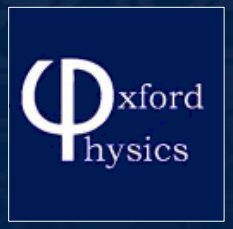


Challenges and Opportunities for Massive Parallism

Stef Salvini

Stef.salvini@oerc.ox.ac.uk

Talk Outline

- Needs for HPC (High Performance Computing)
  - Scientific Opportunities
- HPC: current status and trends
  - Systems layout and architectures
  - Need for Massive Parallelism
- Challenges
  - Software
  - Energy
  - Fault tolerance
- The Square Kilometre Array
  - An extreme MP computational challenge

Questions …

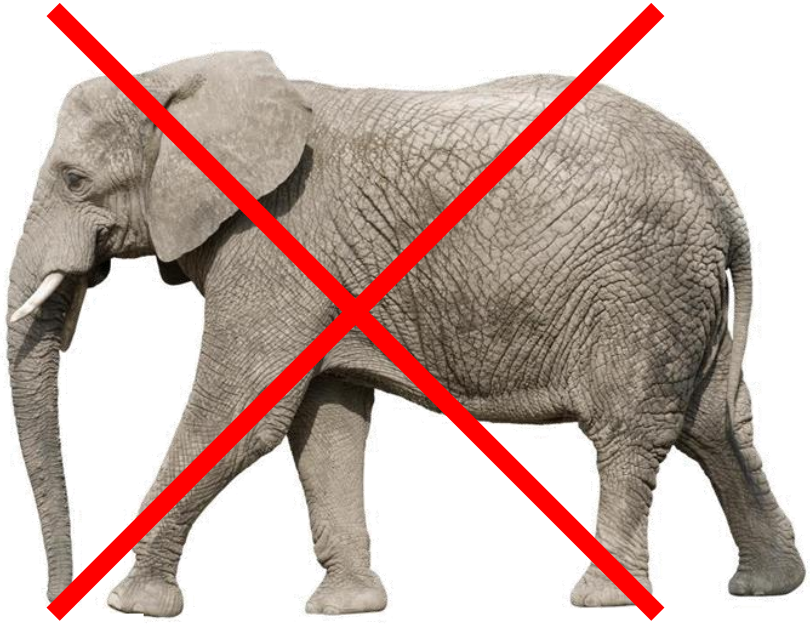
How many cores in your smart phone?

Ans: 10s

How many cores in your Mac Book Pro?

Ans: 100s

MPP is the art of moving heavy objects with ….

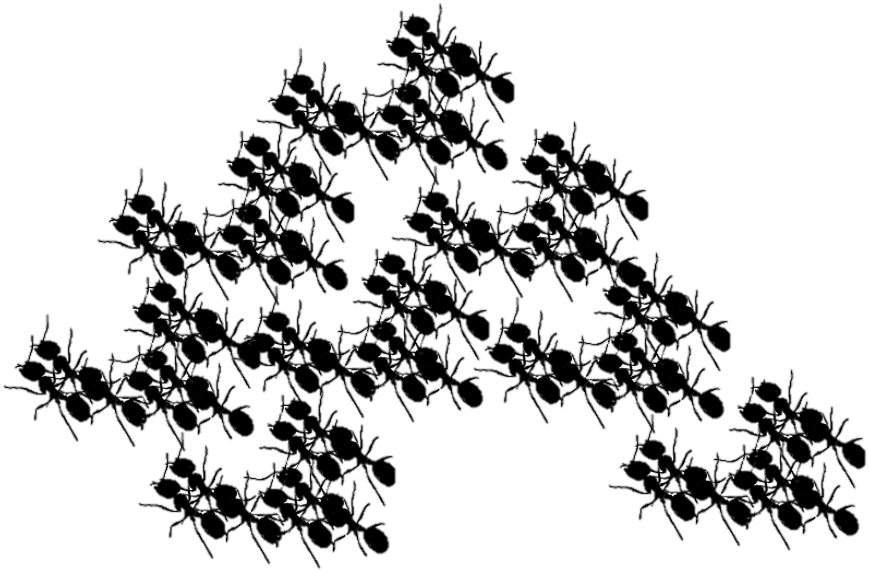


Elephants ..?

MPP is the art of moving heavy objects with ….

Mice …?

MPP is the art of moving heavy objects with ….



Ants!!!

Lots of them!

Need for HPC

- Increasingly large scientific computation
  - Cell biology
  - Climate
  - Simulations (nuclear, aerospace, energy, etc.)
  - Nano technology
  - Multiscale science
  - Physics, Chemistry, Material Science, etc
- Big Data
  - Powerful instruments → scientific data products (knowledge)
  - Humanities
  - Finance
  - Government (including classified …)

What makes up high-end HPC systems

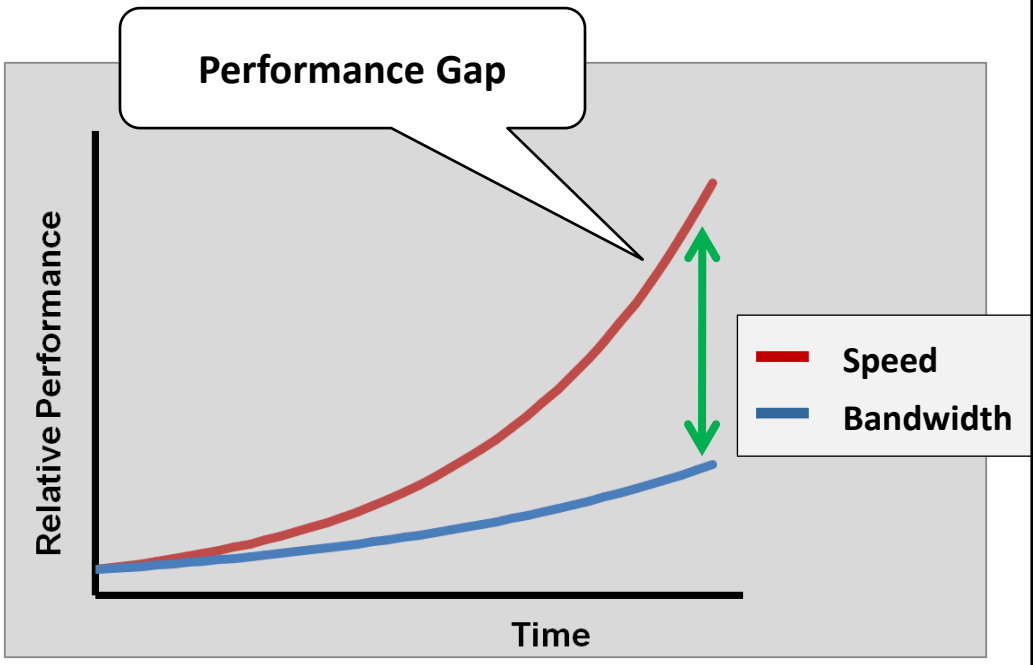
- Large concentration of small components: why?
- Energy
  - Power consumption ~ clock^3
- Design
  - 1,000,000,000s components per chip (22-14-…-8 nm)
  - much easier to use modular design
  - Miniaturisation (8-14-22 nm)
- Hierarchical structure
  - Core ➔ CPU (GPU) ➔ Node ➔ Rack ➔ System

Technological Trends

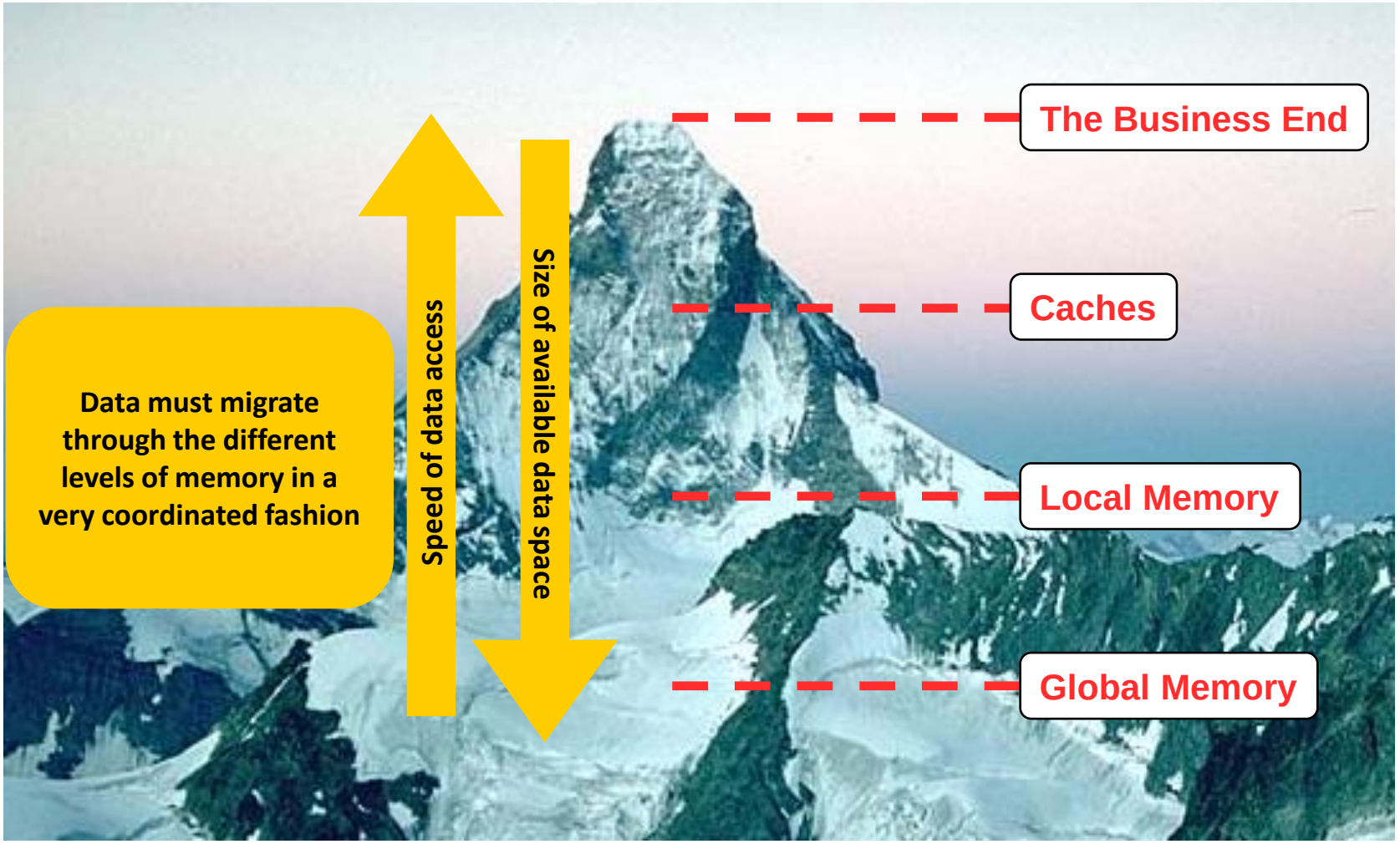
- Moore's Law: CPU speed doubles every 18-24 months
  - Gordon Moore of Intel, 30 years ago
  - Now achieved by multi-cores
  - Trap or achievement?
- Similar growth (but, critically, at a slower pace) in
  - memory size
  - Bandwidth
  - storage capacities
  - network speed

Mind the Gap …

- Moore's Law: CPU speed doubles every 18-24 months
  - Gordon Moore of Intel, 30 years ago
  - Now achieved by multi-cores
  - Trap or achievement?
- Processor speed and Bandwidth are both increasing steadily
  - Bandwidth rate of increase is lower
  - Performance gap is increasing



Data Access or Lack Thereof



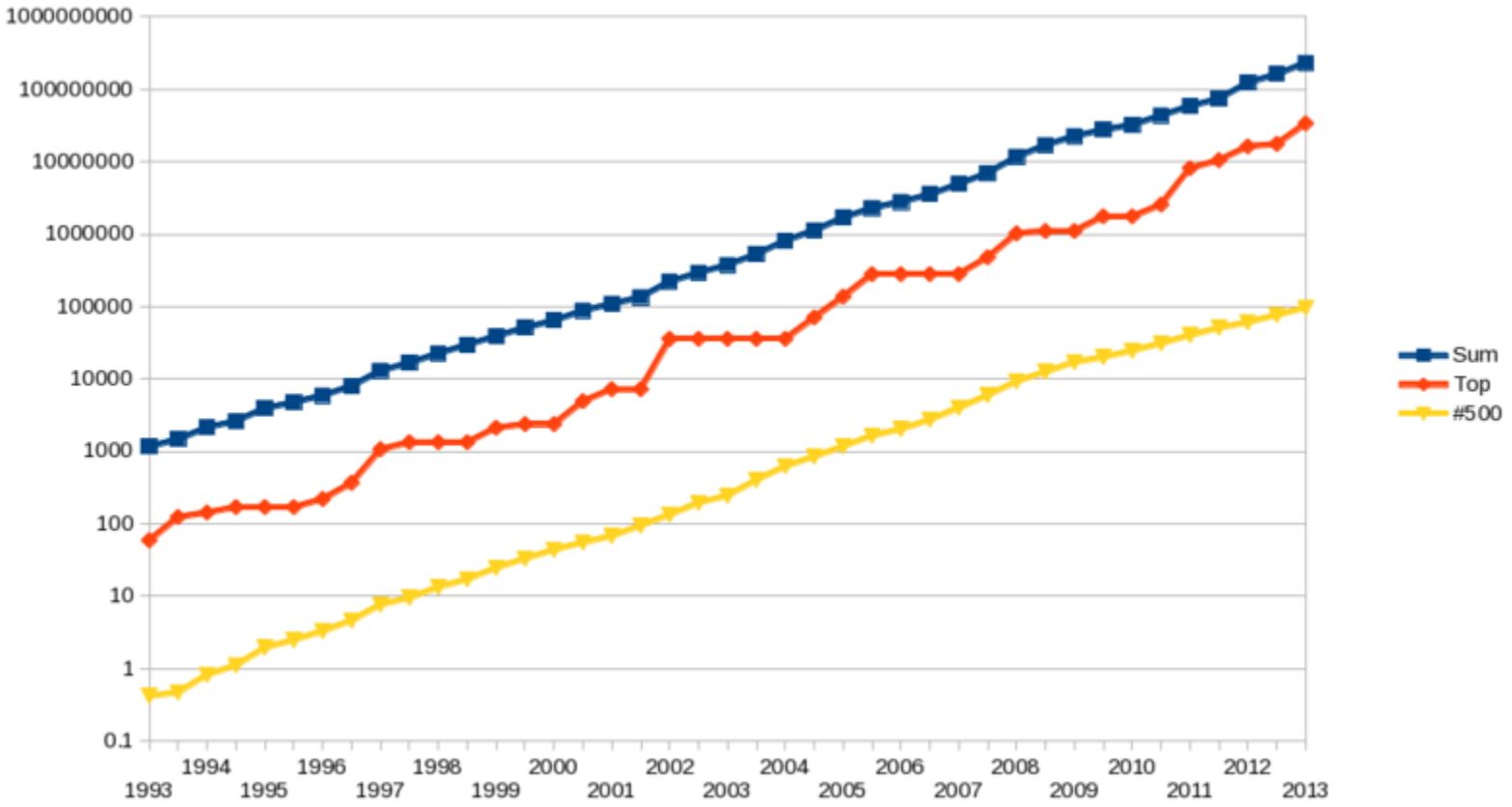
Trends 1

- Energy efficiency **ESSENTIAL**
  - power O(frequency^3)
  - clock race is over!
  - ➔ performance = massive parallelism
- Performance is achieved through parallelism:
  - Moore's Law continues ➔ much more circuitry on each chip
  - More cores per chip
- Vector processing back in fashion
  - ➔ cores work in small groups: same ops, different data
- NVIDIA GPUs
  - cores work in groups of 32 (a thread warp)
- Intel Xeon Phi
  - ~ 60 cores, each 8-16 long vectors

Trends 2

- Multithreading is very important:
  - CPU cores: out-of-order execution and branch prediction to maximise performance and avoid memory stalling
  - many-core chips: simple in-order execution cores, multithreading to avoid memory stalling (n. threads >> n.cores)
- Data movement is key to performance:
  - 200-600 cycle delay in fetching data from main memory
  - many applications are bandwidth-limited, not compute limited
  - (in double precision, given 200 GFlops and 80 GB/s bandwidth: 20 ops/variable to balance computation and I/O
  - Power / time to move data across a chip > 1 Flop

Top 500


1000000000
100000000
10000000
1000000
100000
10000
1000
100
10
1
0.1
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Sum
Top
#500

Current top 5

System	Where	Node	No. Nodes No. Cores	PFlops
Tianhe-2	Guangzho, China	2 x Intel Xeon 12C + 3 x Intel Xeon Phi (57 cores)	16,000 3,120,000	33.86
Titan	ORNL, USA	1 x AMD Opteron 16C + Nvidia K20x (14 SMS/ “cores”)	18,688 560,640	17.59
Sequoia	LLNL, USA	IBM Blue Gene/Q, PowerPC 16C	98,304 1,572,864	17.17
K Computer	RIKEN, Japan	SPARC64 8C	88,128 705,024	10.51
Mira	ARNL, USA	IBM Blue Gene/Q, PowerPC 16C	49,152 786,432	85.87

Intel Xeon Phi (Tienhe-2)

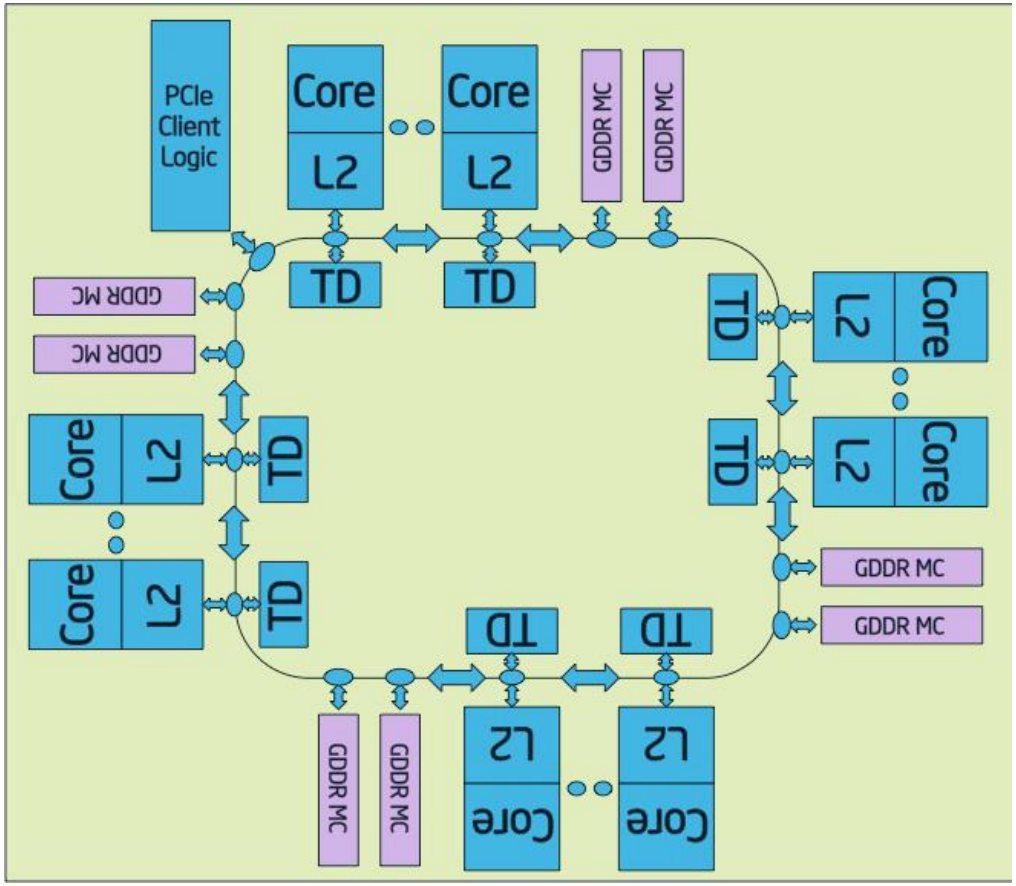
Intel Xeon Phi 31S1P

8GB global memory

57 cores

512b-long vectors

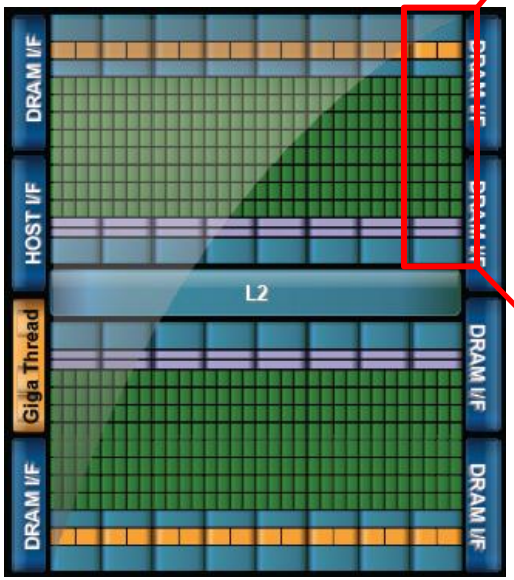
Peak performance: ~ 1,003 GFlops



Nvidia Tesla GPU (1 per node in Titan)

Kepler Tesla K20

6GB global memory
48kB shared memory
192 cores per SMX
14 SMX
Peak performance: 1,310 GFlops



Massive Parallelism: good and less good

- **GOOD**: Decomposition in many independent tasks
  - Google? Communications networks? CERN Grid?
- **OK**: Semi-decomposable tasks
  - Low rate of inter-partition communications
  - Small data “perimeter” to “area”
  - Lattice-Boltzmann, etc.
- **BAD**: “Monolithic” tasks
  - “all-to-all communications” and data exchange
    - 3-D FFTs (e.g. long-range interactions in particle dynamics)

Challenge 1: Computational Organisation & Algorithms

- Inhomogeneous computation
  - CPU, GPU, Xeon Phi and their ilk, FPGA, etc
- Multi-level hierarchy
  - Not adequately mapped to software
- Sustainability (“Future-proofing”)
  - Need to re-use software components!
  - Need to reuse software on new generation hardware
- Fault-tolerance
  - Components will fail. How to cope with that?
- Algorithms
  - New generation required
  - Multi-level algorithmic/data decomposition

Challenge 2: Software and software tools

- Programming languages
  - Too close to hardware level (cfr. CUDA, OpenCL)
    - Very labour intensive
    - Needs considerable reworking for new hardware
- Auto-tuning?
  - The philosophal stone!
- Development tools for 1,000,000 computing engines:
  - Debuggers
  - Profilers (to identify reliably bottlenecks, etc)

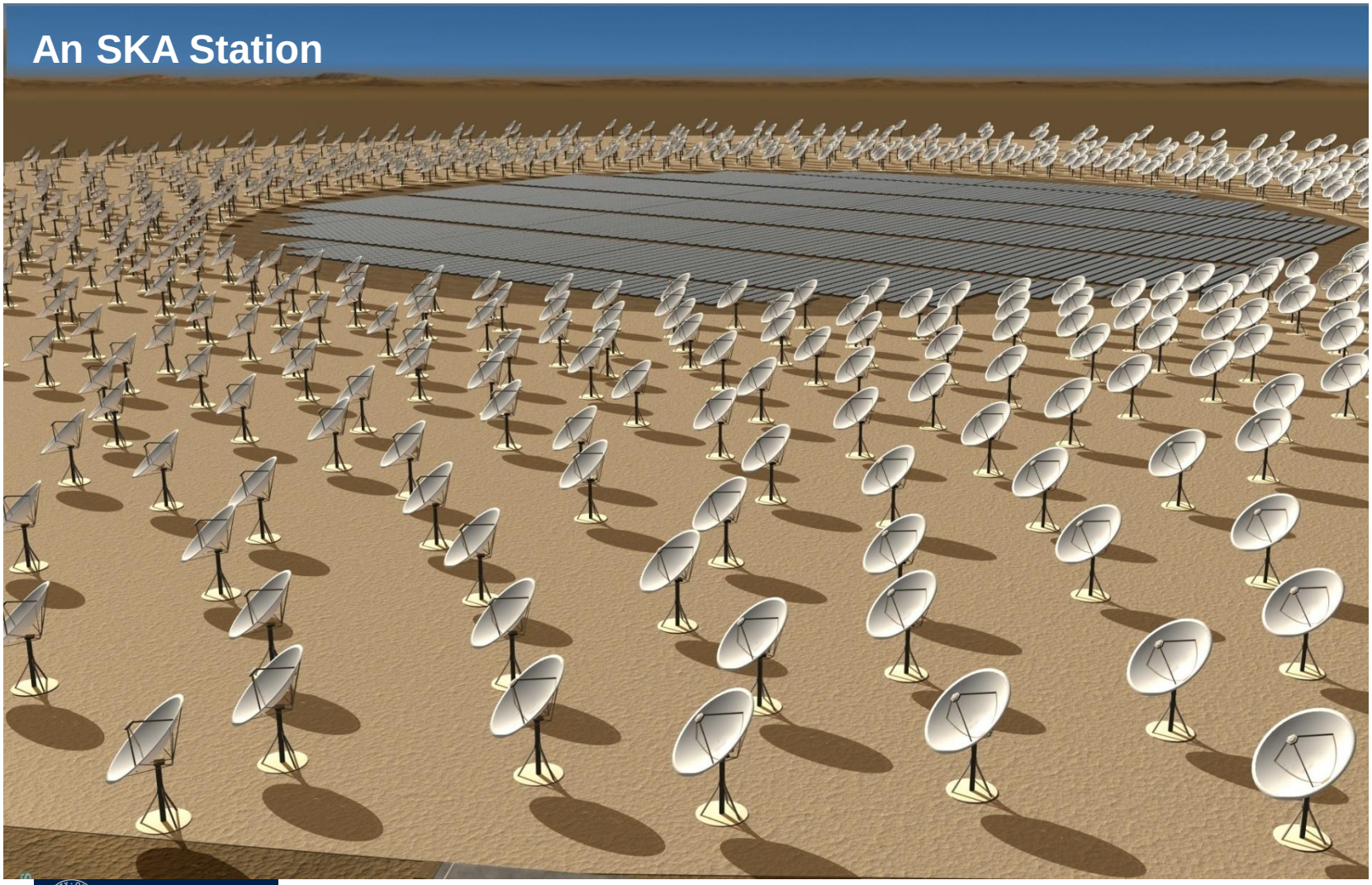
To Exaflop or not to Exaflop

- Exascale or Exaflops
  - Many problems are Exascale
    - E.g. SKA
  - Is there a real need for Exaflops boxes?
- EESI European Exascale Software Initiative
- International Exascale Software Project (IESP)
- Power envelope all important
  - Current technology requirements (excluding memory, etc.)
    - “Standard” CPUs (~ 1 Watt / Gflop) ➔ 1,000 MW
    - GPUs (~ 0.1 Watt / Gflop) ➔ 100 MW

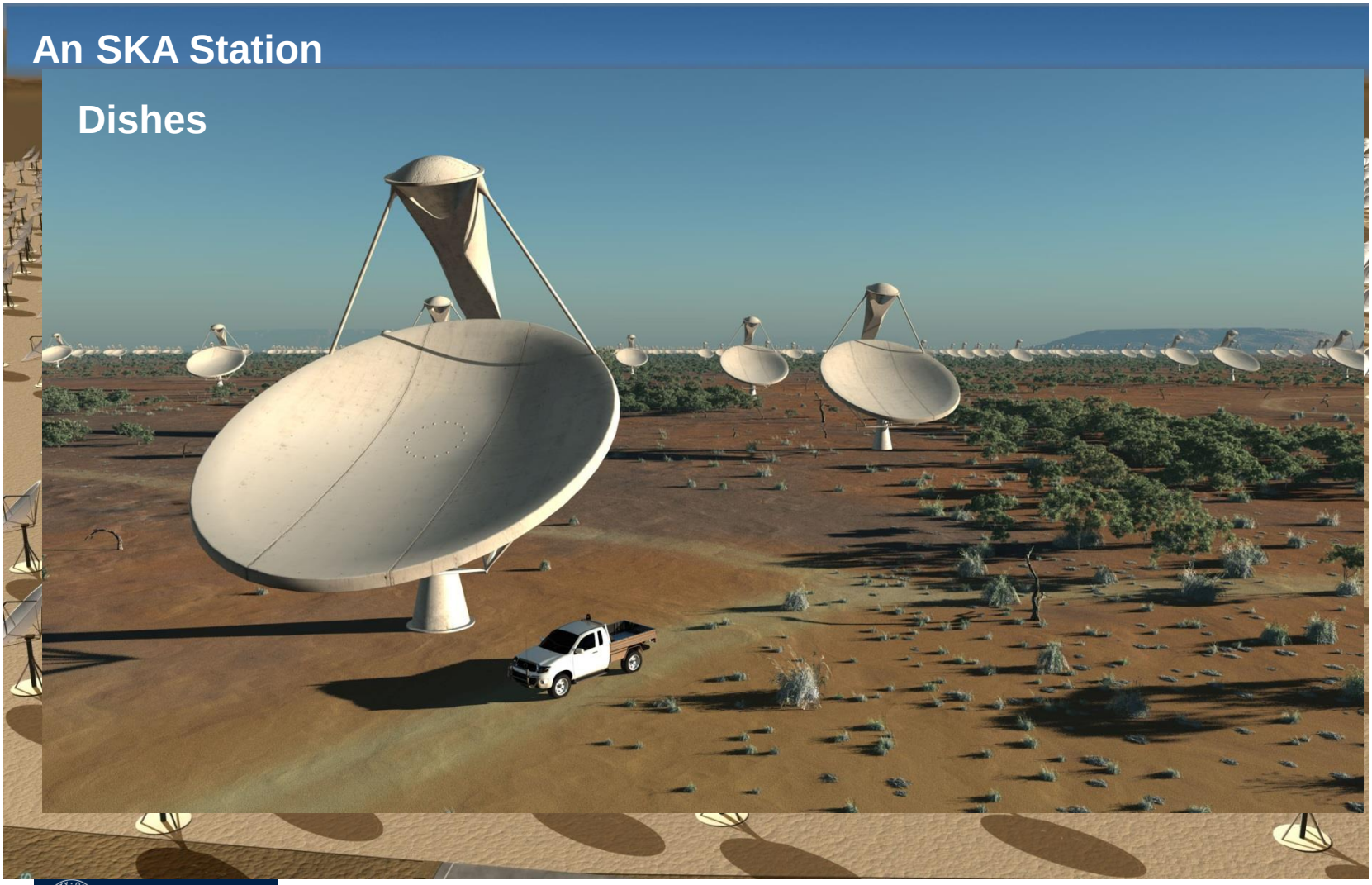
What is the Square Kilometre Array (SKA)?

- Next Generation radio telescope – compared to best current instruments it is ...
  - ~100 times sensitivity
  - ~ 106 times faster imaging the sky
  - More than 5 square km of collecting area on sizes 3000km
- Will address some of the key problems of astrophysics and cosmology (and physics)
- It is an interferometerMajor ICT project

SKA: Artist's Impression

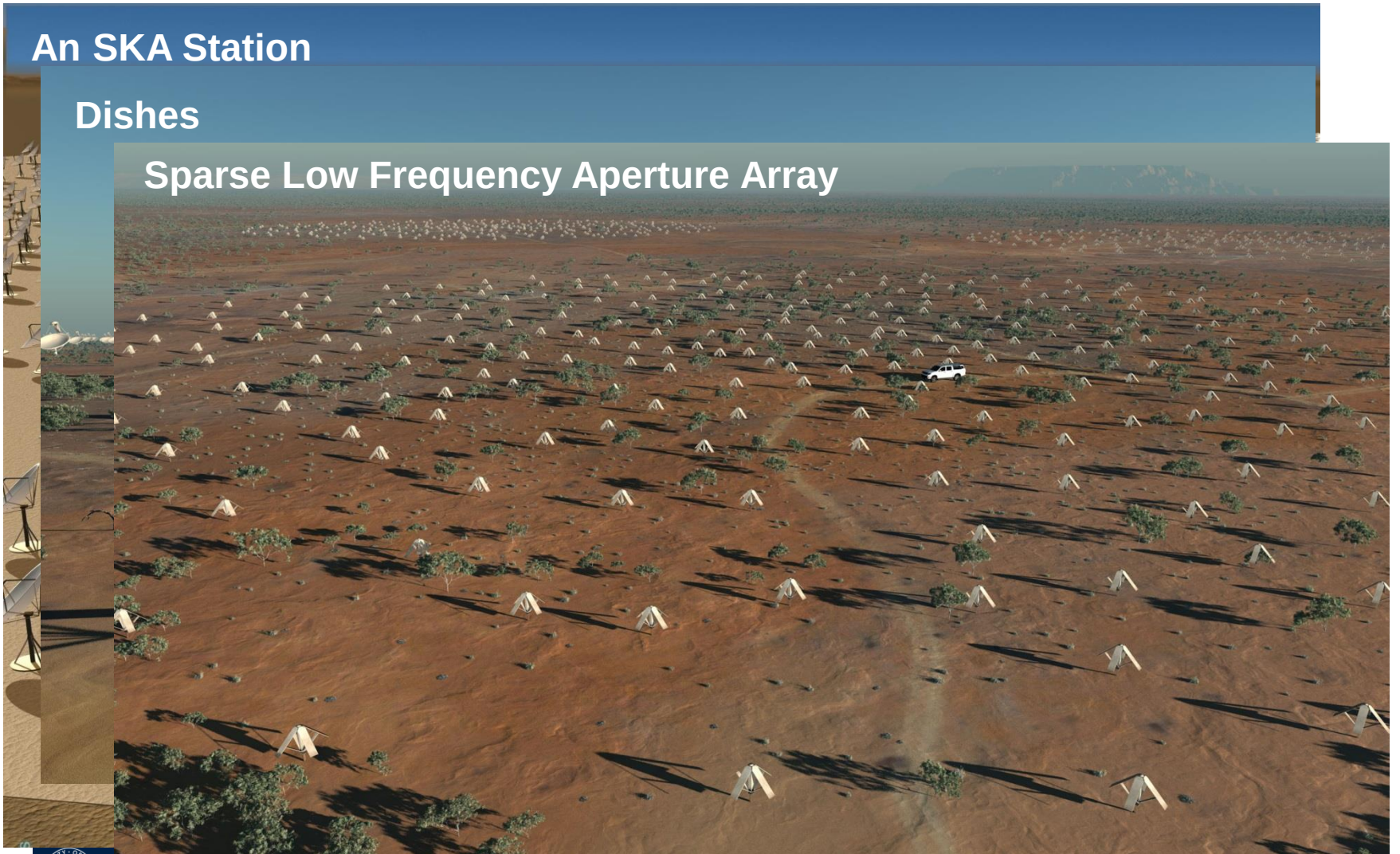


SKA: Artist's Impression

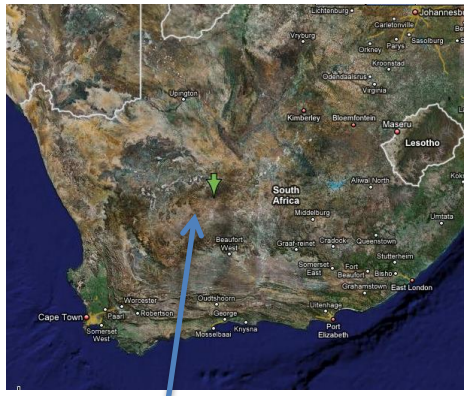


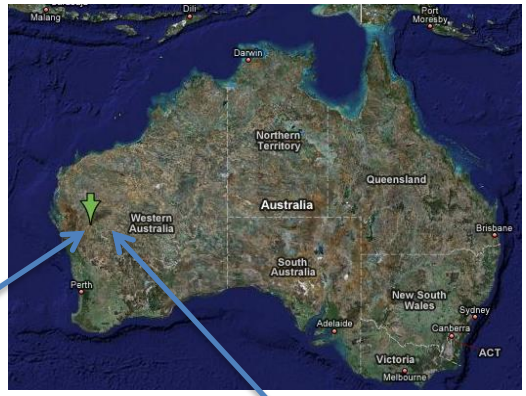
UNIVERSITY OF OXFORD

SKA: Artist's Impression



SKA Phase 1 Implementation





SKA1_Mid incl MeerKAT

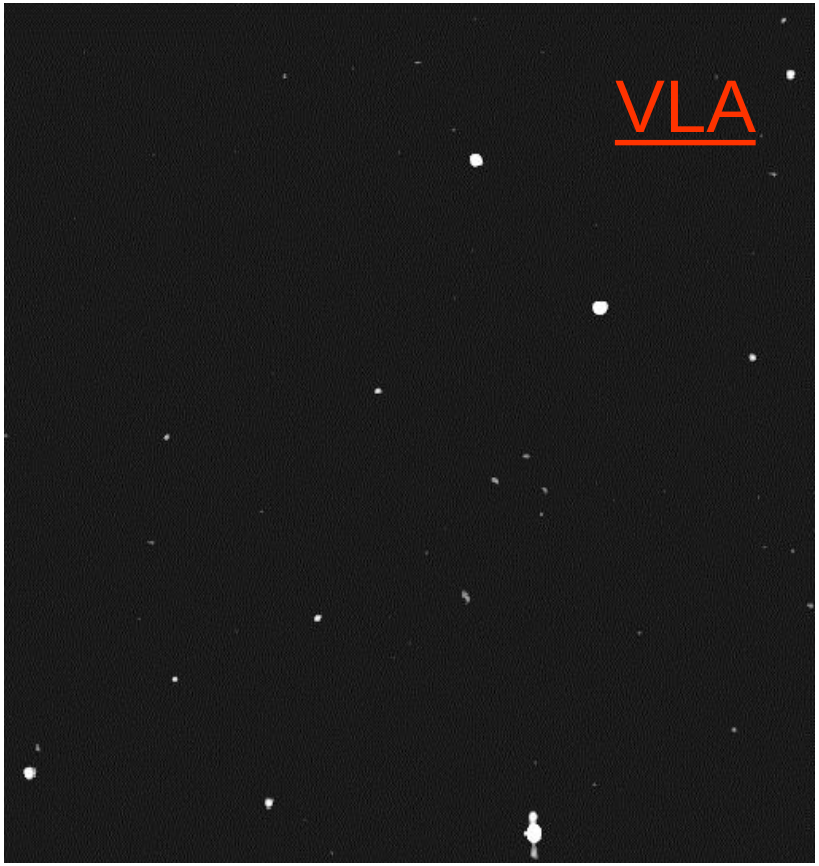
SKA1_Low

SKA1_AIP_Survey incl ASKAP

	SKA Element	Location
Dish Array	SKA1_Mid	RSA
Low Frequency Aperture Array	SKA1_Low	ANZ
Survey Instrument	SKA1_AIP_Survey	ANZ

SKA science drivers: Galaxy evolution back to z ~ 10?

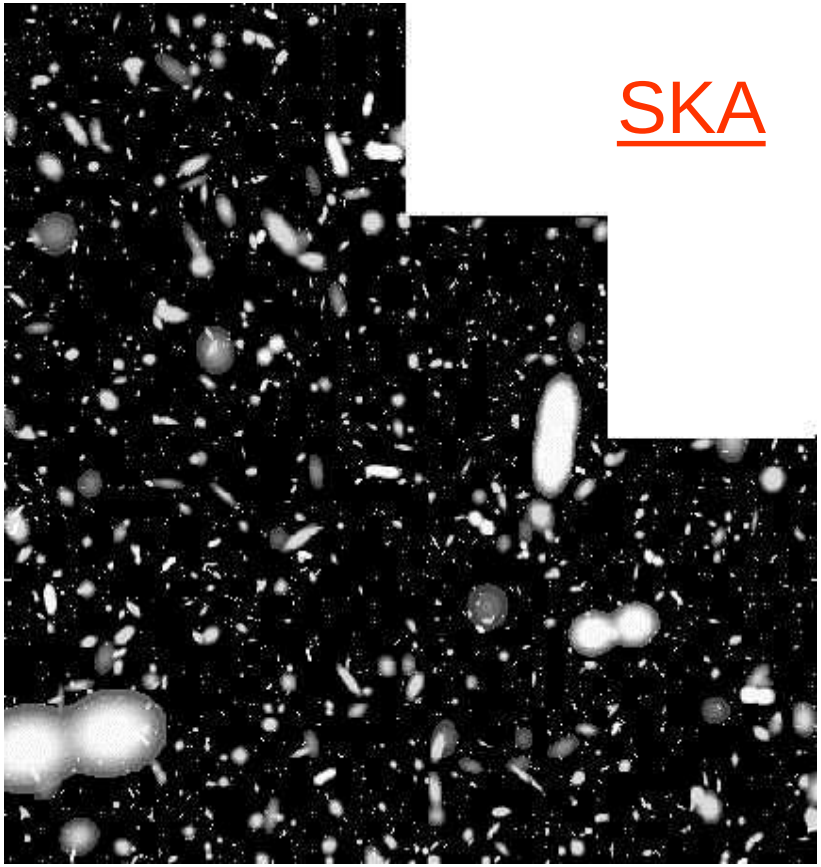
HDF



VLA

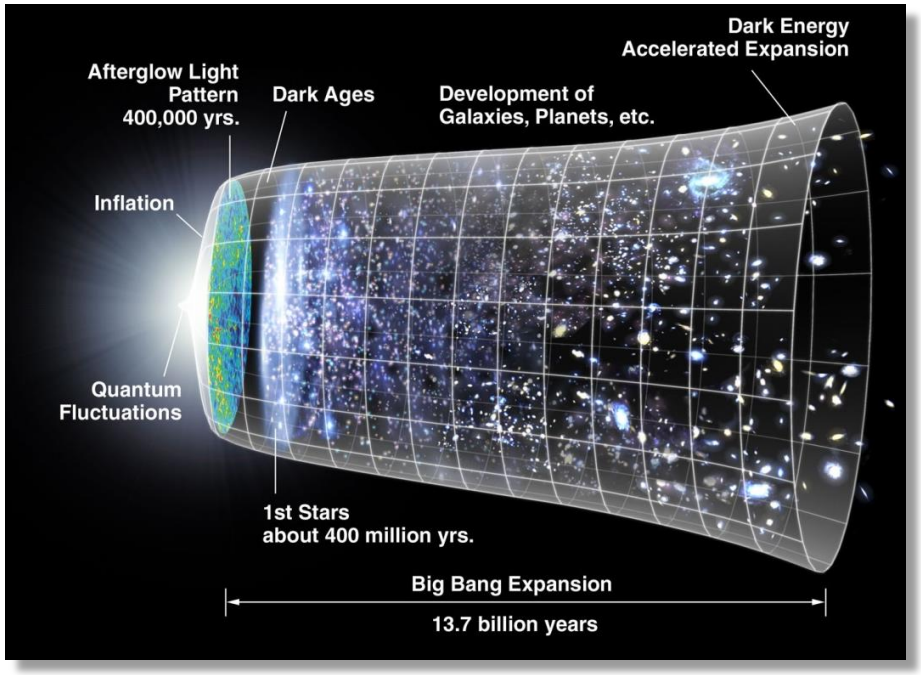
SKA science drivers: Galaxy evolution back to z ~ 10?

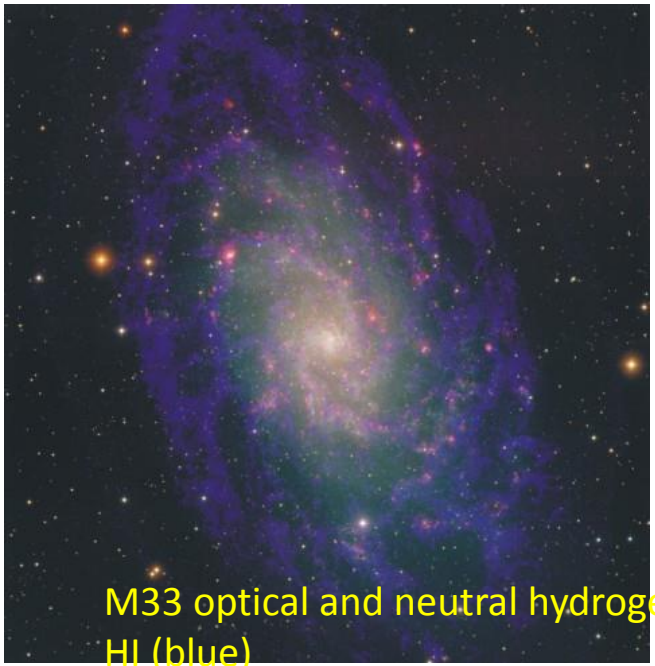
HDF



SKA

Cosmology and the History of Hydrogen





M33 optical and neutral hydrogen, HI (blue)

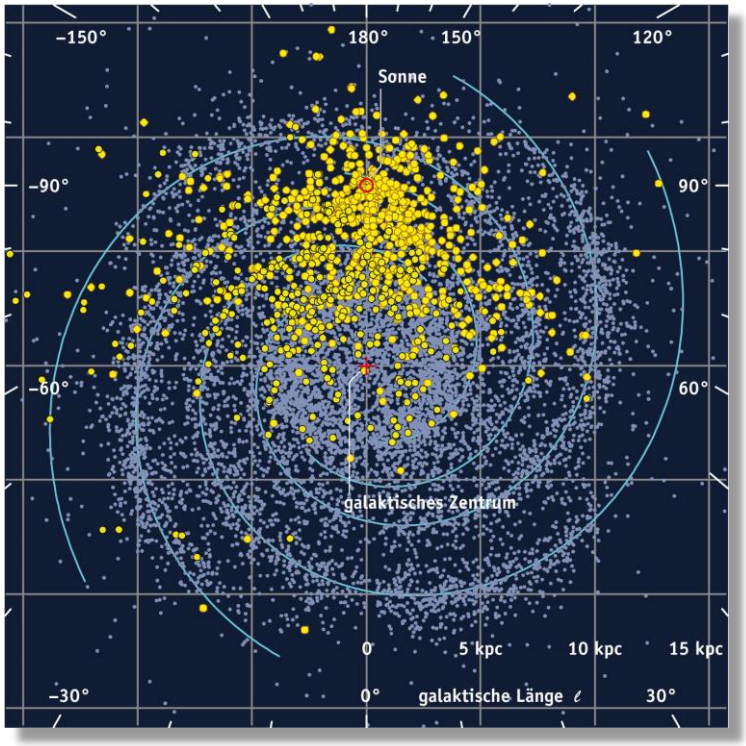
- After recombination (CMB) Universe if neutral, but we know that hydrogen (not in galaxies) is hot and ionised
- Re-ionization occurs when first objects (galaxies and AGN) form via UV- and X-ray emission
- Epoch of Reionisation – EoR next major challenge for Cosmology

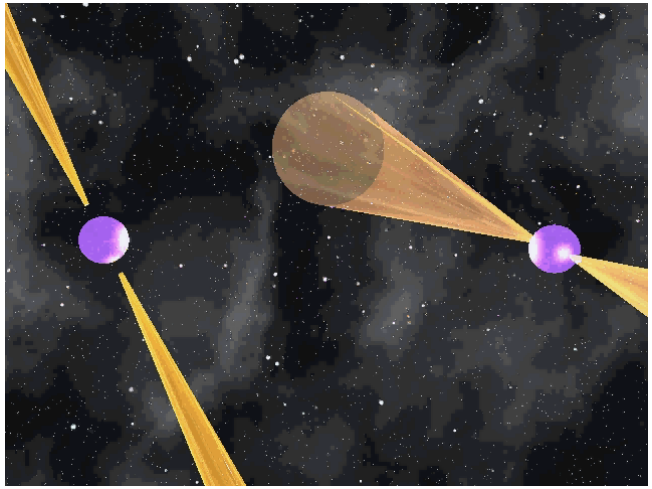
- $^2S_{1/2}$ ground state of HI split by the effects of nuclear spin
- $\Delta E = 5.8 \times 10^{-6}$ eV

21-cm line at 1420 MHz.

Pulsar survey: testing gravity

- The SKA will detect around 20,000 pulsars in our own galaxy
- Relativistic binaries give unprecedented strong-field test of gravity
- Timing net of ms pulsars to detect gravitational waves via timing residuals





UNIVERSITY OF OXFORD

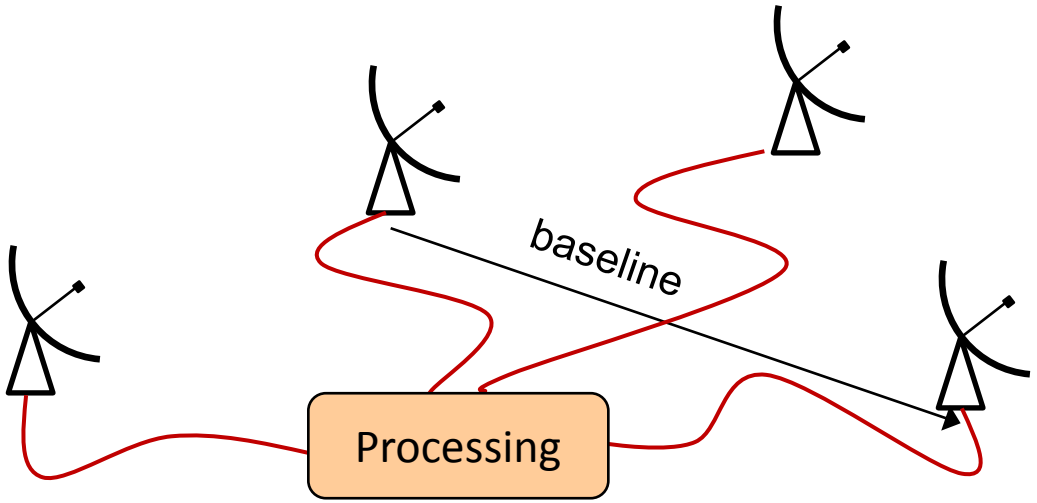
And much more …

- Star formation history of the universe
- Physics of active galaxies, accretion discs, pulsars, jets, ...
- Radio emission from decaying dark matter
- The cosmic web and structure formation
- Measuring changes in the fine structure constant
- Interstellar medium of galaxies
- Astrobiology – direct detection of large molecules
  - ... and evidence for intelligent life?

Most importantly what we haven't predicted

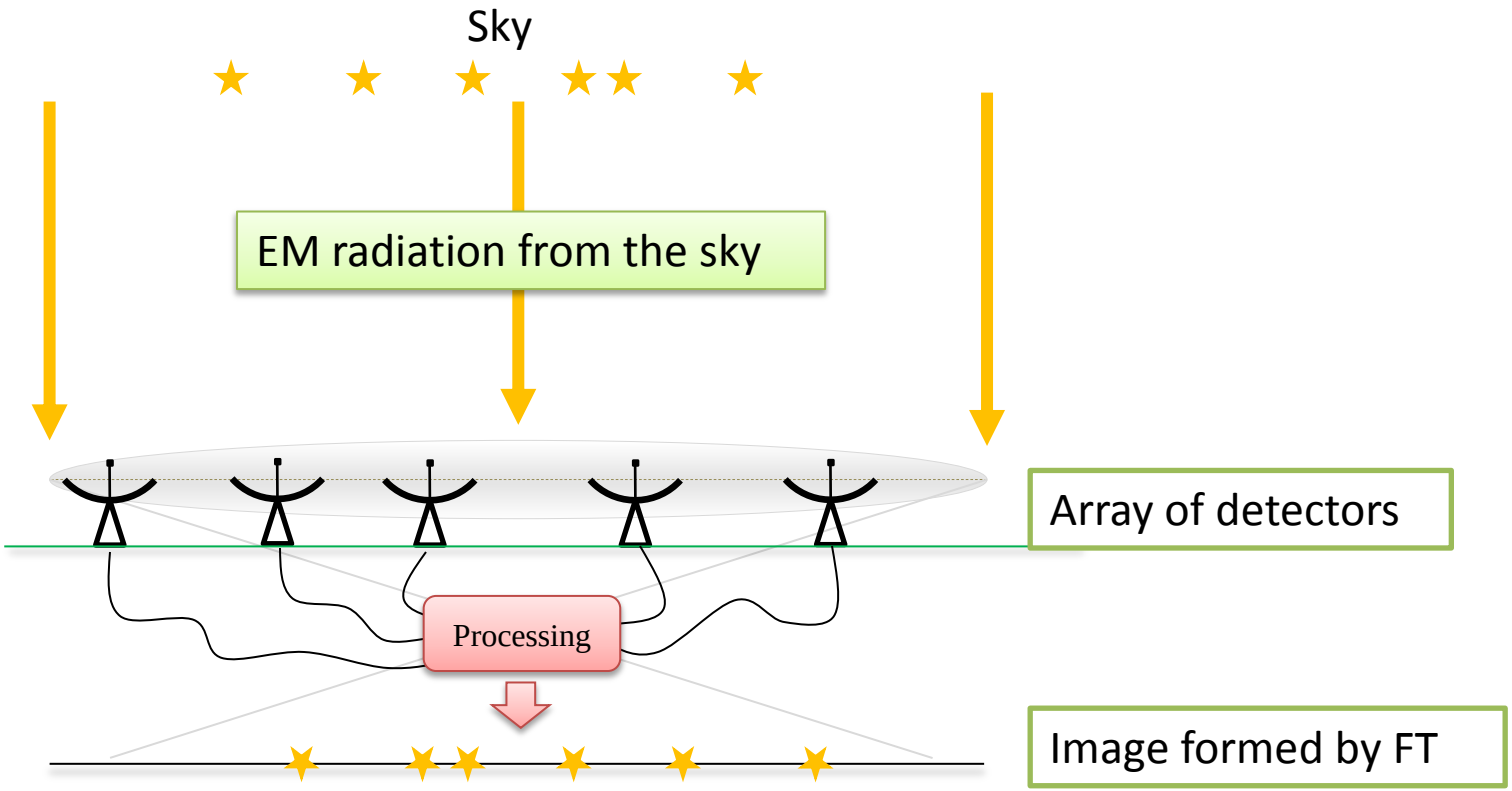
Interferometry

- Combine detectors to increase resolution
- Differential measurement of the same wave front from different and distant receivers
- Time average to reduce noise
  - Earth rotation helps with baseline coverage



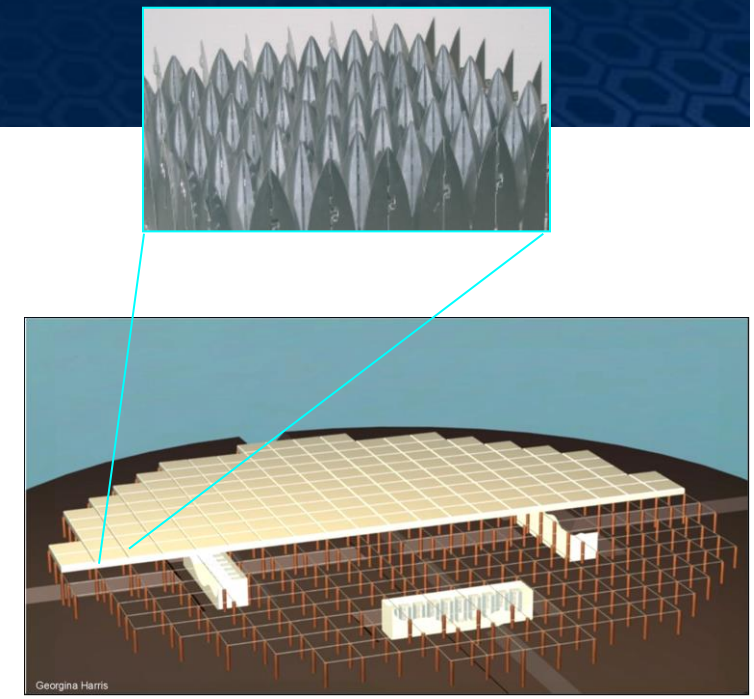
How does this compare to an optical system?

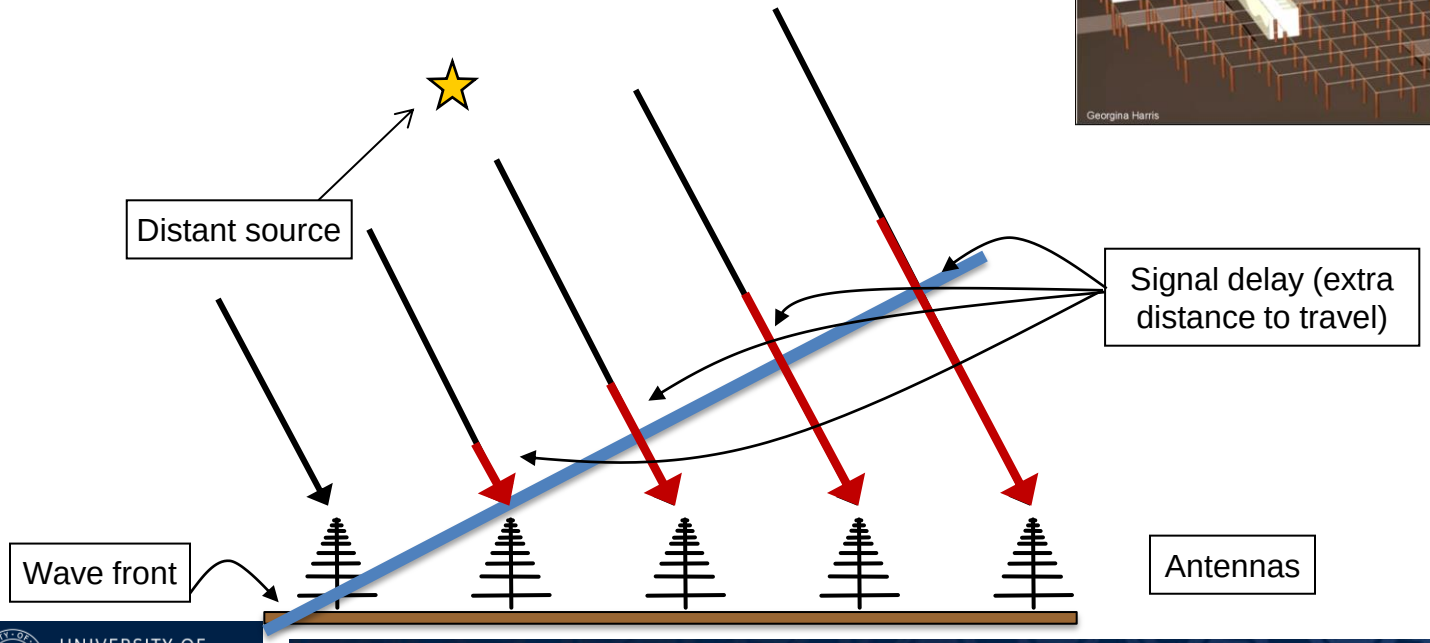
- A radio interferometer samples the wave-front in the Fourier plane and image formation is performed in post processing.



Aperture Arrays

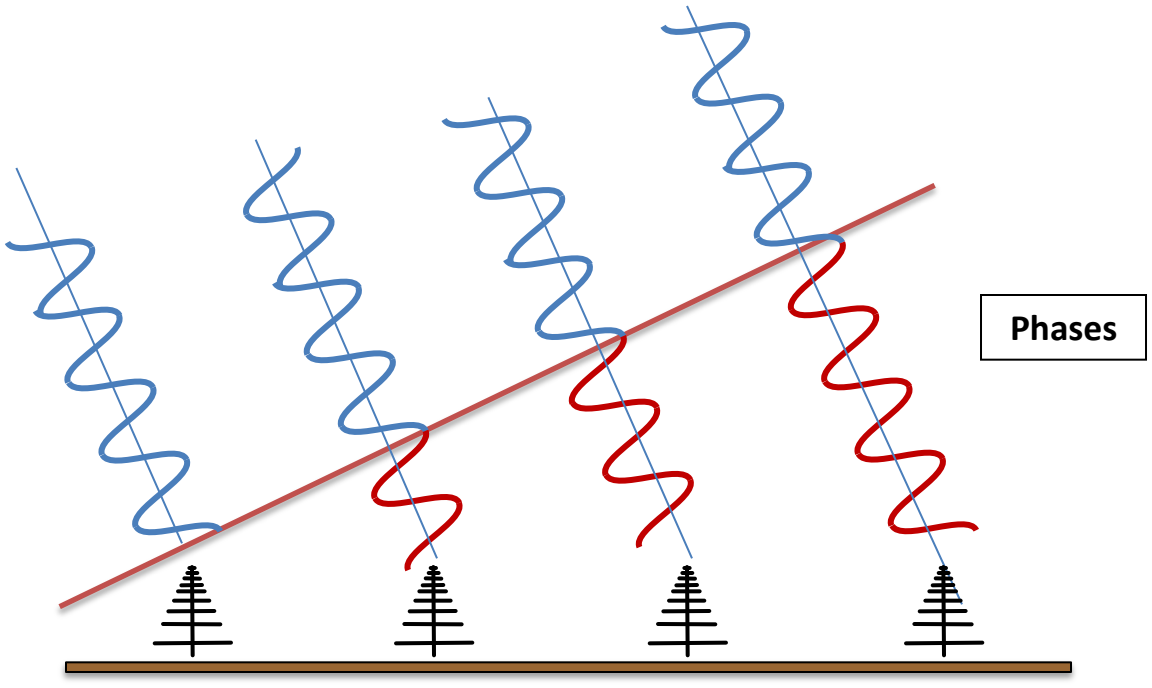
- Collection of omni-directional antennas
- Used technology
  - Radars for avionics
  - Use in RA starting (LOFAR)
- Beam from combination of antenna signals
  - (See next slides)





Frequency Domain Beamforming

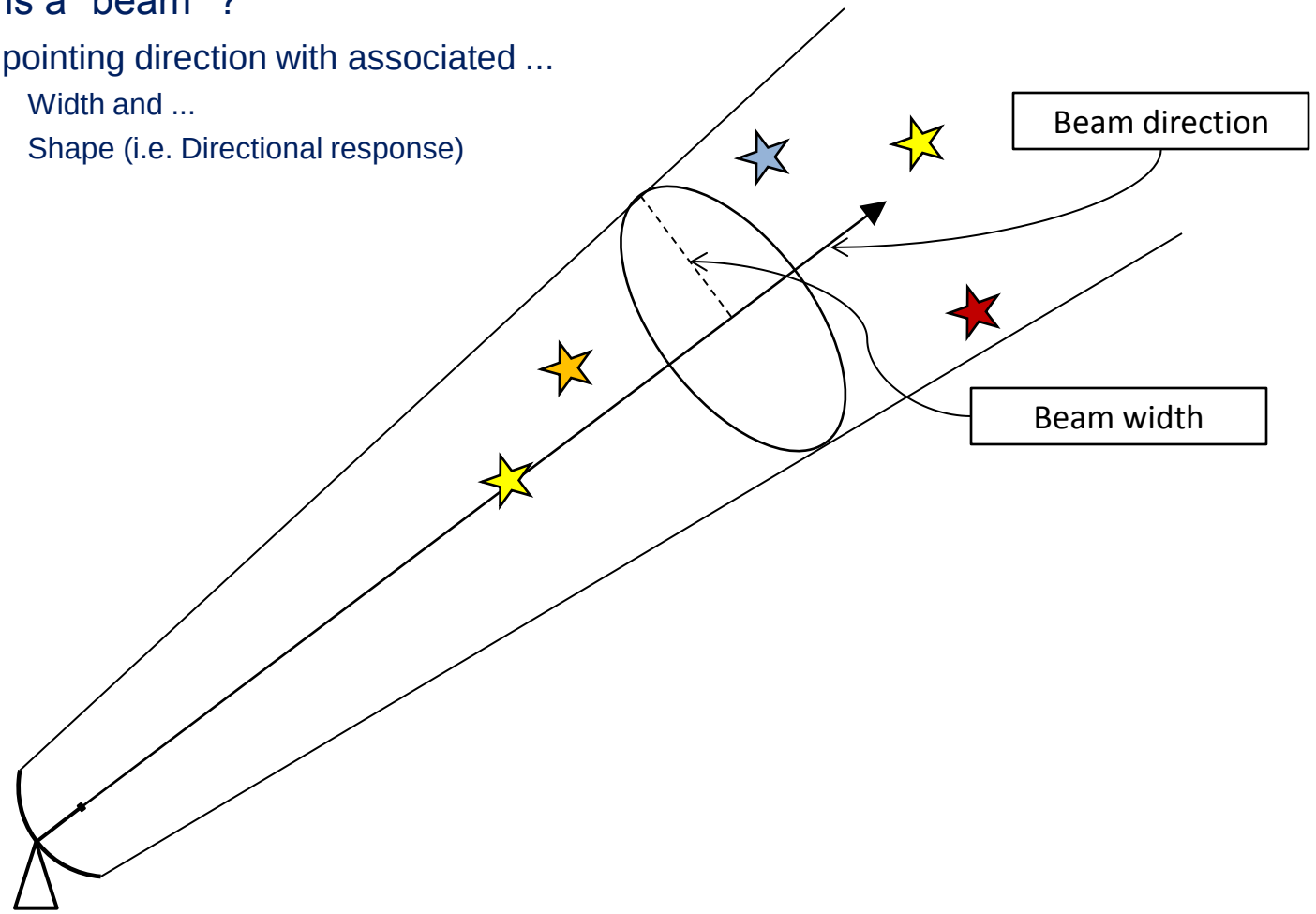
1. Antennas measure broadband signals
2. Split antenna signals into sum of Fourier components
3. Delay == phase factor: $A\, e^{i(\omega t + \varphi)}$
4. Can point in different directions at the same time (multiple beams)



UNIVERSITY OF OXFORD

Beams

- What is a “beam” ?
  - A pointing direction with associated ...
    - Width and ...
    - Shape (i.e. Directional response)



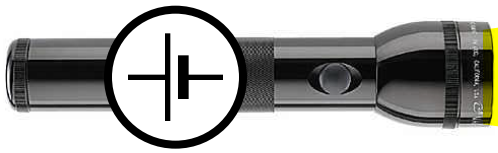
Beams

- What is a “beam” ?
  - A pointing direction with associated ...
    - Width and ...
    - Shape (i.e. Directional response)

An (imperfect) analogy

... To beam

From voltage ...



Beams

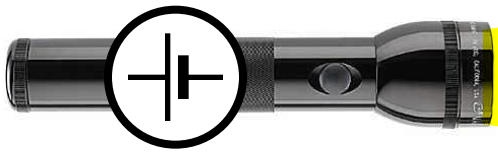
- What is a “beam” ?
  - A pointing direction with associated ...
    - Width and ...
    - Shape (i.e. Directional response)

An (imperfect) analogy

Reverse Time

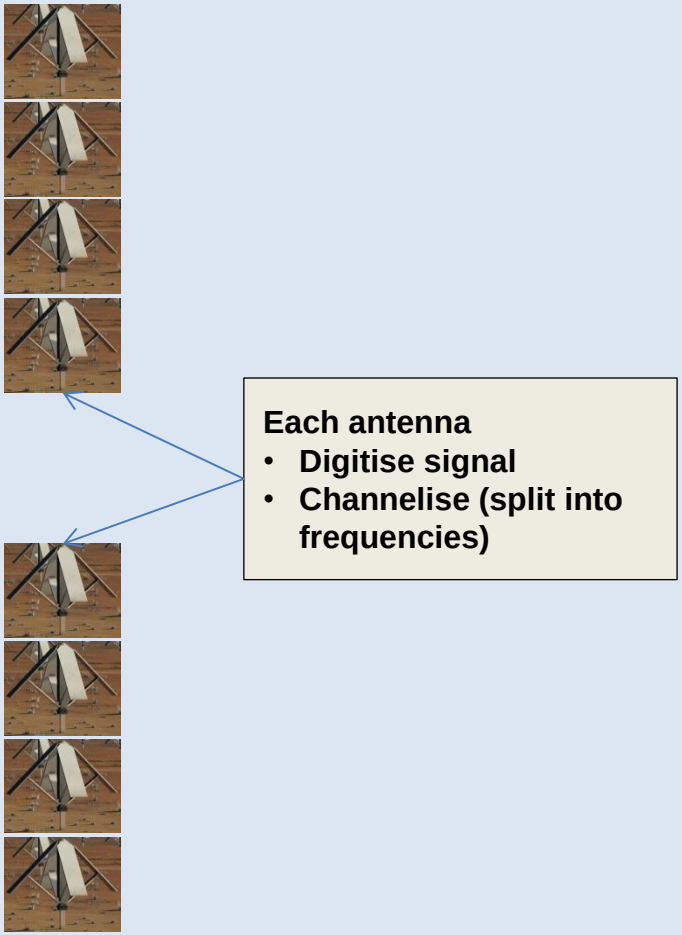
From beam ...

... to voltage



Low Frequency Aperture Array Imaging Pipeline

Low Frequency Aperture Array Imaging Pipeline

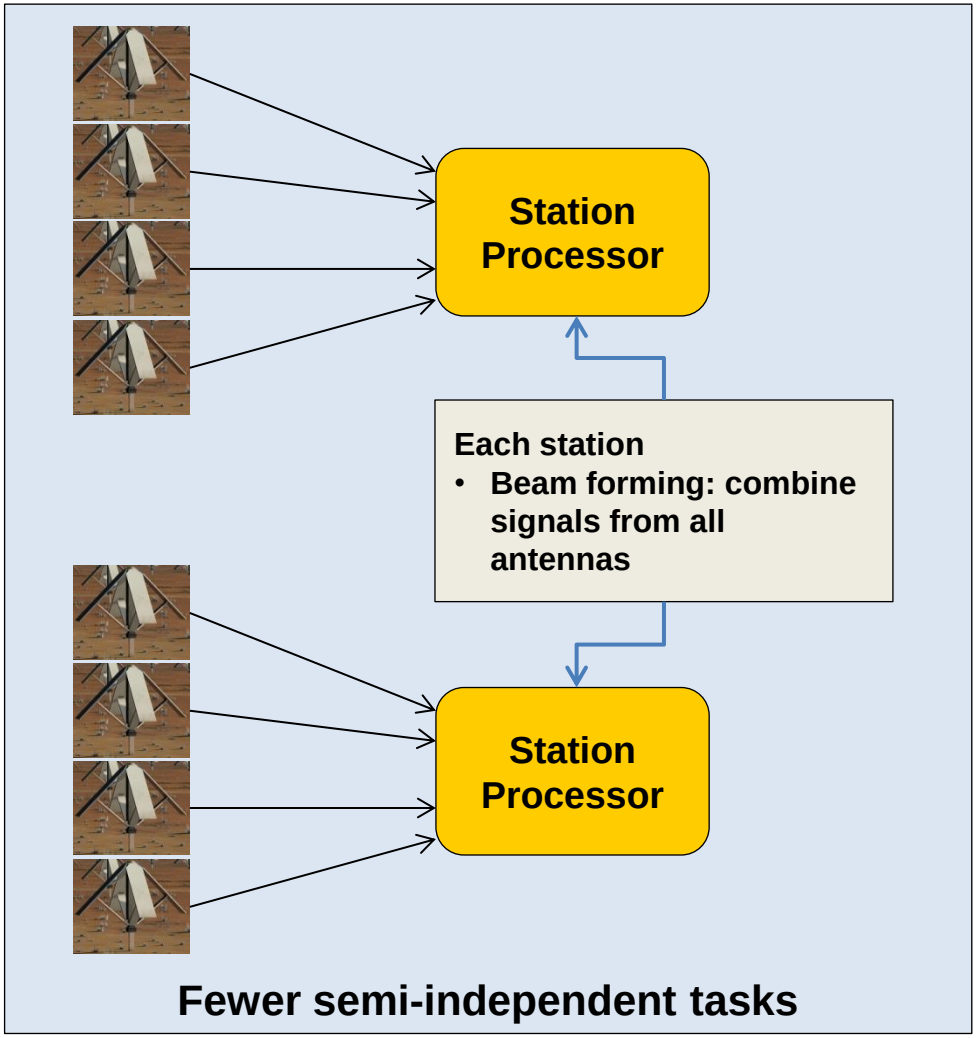


Each antenna
- **Digitise signal**
- **Channelise (split into frequencies)**

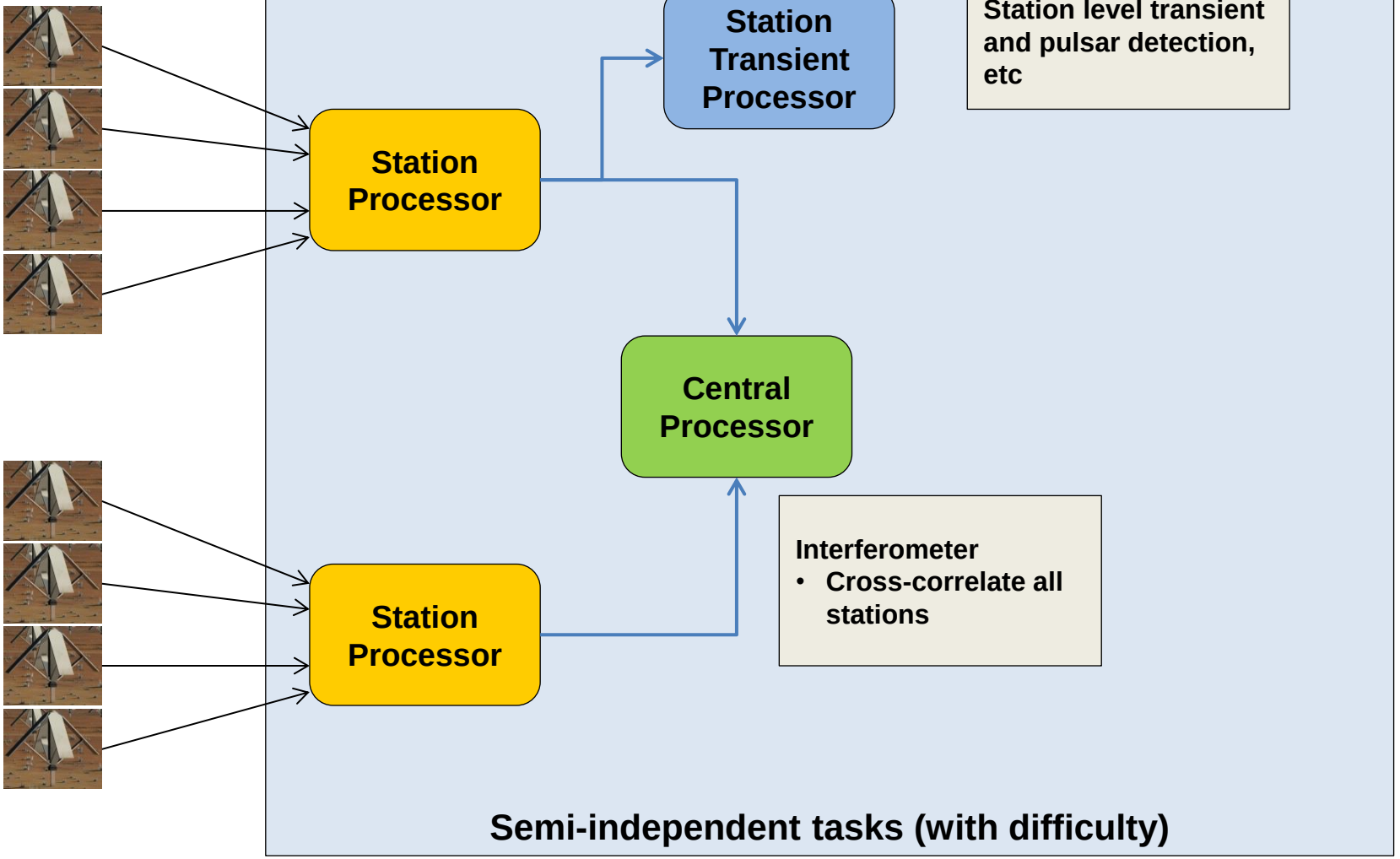
Many independent tasks

UNIVERSITY OF OXFORD

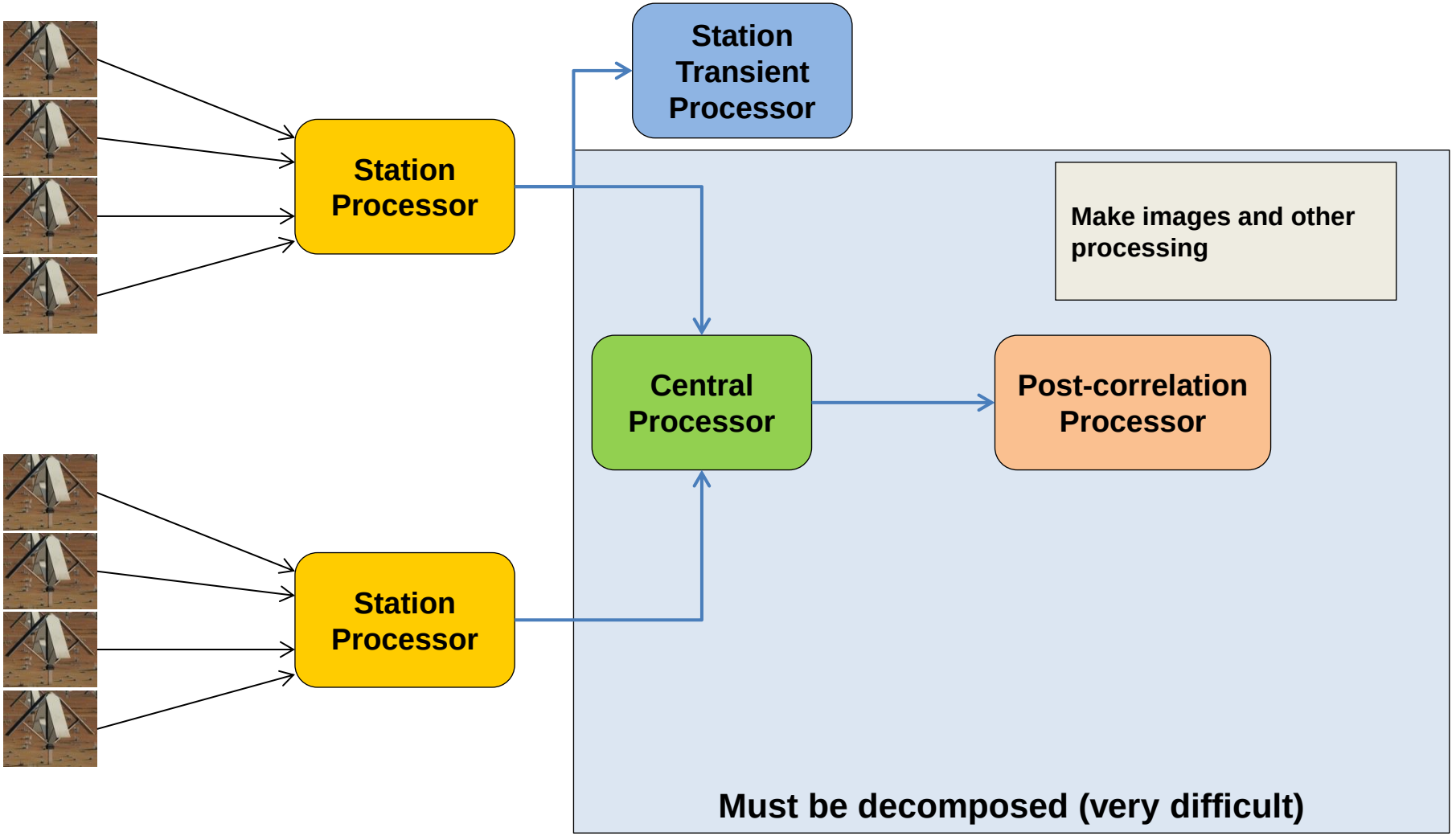
Low Frequency Aperture Array Imaging Pipeline



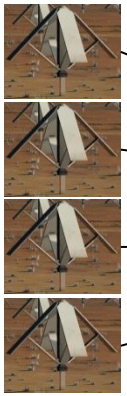
Low Frequency Aperture Array Imaging Pipeline



Low Frequency Aperture Array Imaging Pipeline



Low Frequency Aperture Array Imaging Pipeline



Station Processor

Station Transient Processor

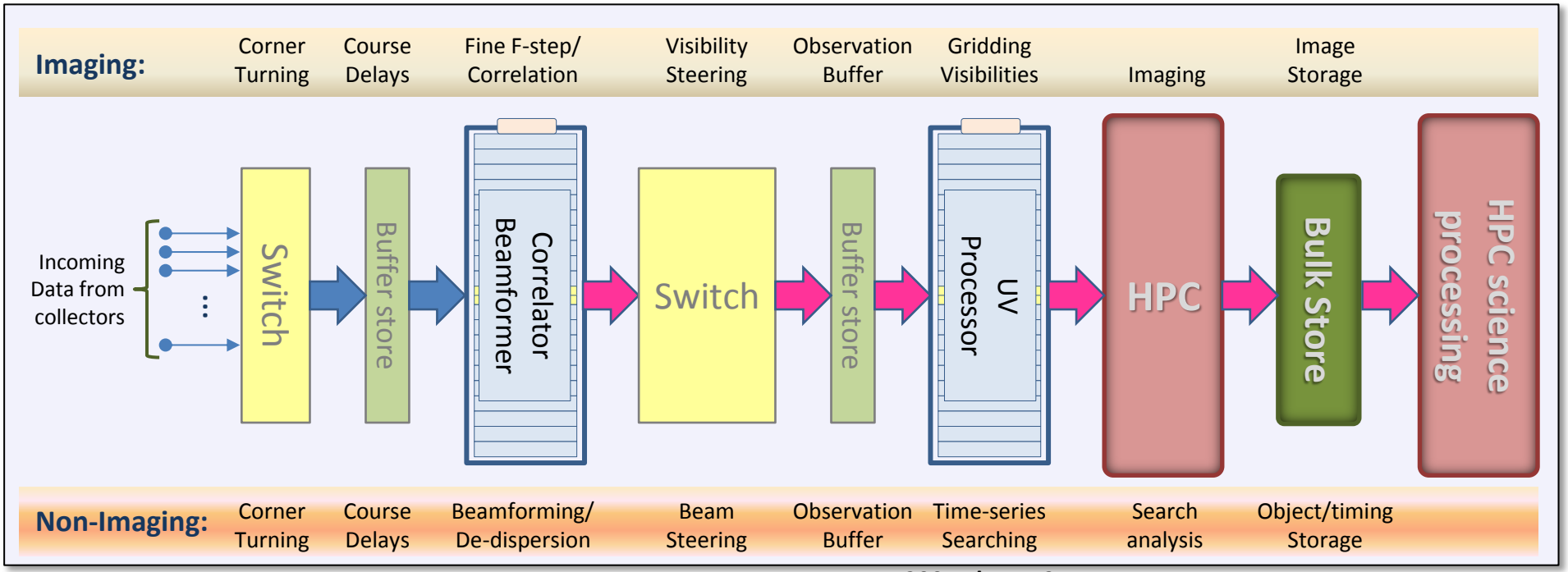
Central Processor

Post-correlation Processor

Station Processor

The outside World
- Other processing
- Repositories
- Mirrors
- etc

Very Simplified Data Flow



Software complexity

Thank you!

Any questions?

Thanks to Paul Alexander (Cambridge), Mike Giles (Oxford) for several slides