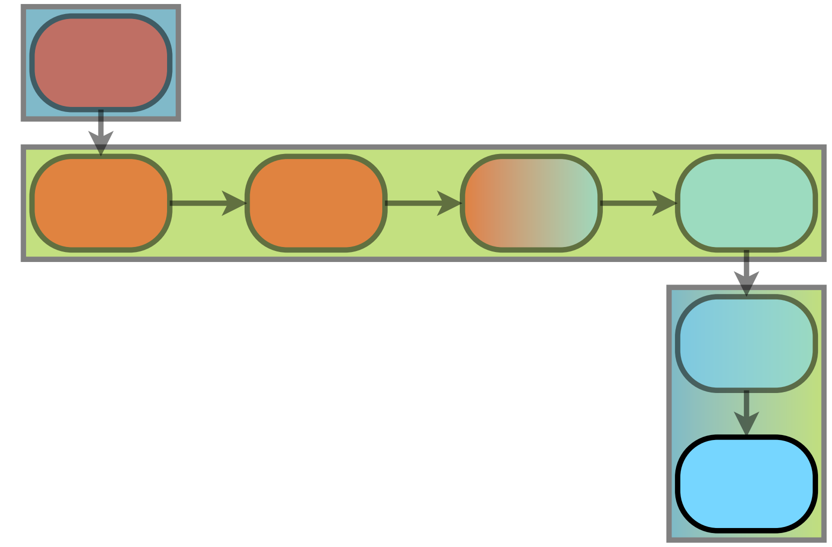




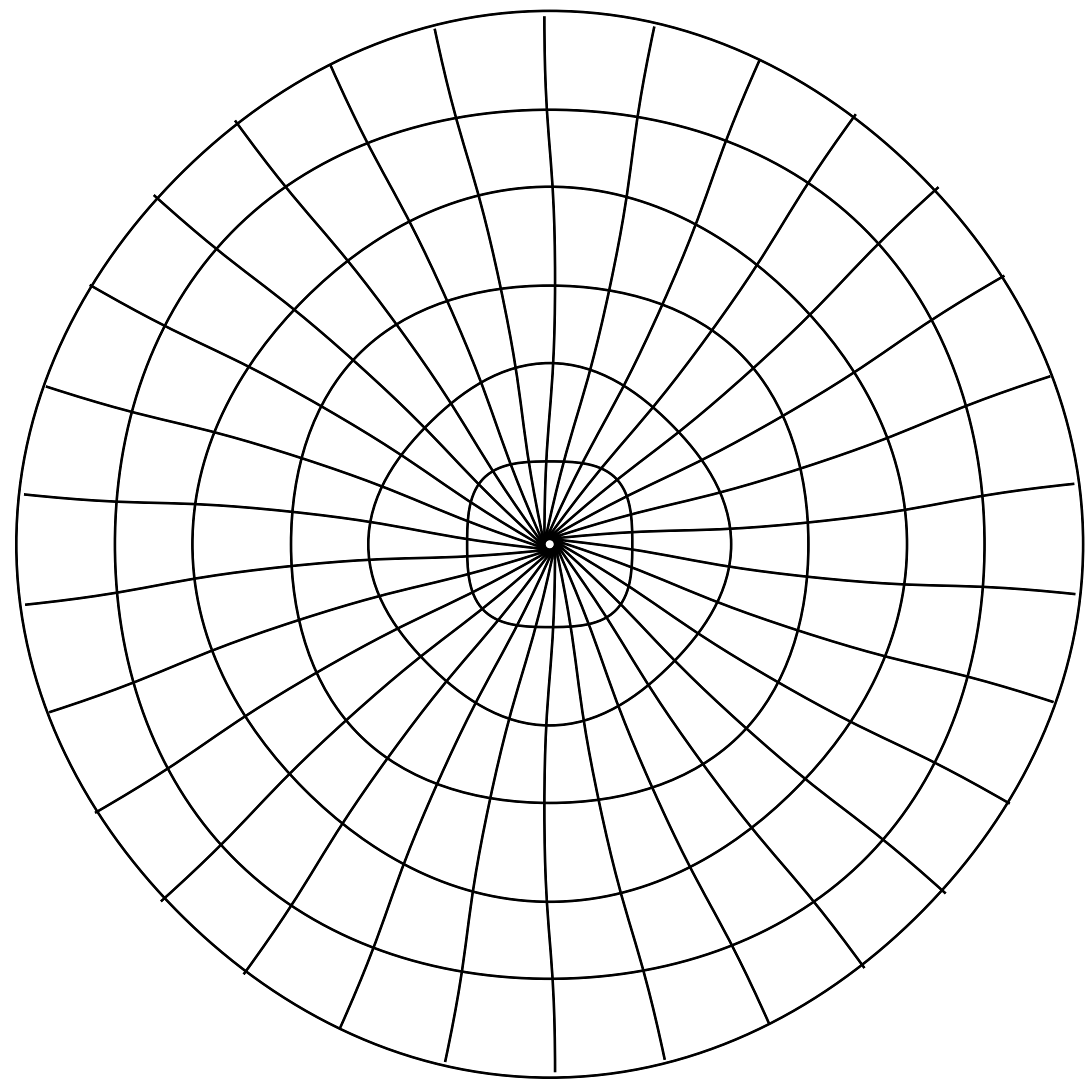
Current production implementation is simple wgt'd avg. Require flux-conserving interpolation

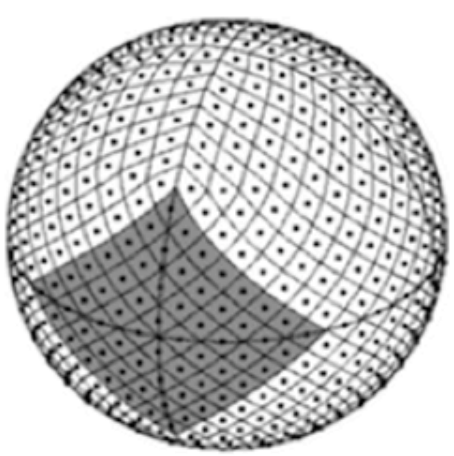
Adapted from Richard Edgar

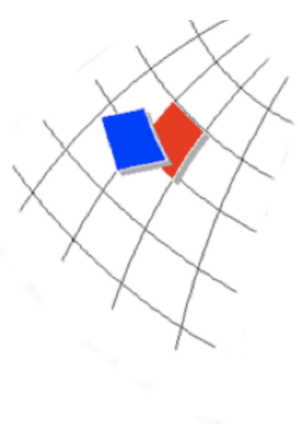
Regridding Images

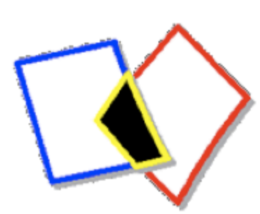


- ionospheric distortion
- distortion due to wide field of view
- sky curvature (use HEALPIX frame)
- Heterogeneous computing model
  - vertex overlaps predictable
    - computed on CPU
    - applied on GPU









Current production implementation is simple wgt'd avg. Require flux-conserving interpolation

Adapted from Richard Edgar

Benchmarks

- Benchmarks performed on
  - 3 GHz Intel Xeon E5462 (Harpertown)
  - NVIDIA Tesla S1070
- Benchmarks for
  - Single channel
  - 5 calibration sources
  - 30 degree field of view (1600^2 pixels)

Adapted from Richard Edgar

Benchmarks

- Overall speed up 18.7x
  - results based on limited optimizations (CPU, GPU)
  - apples and oranges, but CPUs fail to meet 8 s deadline. Full stop.
- Performance/$ improvement 11.7x
- Performance/W improvement 10.2x
- Further work to be undertaken within MWA
  - Memory optimisations
  - Faster gridding
  - Mixed precision
  - Tailoring to Fermi

Stage	CPU (sec)	GPU (sec)
Acquire Data	1.09	1.08
Send GPU	0.0	0.03
Calibrator Measurement Loop	500.52	3.58
Gridding Preparation	5.13	0.17
Gridding	14.70	1.40
Imaging	3.78	0.34
Receive GPU	0.0	0.05
Deprojection	3.56	0.10
Regridding	4.55	0.49
Cleanup	0.01	0.13
Total	533.34	7.37

Table 3: Comparison of CPU and GPU timings for individual stages of the RTS. Timings are for a 12 channels, with the CML using 50 calibration sources. The gridding convolution function was 24×24 pixels in size, and 1125×1125 pixel images were produced. These timings do not include the precomputations for the Deprojection and Regridding stages.

Adapted from Richard Edgar

Scaling to 10x

- motivation to look to larger computing scales
  - US community plan for next gen. instrument
    - HERA (Hydrogen EOR Array)
    - 10x and 100x "current" apertures c. 2015, 2020
    - endorsement by 2010 astro. decadal survey
- signal processing via HPC backbone
- getting to 10x...

L. Greenhill

Scaling to 10x

- computing dependent on design considerations
  - HPC is the lynchpin for dipole arrays
  - hierarchy of RF array → computing framework
  - but lessons not yet learned w/ current generation
- array characteristics
  - N: antennas or tiles   B: bandwidth (# of ch.)
  - F: field of view   S: array geographic size
- computation
  - correlation
    - $\propto k_0 N^2 F B + k_1 N B$
  - calibration & imaging
    - $\propto k_3 N^2 B + k_4 B (F S)^{1-2}$
    - S scaling can be weakened for compact-condensed array

L. Greenhill

Scaling to 10x

- working memory
  - problem parallelizes over frequency
  - keep data local to GPU (power, compute speed)
  - As N_{ant} grows, N_{ch} per node drops
    - undesirable to have $N_{ch} < 1$ per GPU
  - 6 GB on GPU allows up to $N_{ant} \sim 20000$
    - A_e per element ~ 8-20 $m^2 \rightarrow N_{ant} = 5000\text{-}12000$
    - memory volume is likely not a problem, but BW may be
- are I/O and ops. rates manageable ?

Scaling

N_{ant}	Correlator Tb s^{-1}		X-corr. (Top s^{-1})	"MWA Cal" (TFlop s^{-1})
	In	Out		
512	1.08	0.084	420	170x_{iter}
1024	2.16	0.33	1700	230x_{iter}
2048	4.32	1.3	6700	480x_{iter}
4096	8.65	5.4	26800	1500x_{iter}
8192	17.3	22	107000	5300x_{iter}
16384	34.6	86	429000	21000x_{iter}

time ↓

10-100 PFlop s^{-1}

PAPER dipole: 8 m^2
N_{ant} ~ 12000
MWA dipole: 20 m^2
N_{ant} ~ 5000

32 PFlop s^{-1} c.2016 comparable in size to Nebula deploy't

Is power budget affordable?

- combine corr. + cal/im on GPU → savings
- e.g., see Clark talk

5 km extent; 25° FOV; 100 MHz bandwidth; 5 bit sampling; 10 kHz channels at correlator; 100 kHz-avg for science; characteristic MWA single pass calibration; peel 50 calibrators; 21x21 gridding kernel

L. Greenhill

Summary

- Direct sensing of the IGM during reionization is a frontier in observational cosmology
- Large, low-frequency radio arrays are central
- Entail an entirely new signal processing model
- HPC will be the lynchpin
  - manycore (GPU) is critical for correl., cal., and imaging of filled apertures w/ wide FoV ... next step is peta-scale
- Astro2010 endorsement of HERA concept
  - design, engineering, shakedown w/ current arrays
  - 10^5 m^2 (10x current) by 2nd half of decade

L. Greenhill

- end -

Benchmarks - CML

Stage	**CPU (ms)**	**GPU (ms)**
Clear Groups	12.1	12.5
Unpeel	1489.6	9.5
Rotate & Accumulate	1397.2	10.3
Scale	70.9	1.1
Measure Ionospheric Offset	349.7	17.7
Ionospheric Correction	97.6	1.3
Measure Tile Response	1116.8	46.6
Peel	506.3	5.9
Total	**5569.6**	**104.6**

Adapted from Richard Edgar

Benchmarks - Gridder

Stage	**CPU (ms)**	**GPU (ms)**
Prepare Spheroid		5.1
Memory		18.2
Locations	41.6	0.3
Bin		0.4
Sort		6.8
Reorder		1.5
Lookup Table		0.1
Convolve	1282.7	152.0
Total	**1324.3**	**186.6**

Adapted from Richard Edgar

Benchmarks - Imager

Stage	**CPU (ms)**	**GPU (ms)**
Conjugates	79.4	1.8
Send	55.4	2.0
FFT	304.7	29.9
Receive	145.7	8.6
Total	**587.9**	**42.3**

Adapted from Richard Edgar

Benchmarks - Gridding Cleanup

Stage	CPU (ms)	GPU (ms)
Make Corrector	26.8	10.0
Apply Corrector	98.1	1.2

Adapted from Richard Edgar

Stokes Conversion

Stage	**CPU (ms)**	**GPU (ms)**
Apply Transform	438.1	6.6
Retrieve Image		21.6
Total	**438.2**	**28.2**

Adapted from Richard Edgar

Regridder

Stage	CPU (ms)	GPU (ms)
Send Regridding Information		54.6
Perform Regridding	730.7	26.9
Retrieve Image		23.2
Total	**730.7**	**104.7**

Adapted from Richard Edgar

δT_b

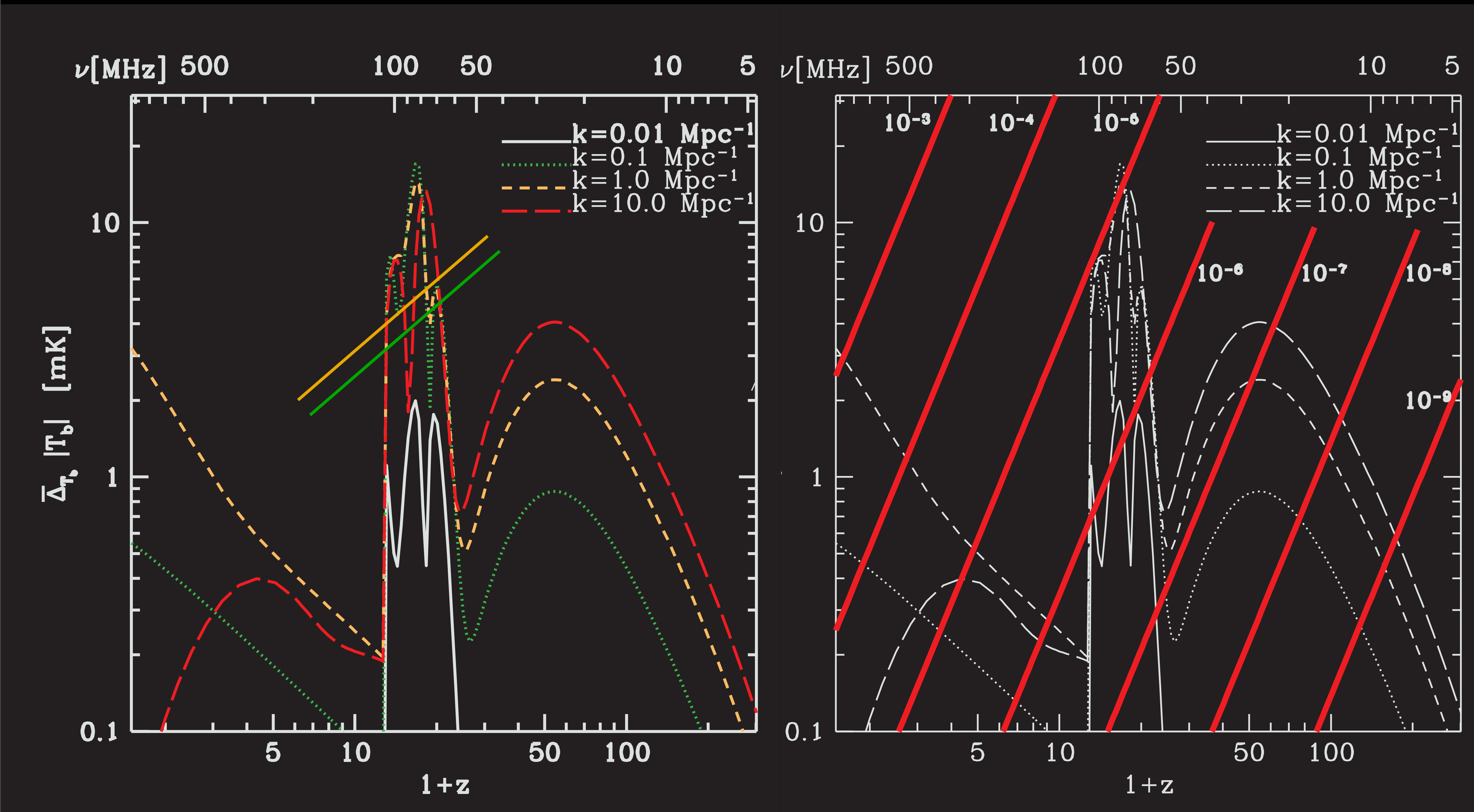
- $T_b \sim 28$ mK $(1+\delta)h^2\, x_{HI}\, [\, 1 - T_S / T_{CMB}]$

  $* [\Omega_b / 0.02]\, [\Omega_m / 0.24]^{-1/2} [(1+z)/10]^{1/2}$

  - δ: density deviation from mean
  - T: temperatures, Brightness, Spin and CMB
  - x_{HI}: neutral fraction of HI

L. Greenhill

AC Signature vs Redshift

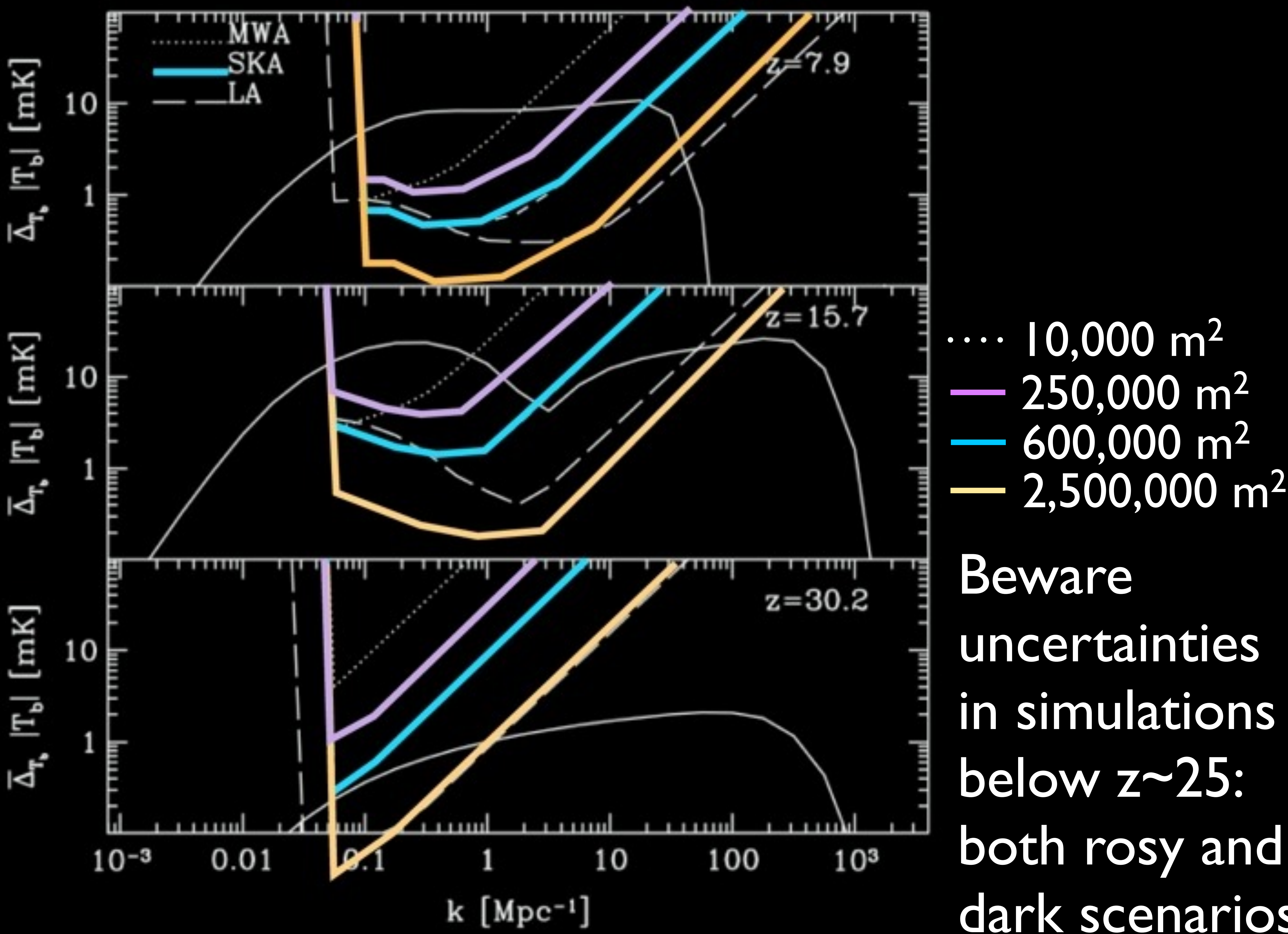


Pritchard & Loeb 2009

$k = 1.0\ \mathrm{Mpc}^{-1}$ @ $10 < z < 20$ ➠ ~ 2′

L. Greenhill

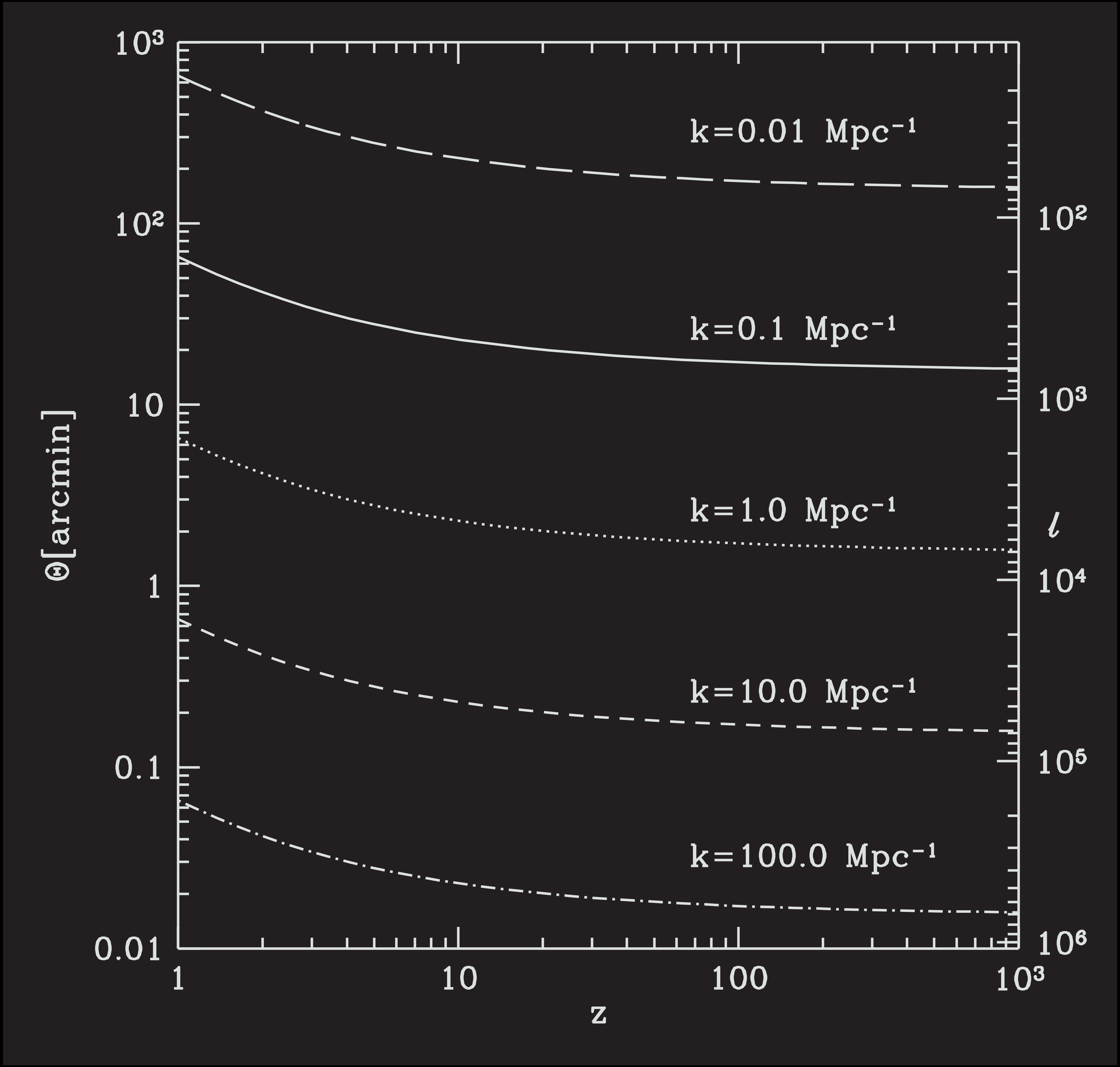
Thursday, 27 January 2011



Beware uncertainties in simulations below z~25: both rosy and dark scenarios

Pritchard & Loeb 2009; adapted by Koopmans

L. Greenhill



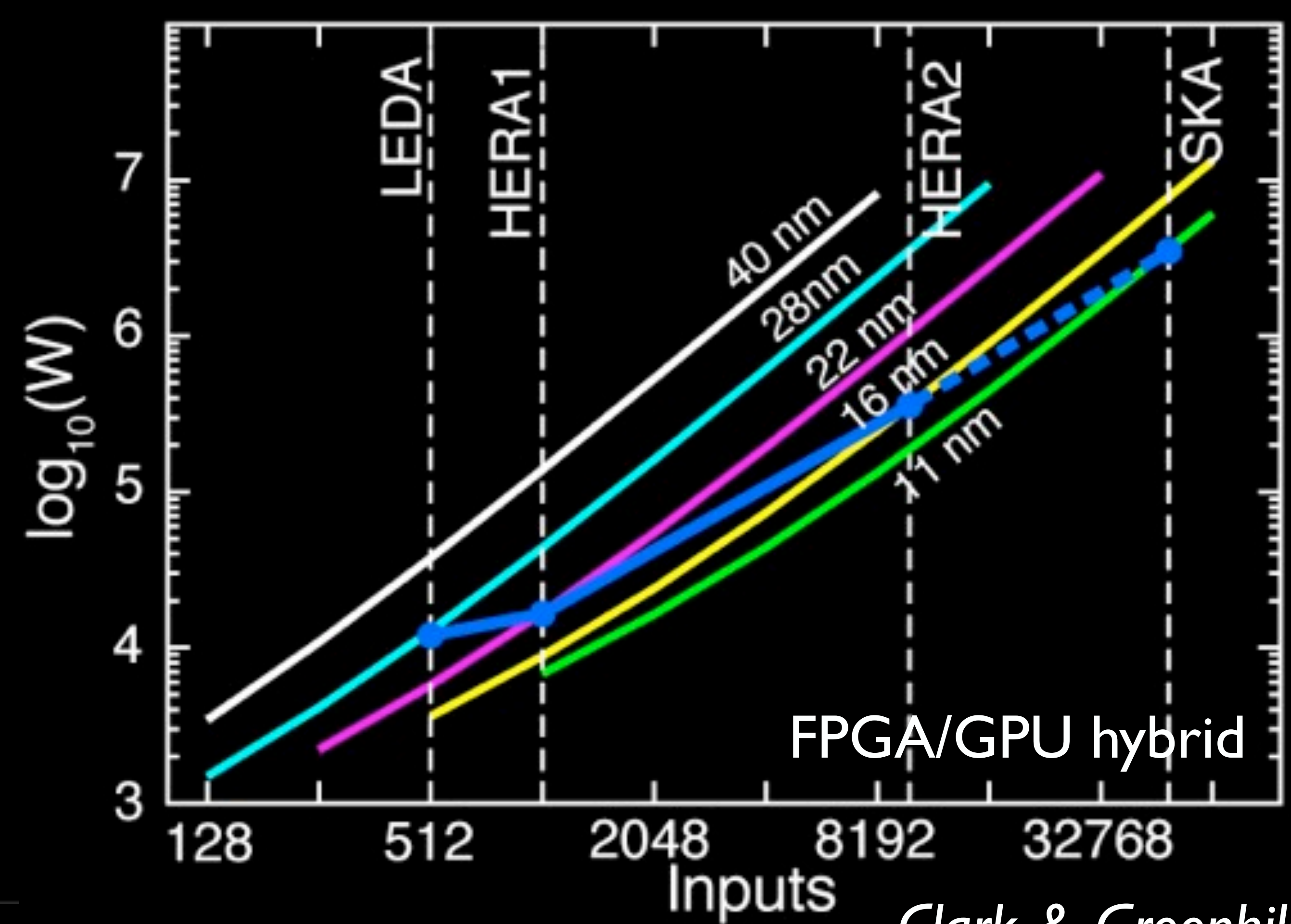
k=0.01 Mpc-1
k=0.1 Mpc-1
k=1.0 Mpc-1
k=10.0 Mpc-1
k=100.0 Mpc-1
Θ[arcmin]
z
l
10³
10²
10
1
0.1
0.01
1
10
100
10³
10²
10³
10⁴
10⁵
10⁶

Practical DSP

e.g., correlation cost

Parsons

e.g., correlation power

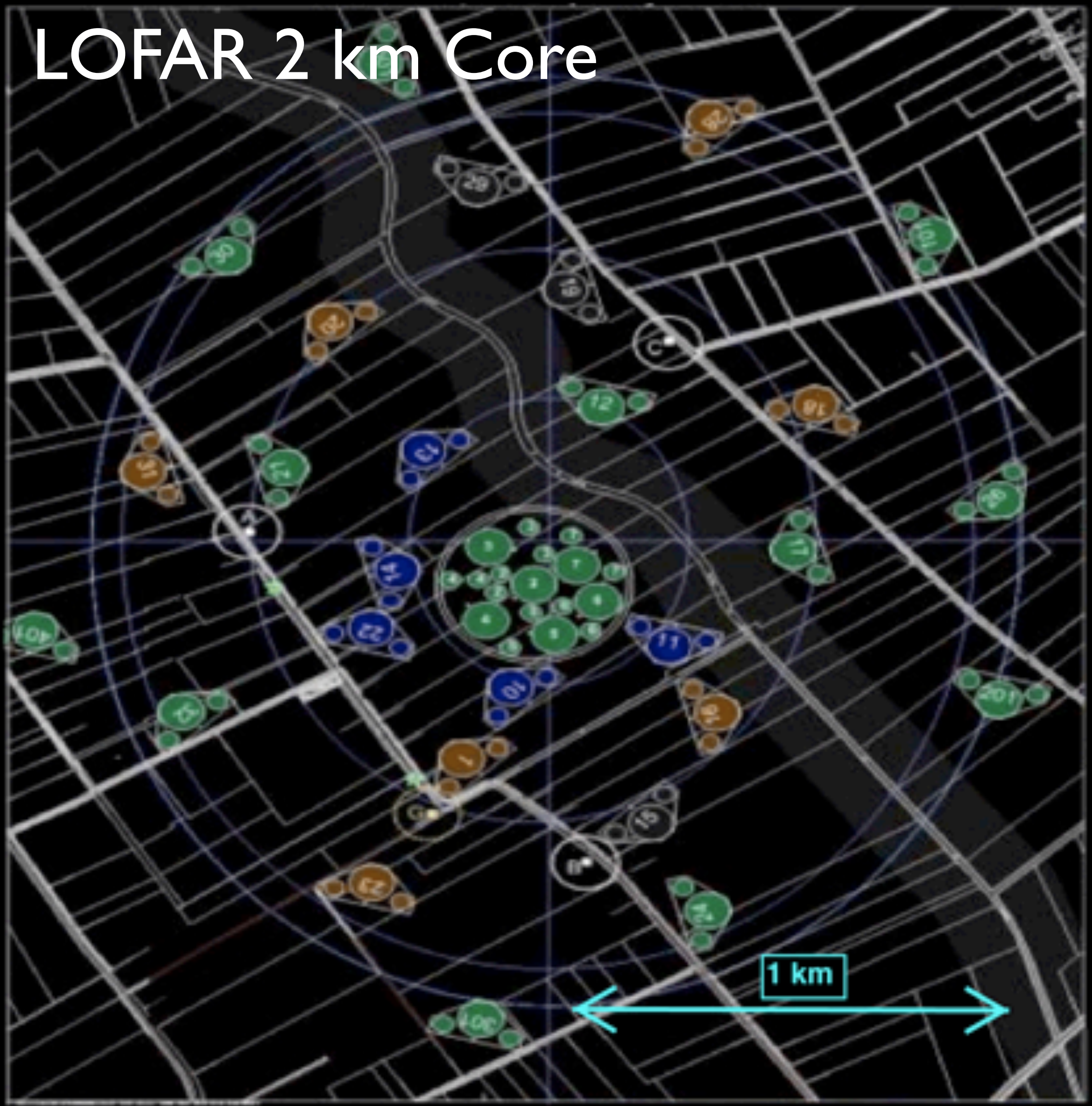


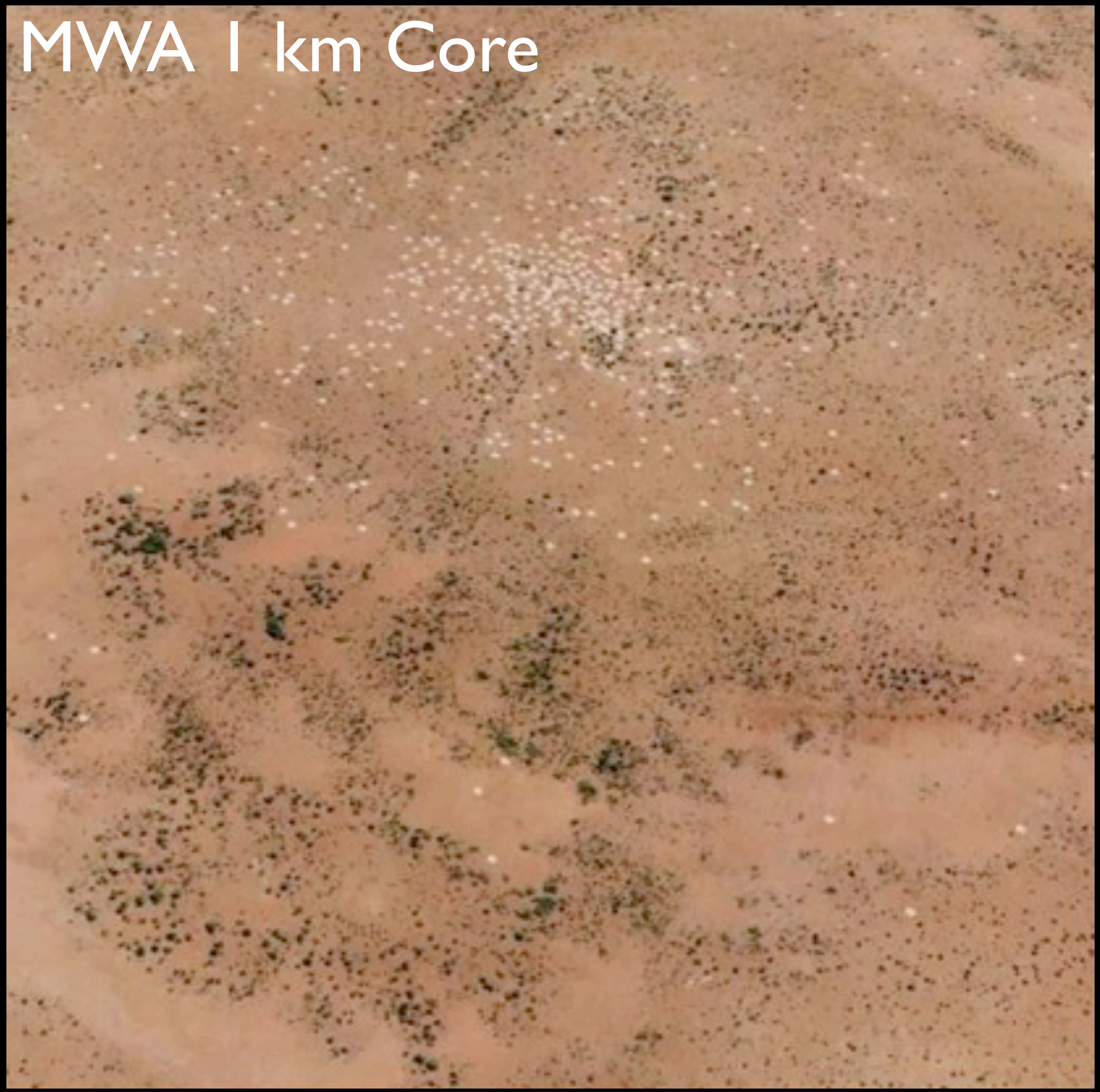
Clark & Greenhill

Synchronize deployment to hardware N-folding times

L. Greenhill

Hierarchical Layouts





Spec.	Driver	SKA_1-Lo	HERA-II
A_{core}		>250,000 m^2 LOFAR-like layout	100,000 m^2 full correlation possible
$T_{rx+ant} < T_{sky}$	sky noise dominated	$< 290\ (^{\nu}/_{150})^{-2.6}$ K	$< 290\ (^{\nu}/_{150})^{-2.6}$ K
$B_{core\ (max)}$	EOR PS O(10^3) h 150 MHz	5 km	3 km
$B_{outer\ (max)}$	point sources & ionosphere	200 km	N/A
A_{core}/T_{sys}	power spectra & some imaging	O(10^3)	~ 350
$FoV_{150\ MHz}$	sidelobes, variance, ...	N x (5 - 20°)	30°
$\theta_{PSF\ 150\ MHz}$	EOR PS	1.5´	3´
Bandwidth	EOR PS	(50) 70 - 200 (450) MHz z ~ 6 - 19 (27)	80 - 200 MHz z ~ 6 - 17
Spectral resolution	RFI, Faraday Rot.	1 kHz	10 kHz

Adapted from Koopmans - PrepSKA WP2 presentation of SKAφ1

L. Greenhill

Scale of HERA-II Cost

Sub-system	units	-$2009
RX	625	$15M
4x4 tile + balun + screens	5000	$8.0M
clock	625	$0.7M
FX corr.	1	$5M
real-time computer	1	$5M
beamformer	5000	$3.8M
cables	--	$1M

- system model: MWAx10 *(fiducial)*
  - MWA $/m^2 < PAPER $/m^2
  - c. 2009 estimates
- construction ~ $40 M
- management ~ $1.5 M
- operations ~ $6 M (3 yr)
- science ~ $7.5 (3 yr)
- reserve ~ $2 M
- R&D NRE ~ $20 M (2 yr)
- infrastructure ~ O($20M)?

~ $100M

L. Greenhill

Layouts

- requirement: filled *u,v* plane (e.g., MWA ~ 300m in 8^s)
  - via snapshot (MWA)
  - via synthesis (LOFAR)
- single-tier compact array PAPER x 100
- two-tier compact array MWA x 10
- multi-tier extended synthesis array LOFAR x 10
- *independent* compact arrays 100 x PAPER
  - boosts area, not dynamic range & FOV; "super-superterps"
- outriggers to compact core(s) e.g., LOFAR
  - different core/periphery apertures

L. Greenhill

Computation as Linchpin

	LOFAR (c)	MWA 512T	HERA-II
Correlation	44 TFlop s^{-1}	160 TFlop s^{-1}	16 - 120 PFlop s^{-1}
Calibration/ imaging	10-100 TFlop s^{-1} ? post real-time	50-200 TFlop s^{-1} real-time	10 - 100 PFlop s^{-1} real-time / post real-time ?

- array characteristics
  - N: antennas or tiles
  - B: bandwidth (# of ch.)
  - F: field of view
  - S: array geographic size
- correlation $\propto k_0 N^2 F B + k_1 N B$
- calibration & imaging $\propto k_3 N^2 B + k_4 B (F S)^{1-2}$
- storage/data management: 1.5 km array, 512 ant, 30° FOV ➟ 3 PB/week
  - image plane analysis becomes attractive BUT
  - output rate can be ~ input rate from correlator - depends on informaton-loss tolerance
  - image-based algorithms require extensive develoment

L. Greenhill

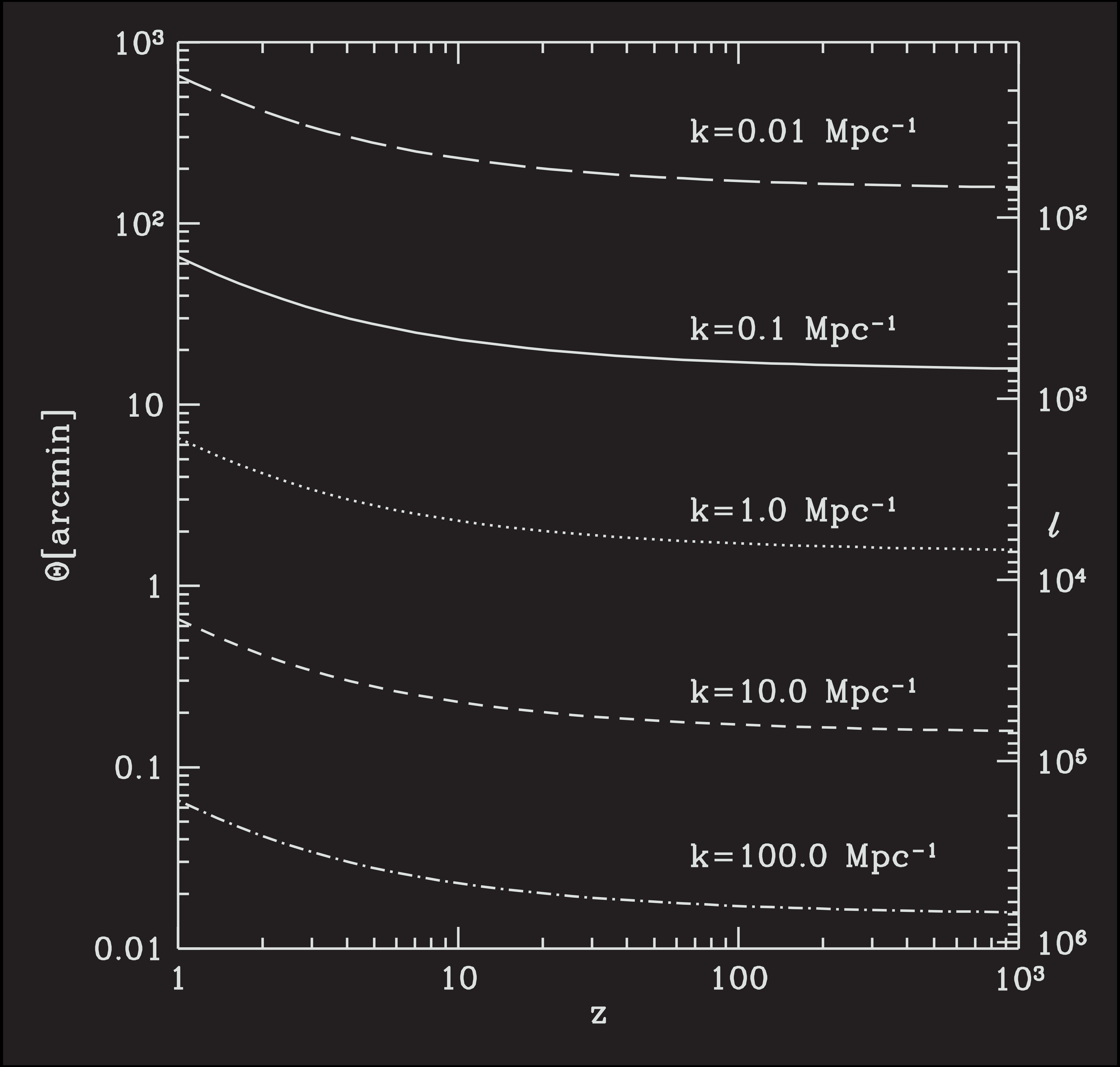
Getting to HERA

- Outside SKA framework
  - construction timelines relatively similar
    - neatly channel dual-track efforts into creation of SKAφ2
  - HERA operation on non-selected SKA site
    - requires host investment in infrastructure
    - HERA as vehicle for win-win scenario
  - International cooperation: reviews by HERAtics, SKAers (?)
    - coordination & cooperation enables transition to φ2
    - joint technical reviews once project plans in place
  - provides time to expand recognition of SKA brand in US

L. Greenhill

HERA Immediate Open Q.'s

- attack Dark Age / EOR transition?
- baselines > 3-5 km?
- how to merge MWA & PAPER engineering & designs?
- peta-scale DSP, algorithms, computation, and storage
  - maximize use of "off-the-shelf" to minimize cost?
  - to what extent must HERA-II use real-time processing?
  - what time-line does $O(N^2)$ scaling & technology enforce?
    - exclude all but tried / true, lowest-risk approaches?
    - consider Nlog(N) appoaches as early as HERA II?
- source of funding? dovetailing with $SKA_{1,2}$ program



10³
10²
10
1
0.1
0.01
Θ[arcmin]
k=0.01 Mpc⁻¹
k=0.1 Mpc⁻¹
k=1.0 Mpc⁻¹
k=10.0 Mpc⁻¹
k=100.0 Mpc⁻¹
10²
10³
10⁴
10⁵
10⁶
l
1
10
100
10³
z

Getting to HERA

- Within SKA framework
  - construction timelines relatively similar
  - SKA design process begins comparatively early
    - HERA groups seek PAPER & MWA science first
      - limited manpower motivates narrow focus
    - science efforts build design lessons learned second
      - after primary science phase, PAPER & MWA become testbeds for prototypes pointing toward 10^5 m^2 array
      - how to filter-in PAPER & MWA lessons into the process?
  - NSF/AST starved; funding scheme on paper only; ∃ others?

L. Greenhill