Big Data & Big Compute in Radio Astronomy

Rob van Nieuwpoort

netherlands eScience center by SURF & NWO

Universiteit van Amsterdam

ASTRON

VU University Amsterdam Faculty of Sciences

Two simultaneous disruptive technologies

- Radio Telescopes
  - New sensor types
  - Distributed sensor networks
  - Scale increase
  - Software telescopes

- Computer architecture
  - Hitting the memory wall
  - Accelerators

netherlands eScience center

Two simultaneous disruptive technologies

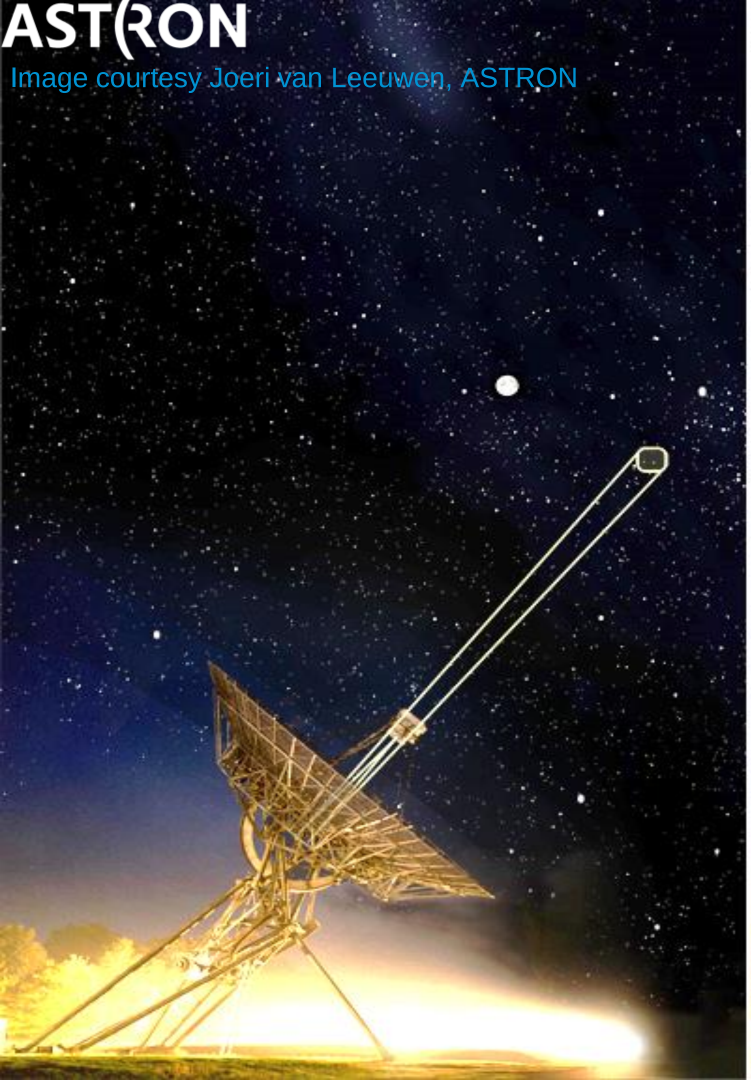
- Radio telescopes
  - New sensor types
  - Distributed sensor networks
  - Scale increase
  - Software telescopes
- Computer architecture
  - Hitting the memory wall
  - Accelerators

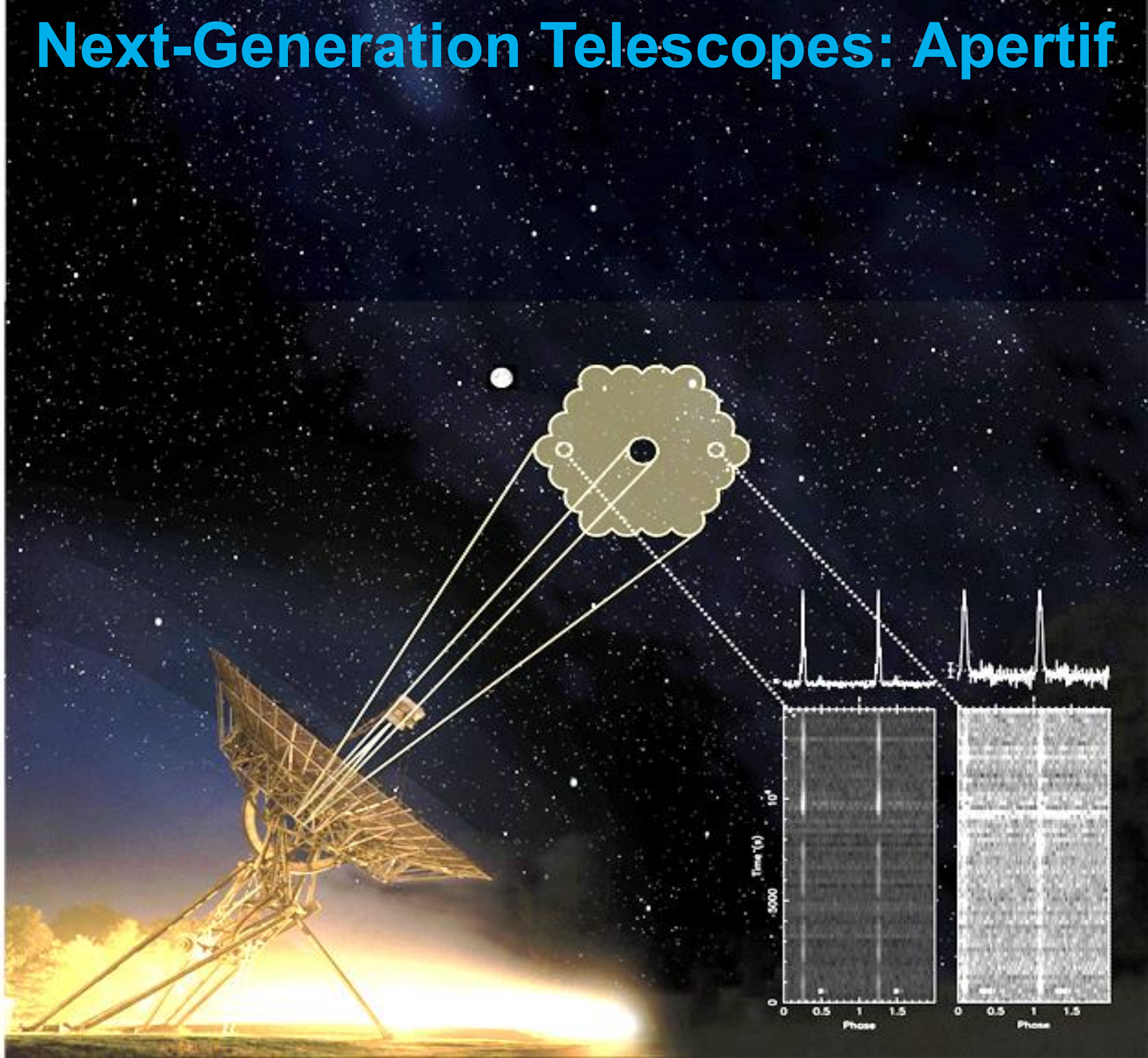
panic!

Next-Generation Telescopes: Apertif

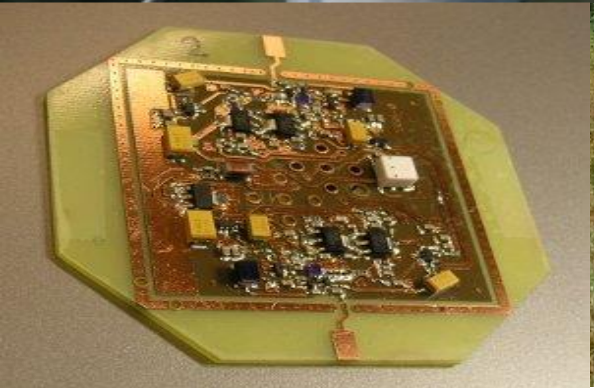
ASTRON

Image courtesy Joeri van Leeuwen, ASTRON

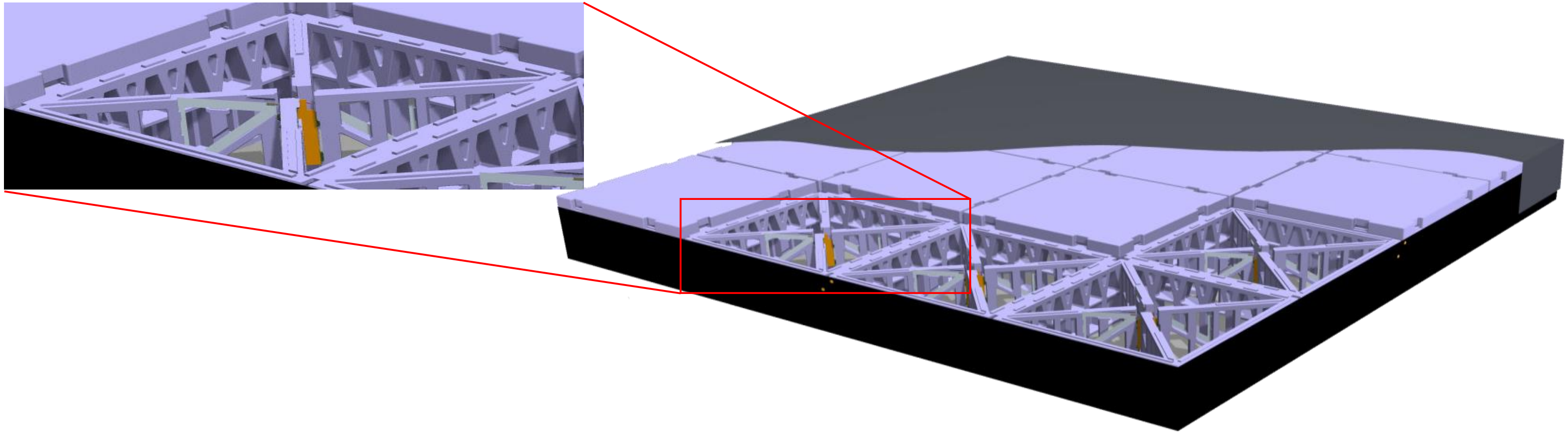




LOFAR low-band antennas



LOFAR high-band antennas



Station (150m)


2x3 km

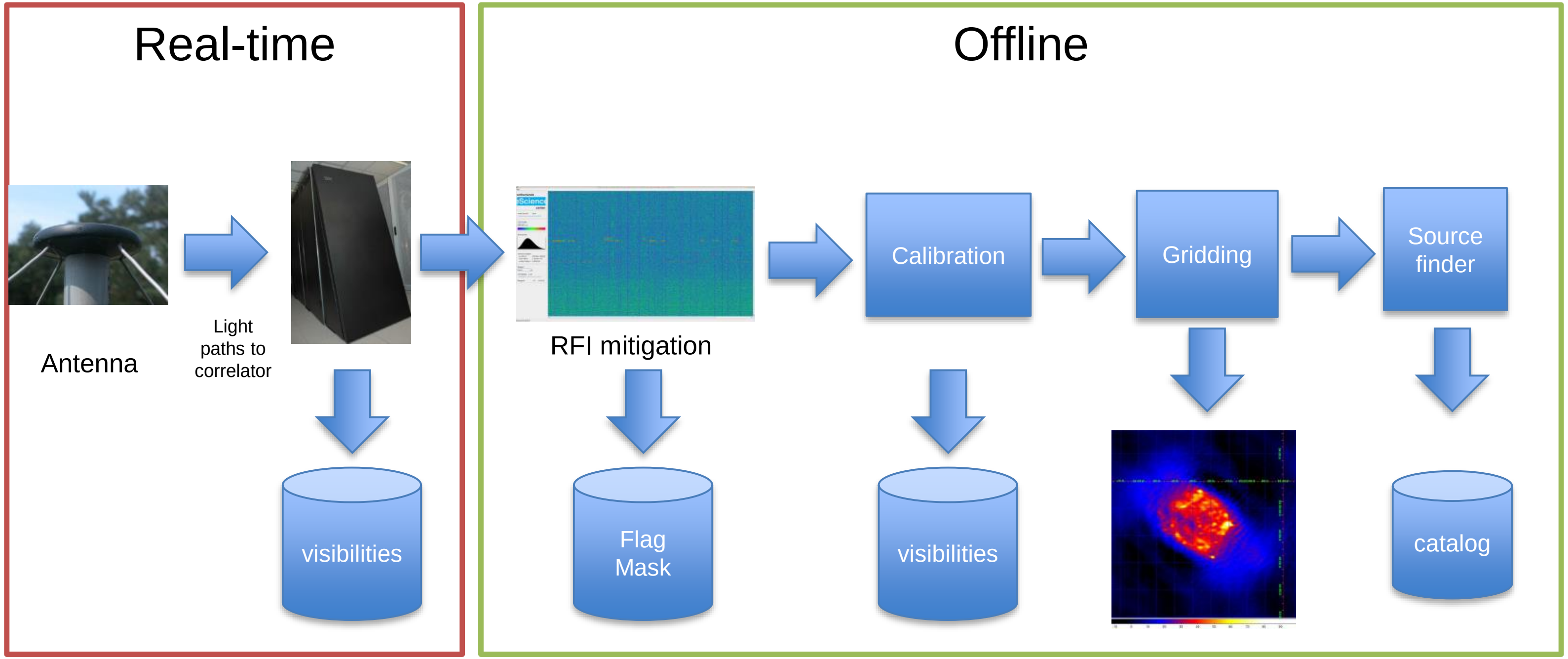
netherlands eScience center

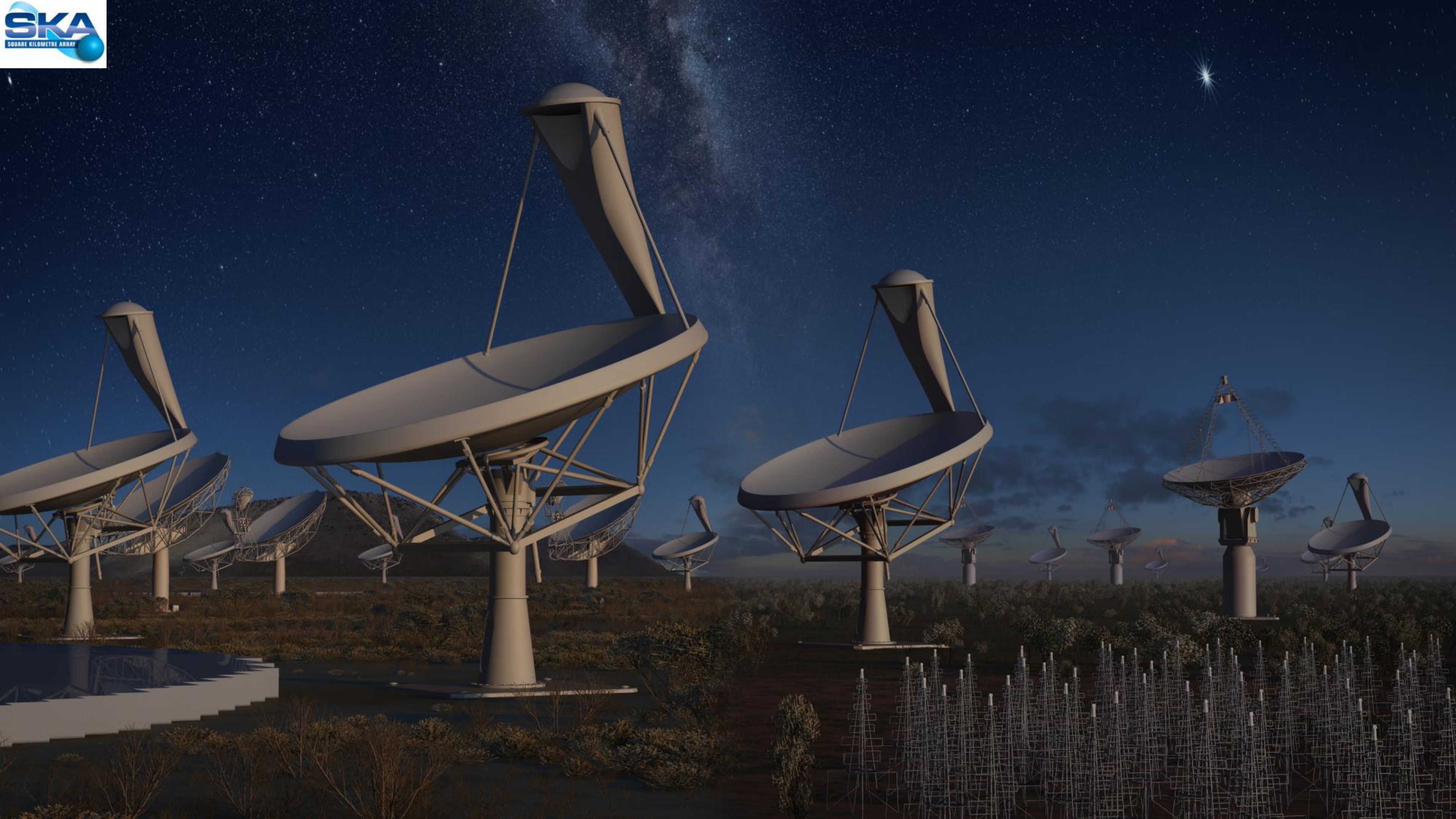
LOFAR

- Largest radio telescope in the world
- ~100.000 omni-directional antennas
- 10 terabit/s, 200 gigabit/s to supercomputer (AMS-IX = 2-3 terabit/s)
- Hundreds of teraFLOPS
- 10–250 MHz
- 100x more sensitive

[John Romein et al, PPoPP, 2014]

Imaging pipeline (LOFAR)





SKA1 MID - the SKA's mid-frequency instrument

The Square Kilometre Array (SKA) will be the world's largest radio telescope, revolutionising our understanding of the Universe. The SKA will be built in two phases - SKA1 and SKA2 - starting in 2018, with SKA1 representing a fraction of the full SKA. SKA1 will include two instruments - SKA1 MID and SKA1 LOW - observing the Universe at different frequencies.

SKA SQUARE KILOMETRE ARRAY

Location: South Africa

Frequency range: **350 MHz** to **14 GHz**

~200 dishes (including 64 MeerKAT dishes)

Total collecting area: **33,000m²** or **126 tennis courts**

Maximum distance between dishes: **150km**

Total raw data output:

2 terabytes per second

62 exabytes per year

SKA1 MID

x340,000

Enough to fill **340,000** average laptops with content **every day**

Compared to the JVLA, the current best similar instrument in the world:

4x the resolution

5x more sensitive

60x the survey speed

www.skatelescope.org | Square Kilometre Array | @SKA_telescope | YouTube The Square Kilometre Array

SKA1 LOW - the SKA's low-frequency instrument

The Square Kilometre Array (SKA) will be the world's largest radio telescope, revolutionising our understanding of the Universe. The SKA will be built in two phases - SKA1 and SKA2 - starting in 2018, with SKA1 representing a fraction of the full SKA. SKA1 will include two instruments - SKA1 MID and SKA1 LOW - observing the Universe at different frequencies.

SKA SQUARE KILOMETRE ARRAY

Location: Australia

Frequency range: **50 MHz** to **350 MHz**

~130,000 antennas spread between **500 stations**

Total collecting area: **0.4km²**

Maximum distance between stations: **65km**

Total raw data output:

157 terabytes per second

4.9 zettabytes per year

SKA1 LOW

Enough to fill up **35,000 DVDs** every second

5x the estimated global internet traffic in 2015 (source: Cisco)

Compared to LOFAR Netherlands, the current best similar instrument in the world

25% better resolution

8x more sensitive

135x the survey speed

www.skatelescope.org | Square Kilometre Array | @SKA_telescope | YouTube The Square Kilometre Array

Did you know?

x 10

+ The dishes of the SKA will produce ten times the global internet traffic.

Did you know?

x 100

+ The aperture arrays in the SKA could produce more than 100 times the global internet traffic.

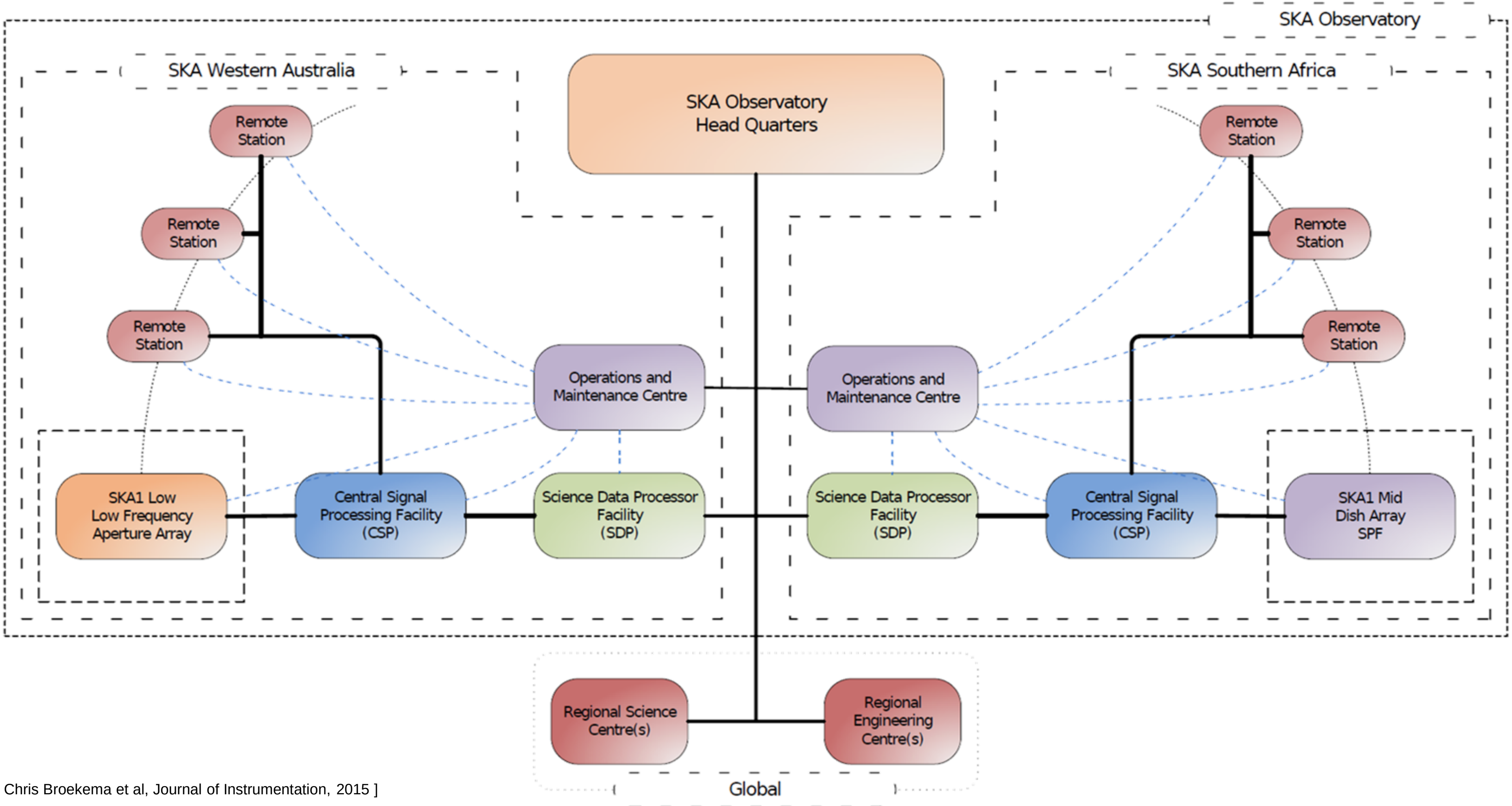
Did you know?

+ The SKA will use enough optical fibre to wrap twice around the Earth!

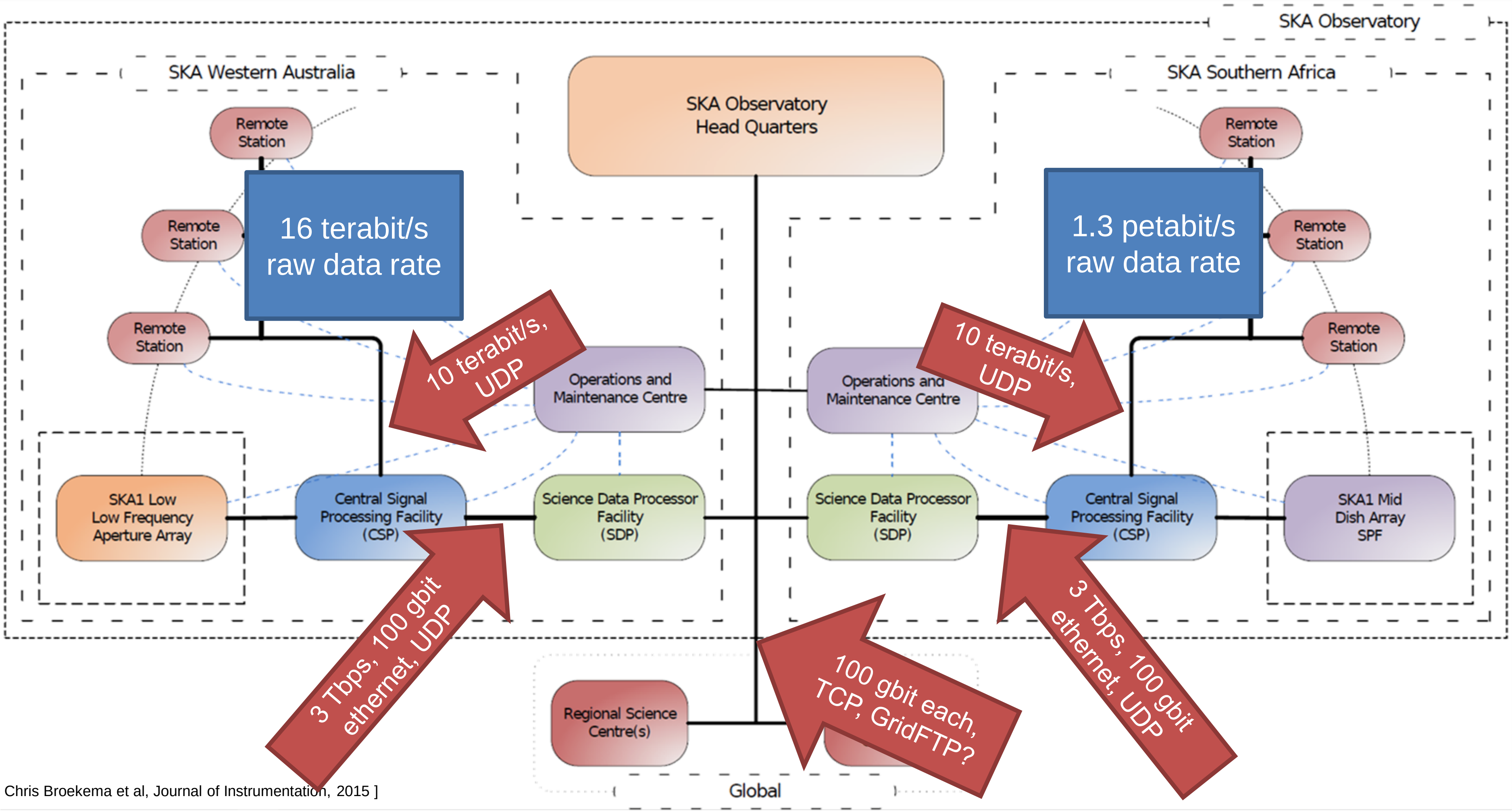
Did you know?

x 3 MILLION

+ The SKA super computer will perform 10^{18} operations per second – equivalent to the number of stars in three million Milky Way galaxies – in order to process all the data that the SKA will produce.

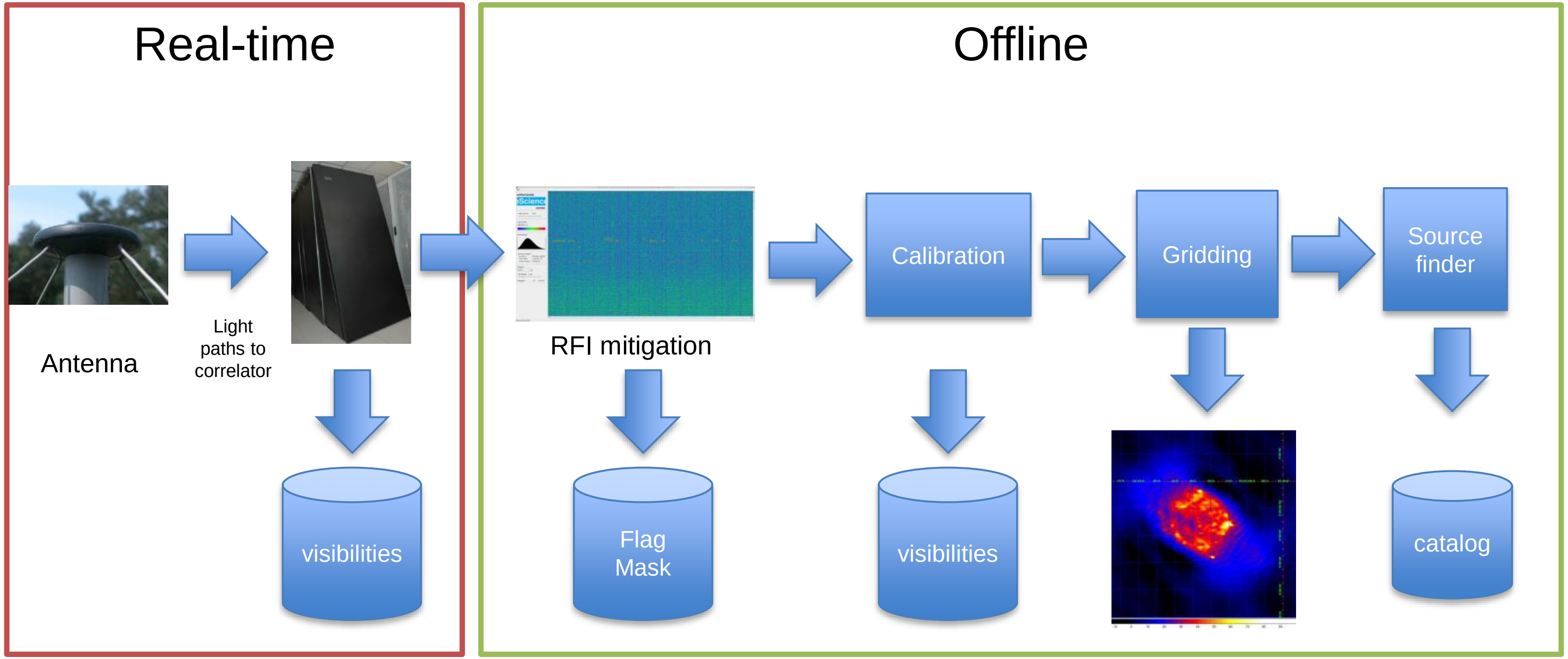


[Chris Broekema et al, Journal of Instrumentation, 2015]

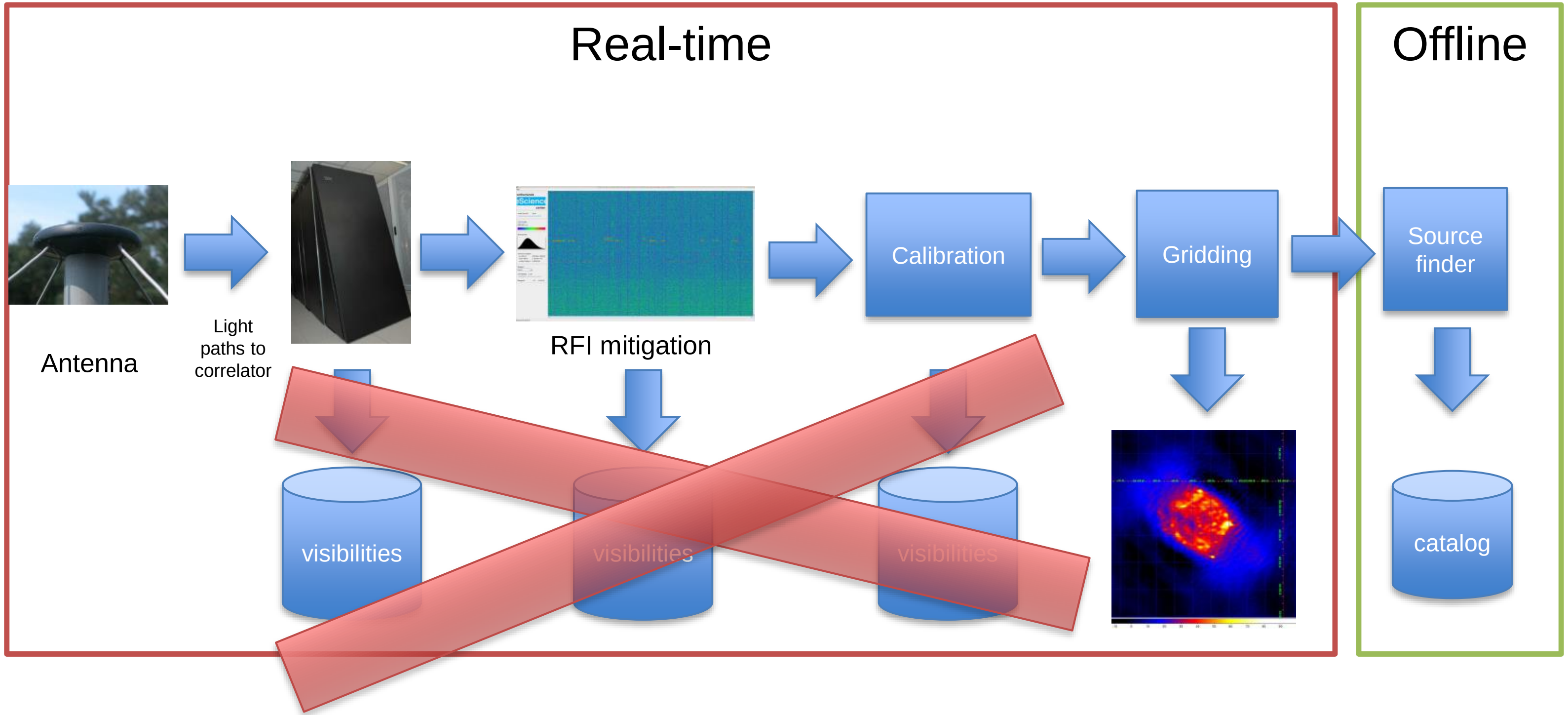


[Chris Broekema et al, Journal of Instrumentation, 2015]

Imaging pipeline (LOFAR)



Imaging pipeline: scaling up to SKA



Meanwhile, in computer science…

Disruptive changes in architectures

Potential of accelerators

- Example: NVIDIA K80 GPU (2014)
- Compared to modern CPU (Intel Haswell, 2014)
  - 28 times faster at 8 times less power per operation
  - 3.5 times less memory bandwidth per operation
  - 105 times less bandwidth per operation including PCI-e
- Compared to BG/p supercomputer
  - 642 times faster at 51 times less power per operation
  - 18 times less memory bandwidth per operation
  - 546 times less bandwidth per operation including PCI-e
- Legacy codes and algorithms are inefficient
- Need different programming methodology and programming models, algorithms, optimizations
- Can we build large-scale scientific instruments with accelerators?

Our Strategy for flexibility, portability

- Investigate algorithms
- OpenCL: platform portability
- Observation type and parameters only known at run time
  - E.g. # frequency channels, # receivers, longest baseline, filter quality, observation type
- Use runtime compilation and auto-tuning
  - Map *specific problem instance* efficiently to hardware
  - Auto tune platform-specific parameters
- Portability across different instruments, observations, platforms, time!

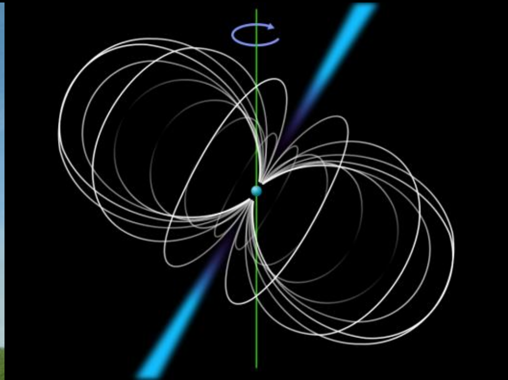
Science Case

Pulsar Searching

Searching for Pulsars

- Rapidly rotating neutron stars
  - Discovered in 1967; ~2500 are known
  - Large mass, precise period, highly magnetized
  - Most neutron stars would be otherwise undetectable with current telescopes

- “Lab in the sky”
  - Conditions far beyond laboratories on Earth
  - Investigate interstellar medium, gravitational waves, general relativity
  - Low-frequency spectra, pulse morphologies, pulse energy distributions
  - Physics of the super-dense superfluid present in the neutron star core

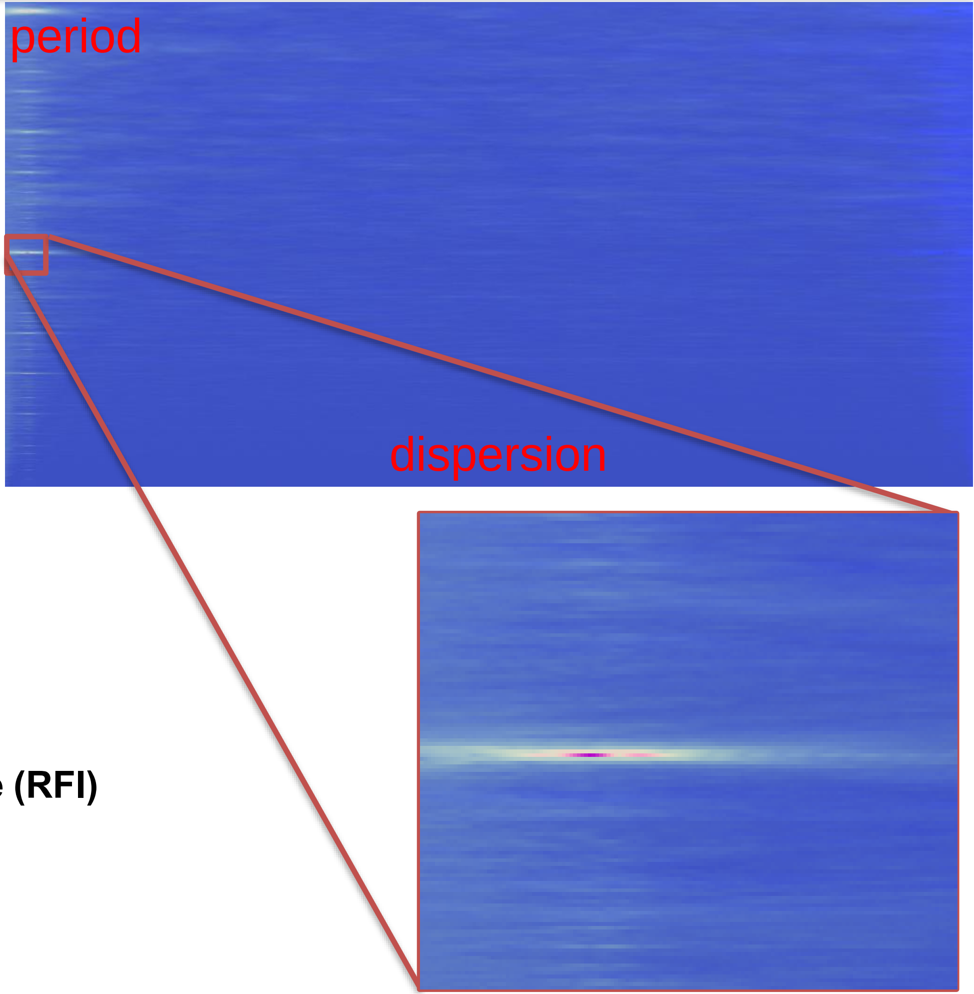


Alessio Sclocco, Rob van Nieuwpoort, Henri Bal,
Joeri van Leeuwen, Jason Hessels, Marco de Vos

[A. Sclocco et al, IEEE eScience, 2015]

Pulsar Searching Pipeline

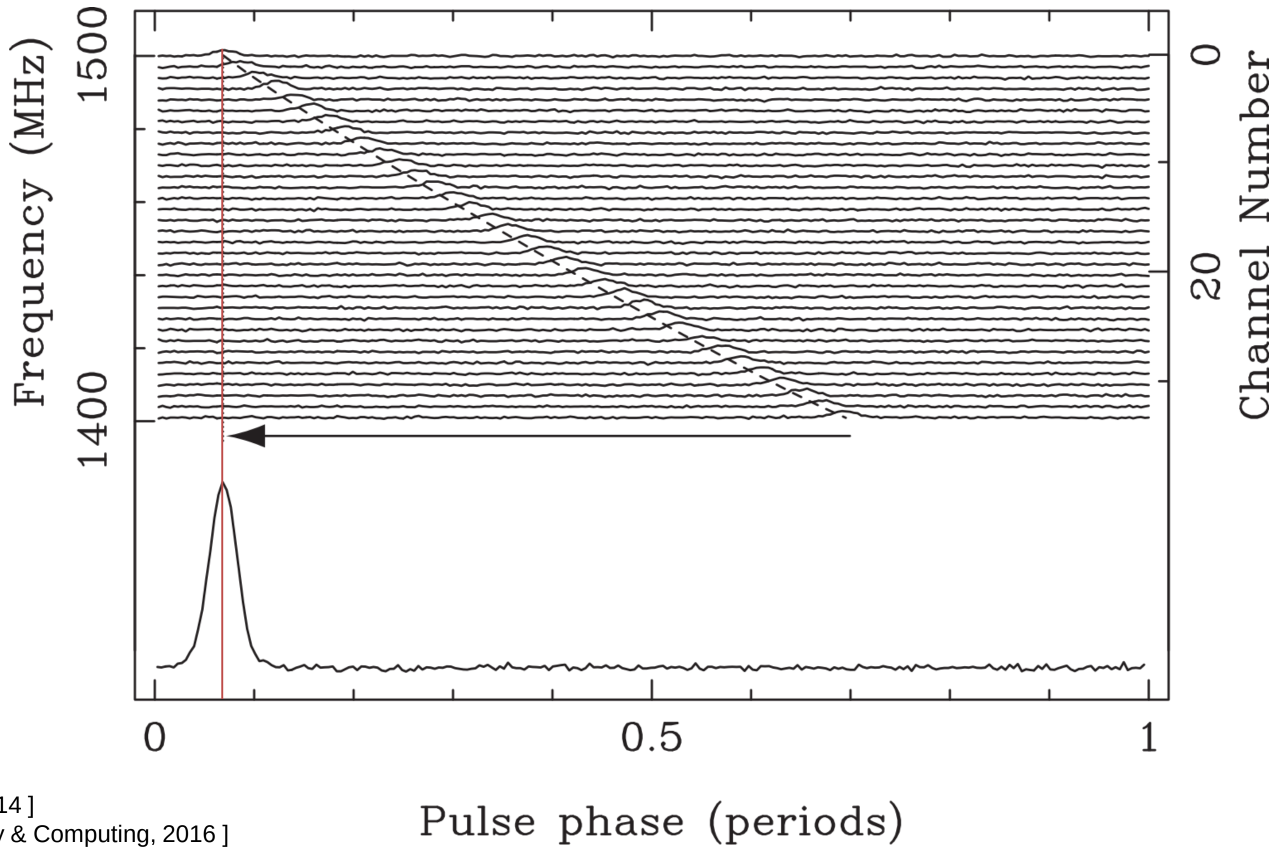
- Three unknowns:
  - Location: create many beams on the sky
    [Alessio Sclocco et al, IPDPS, 2012]
  - Dispersion: focusing the camera
    [Alessio Sclocco et al, IPDPS, 2012]
  - Period
- Brute force search across all parameters
- Everything is trivially parallel (or is it?)
- Complication: Radio Frequency Interference (RFI)
  [Rob van Nieuwpoort et al: Exascale Astronomy, 2014]



An example of real time challenges

Auto-tuning: Dedispersion

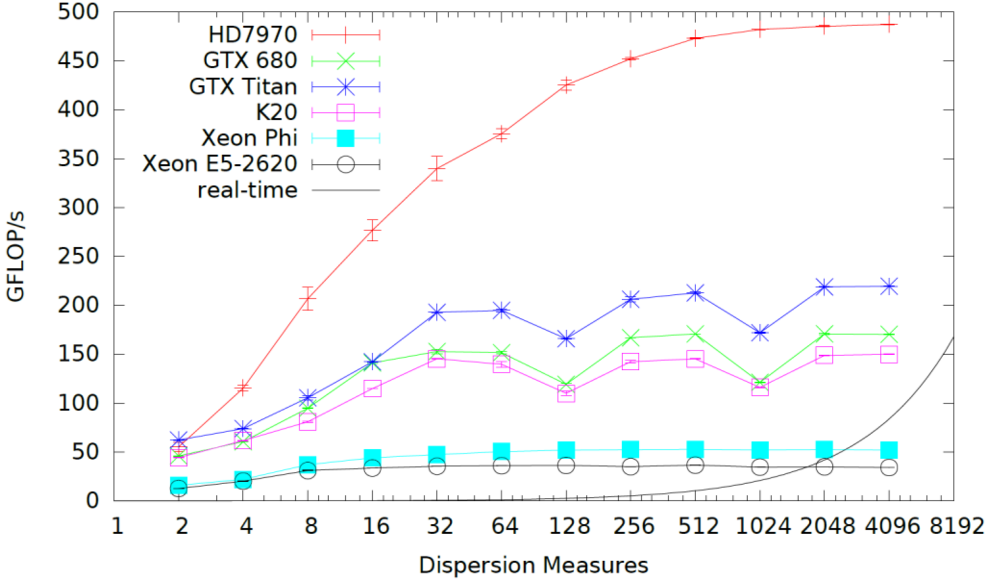
Dedispersion



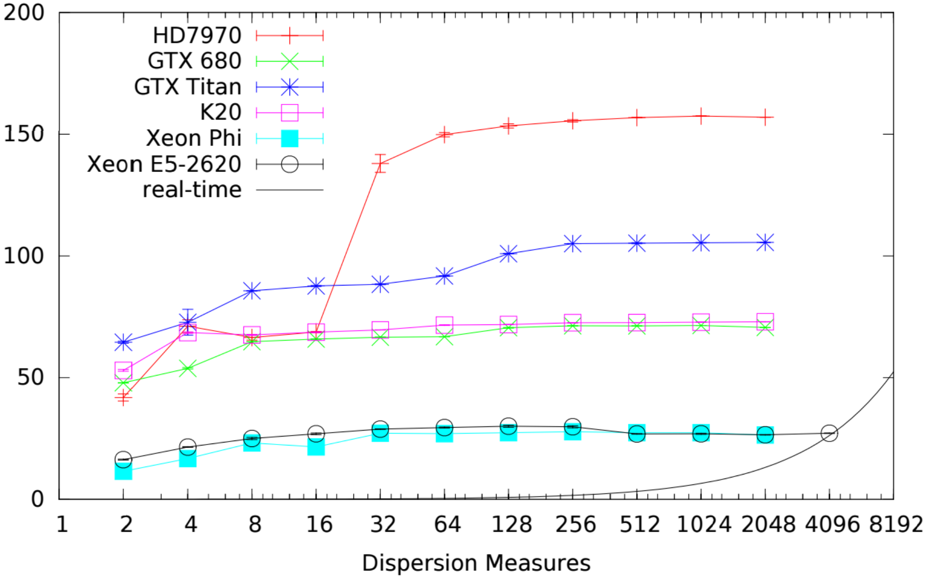
[A. Sclocco et al, IPDPS 2014]
[A. Sclocco et al, Astronomy & Computing, 2016]

Auto-tuned performance

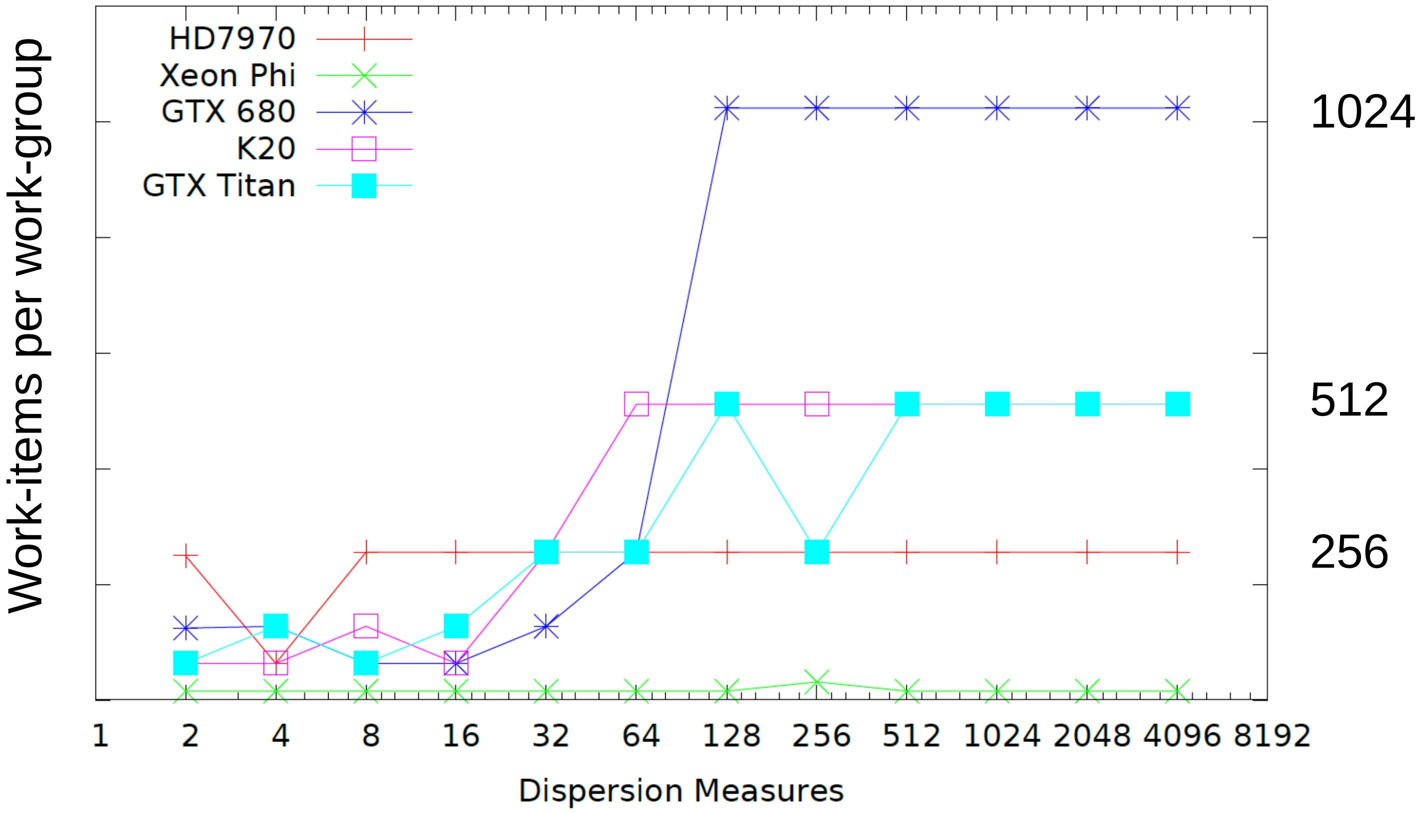
Apertif scenario



LOFAR scenario

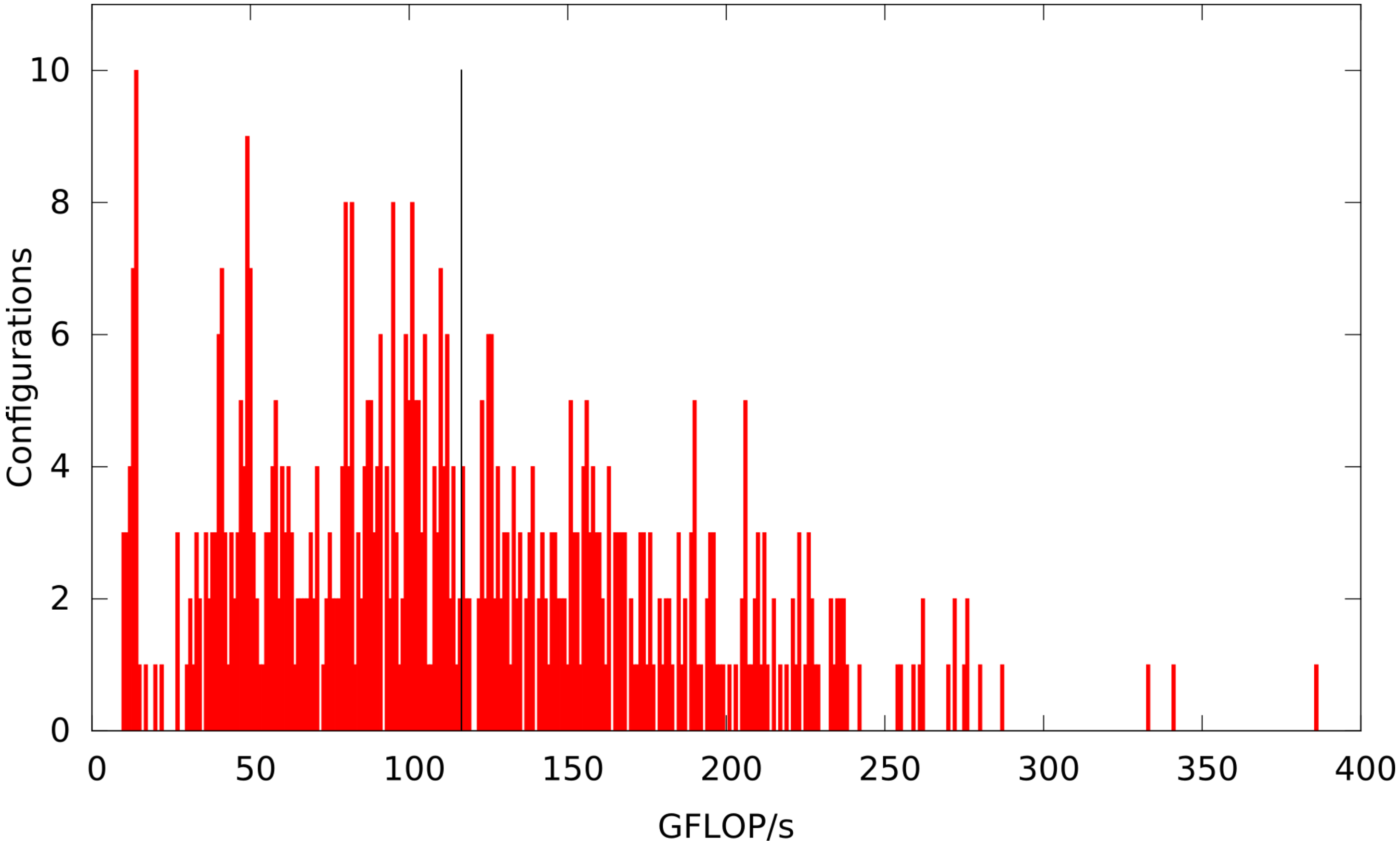


Auto-tuning platform parameters

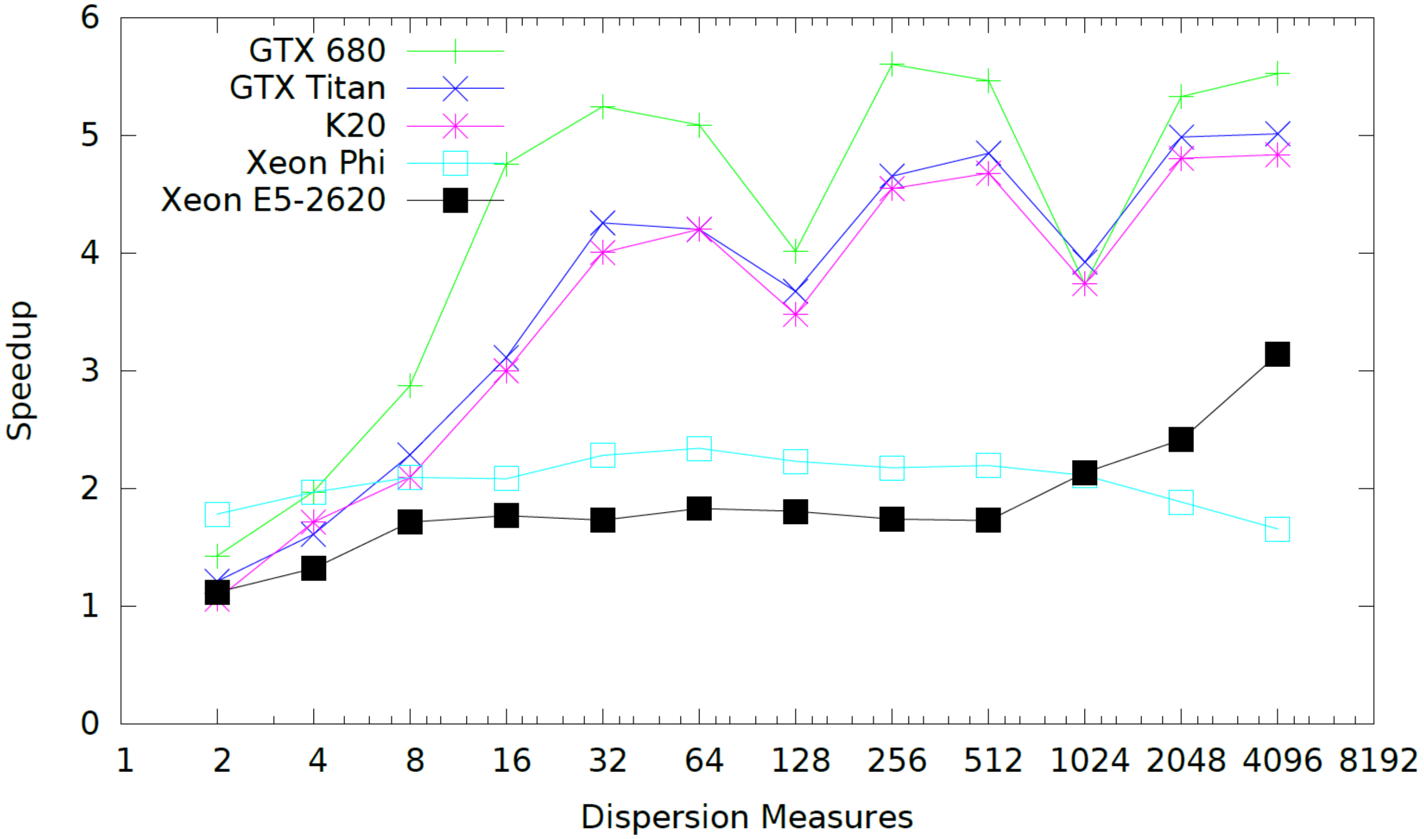


Apertif scenario

Histogram: Auto-Tuning Dedispersion for Apertif



Speedup over best possible fixed configuration



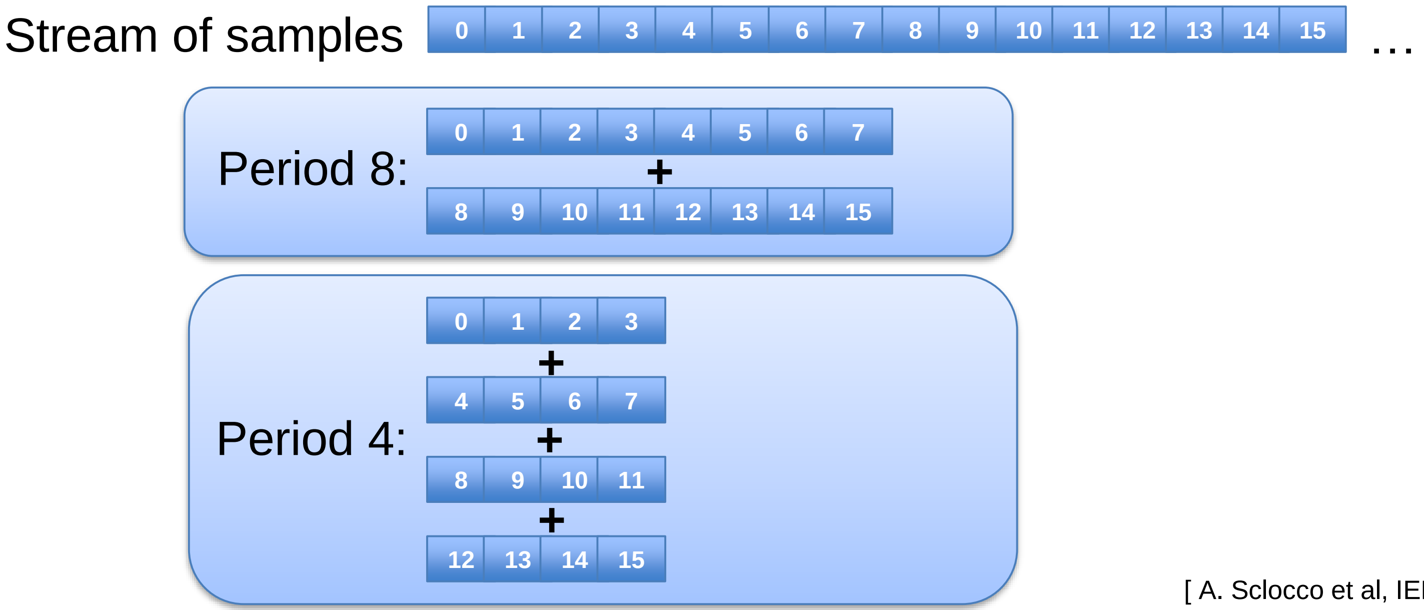
Apertif scenario

An example of real time challenges

Changing algorithms: Period search

Period Search: Folding

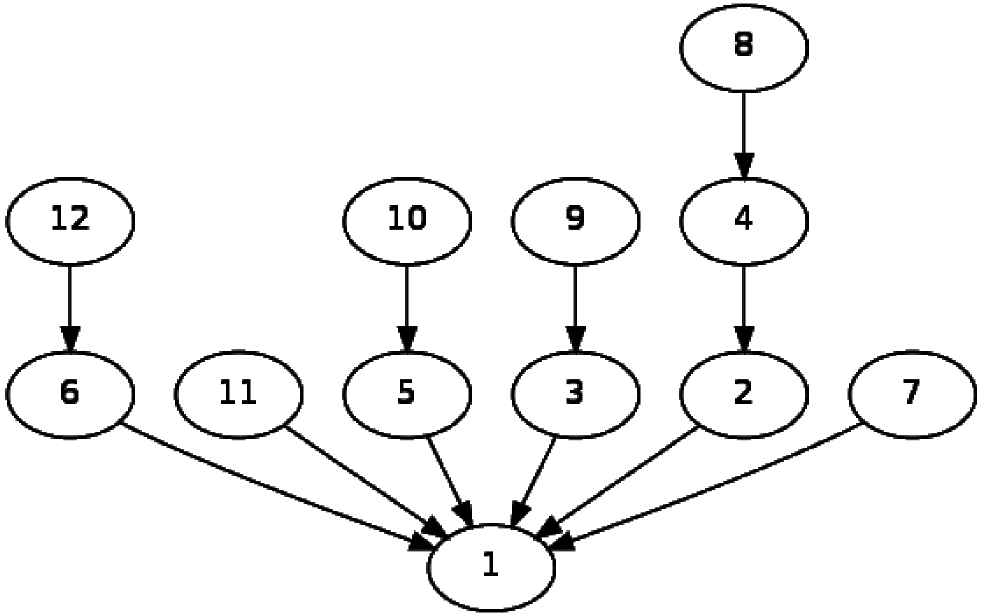
- **Traditional offline approach: FFT**
- **Big Data requires change in algorithm: must be real time & streaming**



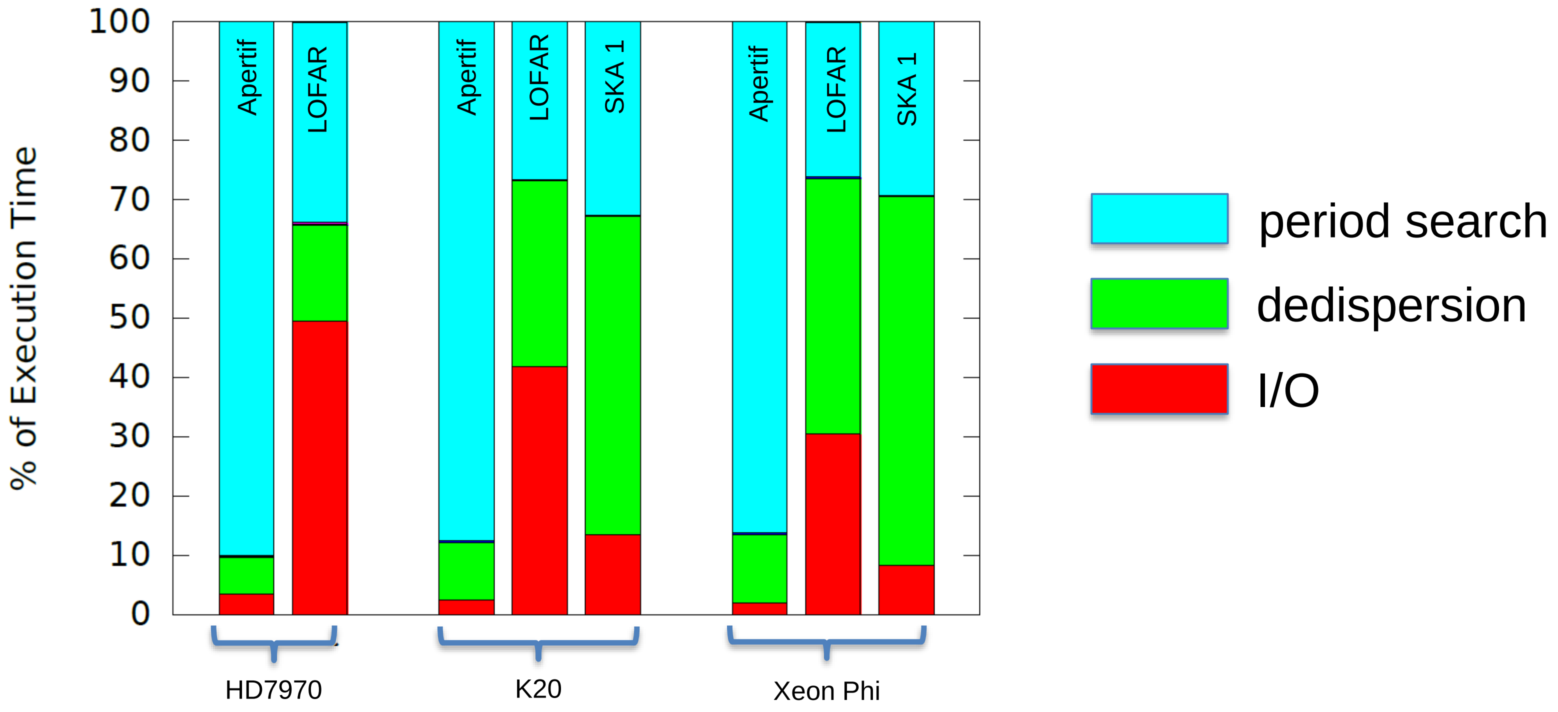
[A. Sclocco et al, IEEE eScience, 2015]

Optimizing Folding

- Build a tree of periods to maximize reuse
- Data reuse: walk the paths from leafs to root

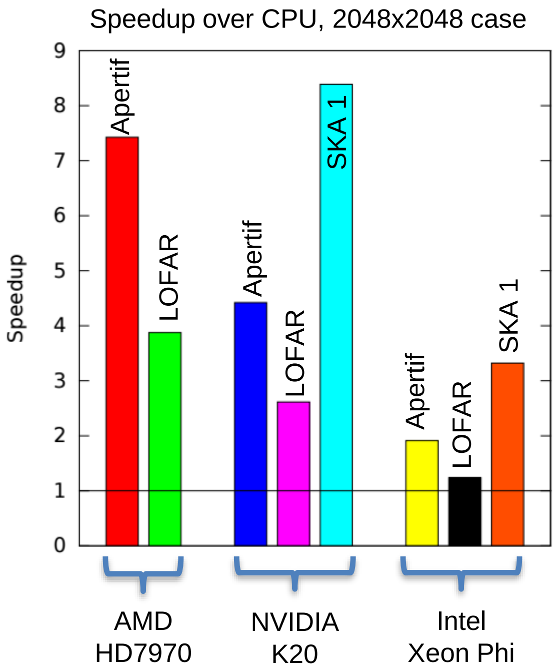


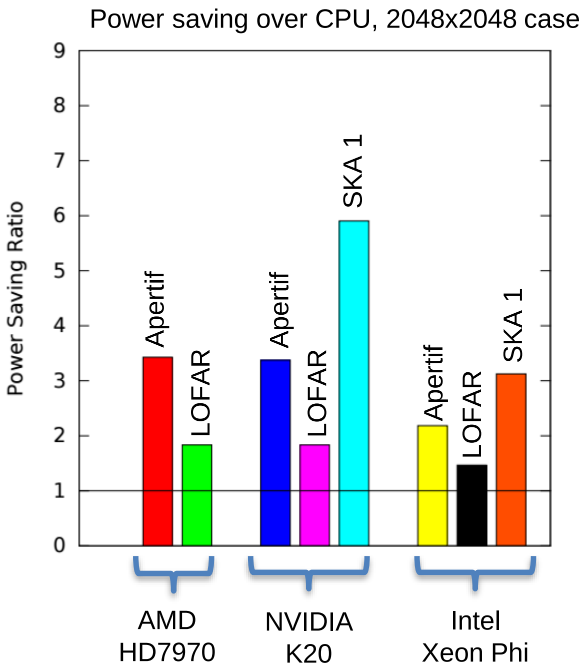
Pulsar pipeline Performance Breakdown



Pulsar pipeline

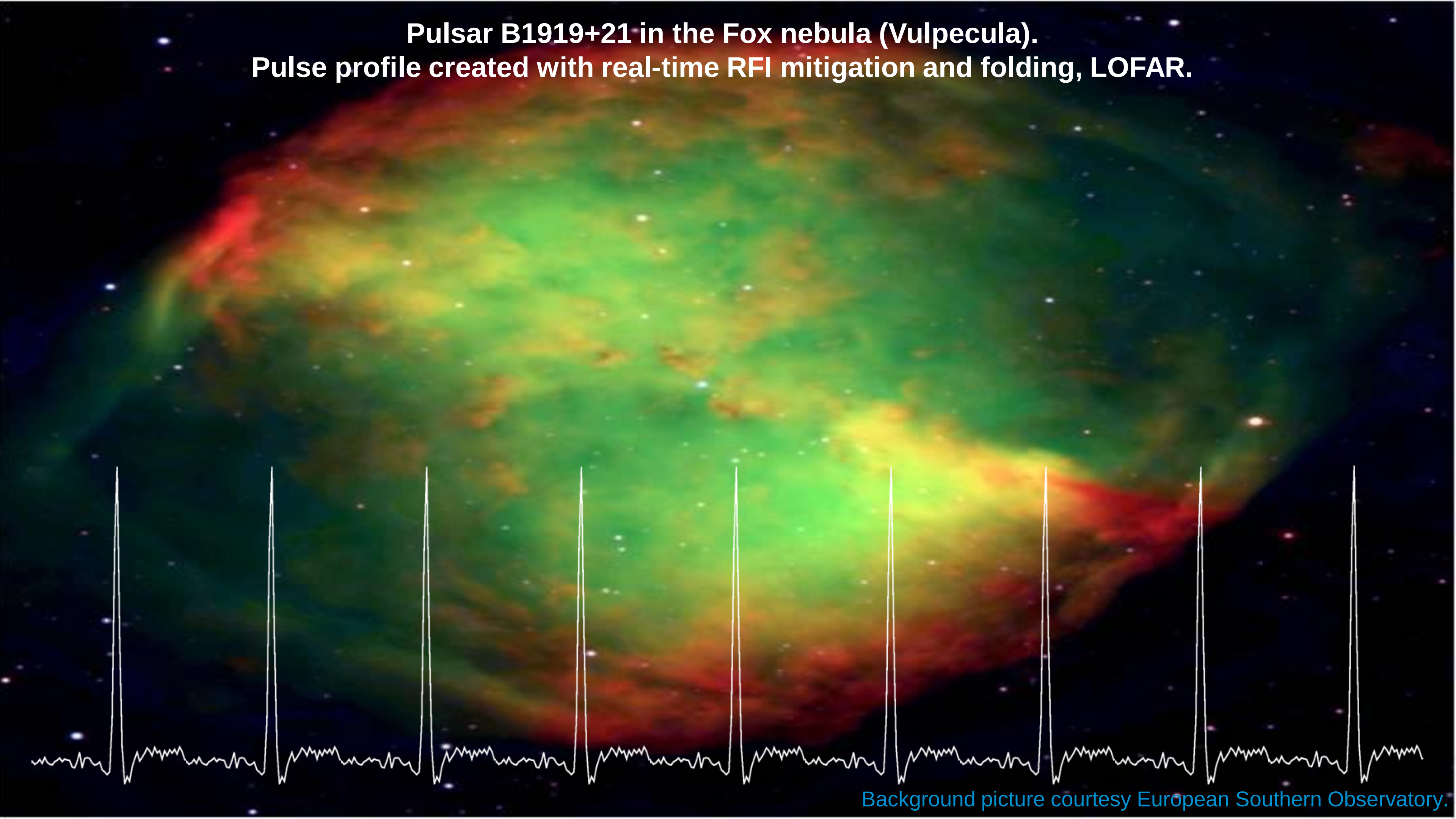
Apertif and LOFAR: real data
SKA1: simulated data





SKA1 baseline design, pulsar survey: 2,222 beams; 16,113 DMs; 2,048 periods.

Total number of GPUs needed: 140,000. This requires 30 MW. SKA2 should be 100x larger, in the 2023-2030 timeframe.


Pulsar B1919+21 in the Fox nebula (Vulpecula).
Pulse profile created with real-time RFI mitigation and folding, LOFAR.
Background picture courtesy European Southern Observatory.

Conclusions: size does matter!

- Big Data changes everything
  - Offline versus streaming, best hardware architecture, algorithms, optimizations
  - Need modular architectures that allow us to easily plug-in accelerators, FPGAs, ASICs, …
  - Auto-tuning and runtime compilation: powerful mechanisms for performance and portability
- eScience approach works!
  - Need domain expert for deep understanding & choice of algorithms
  - Need computer scientists for investigating efficient solutions
  - LOFAR has already discovered more than 25 new pulsars!
- Astronomy is a driving force for HPC, Big Data, eScience
  - Techniques are generic, already applied in image processing, climate, digital forensics