Processing Real-Time LOFAR Telescope Data on a Blue Gene/P

John W. Romein

Stichting ASTRON (Netherlands Institute for Radio Astronomy)
Dwingeloo, the Netherlands

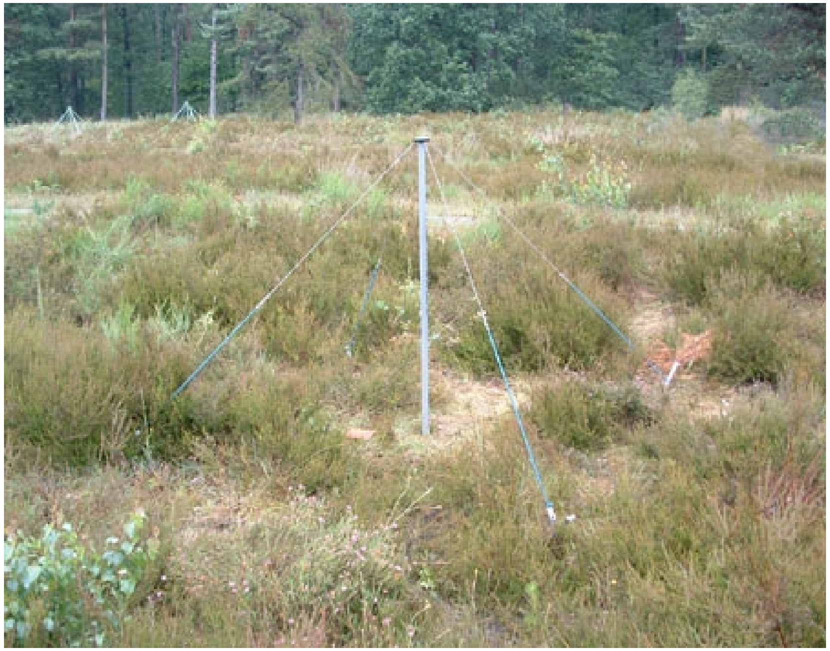
Netherlands Organisation for Scientific Research

LOw Frequency ARray

- radio telescope
- 10–240 MHz
- unexplored
  - dishes infeasible
  - ionospheric disturbance
- new design

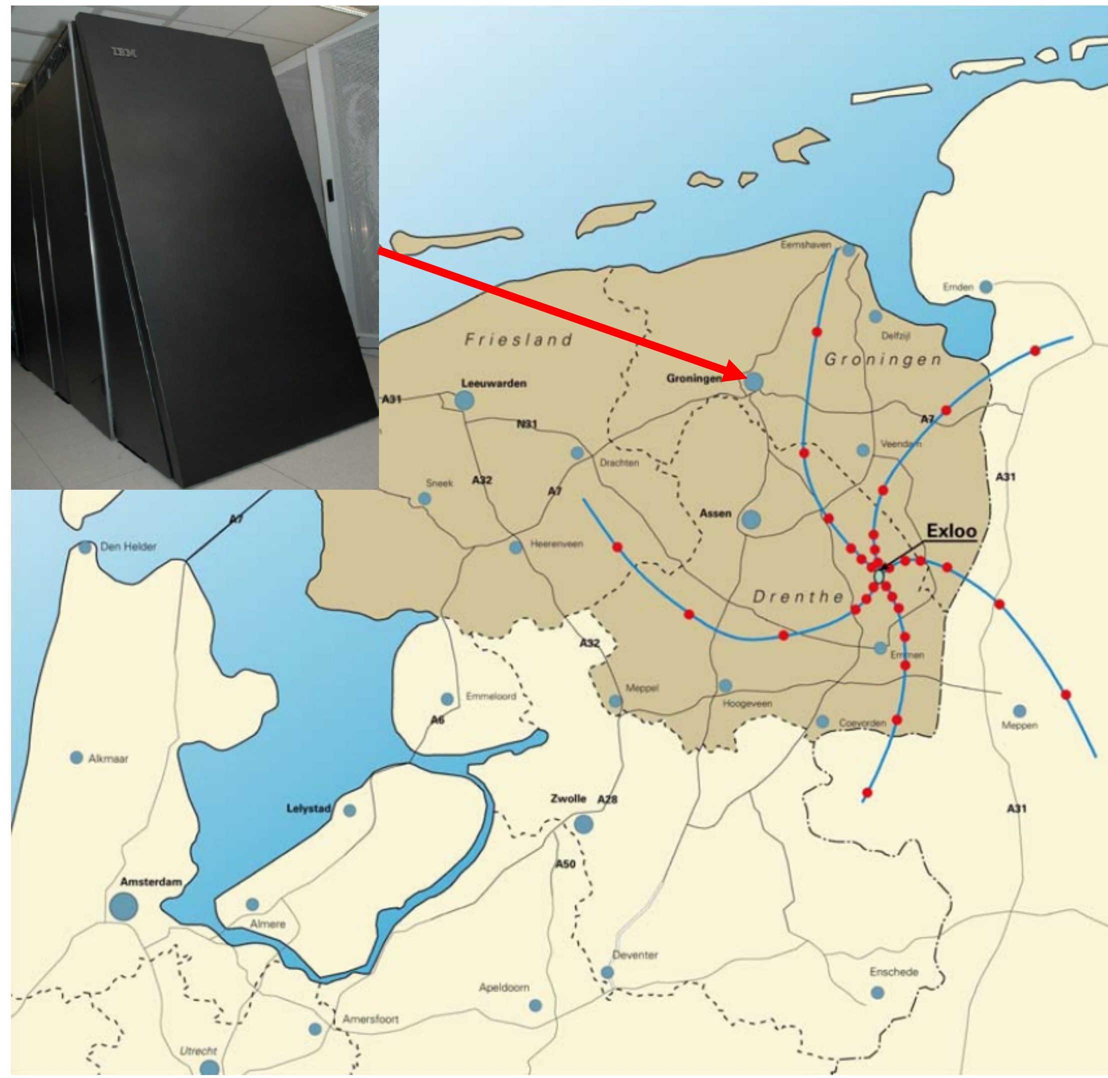
A New Design

- distributed sensor network
- no dishes
  - O(10,000) antennas
  - omni-directional
  - concurrent observations
- software telescope
  - flexible
  - requires supercomputer



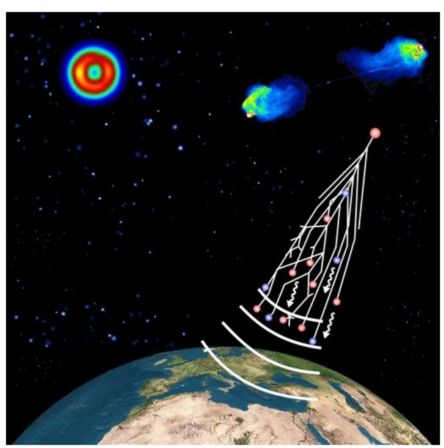
LOFAR Structure

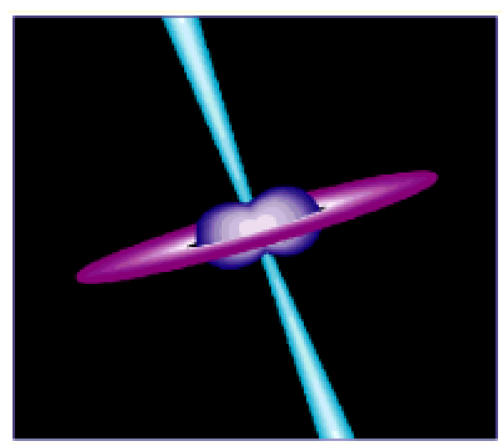
- hierarchical
  - receiver
  - (tile)
  - station
  - telescope
- central core
  - Exloo
- central processing
  - Groningen
  - real time
  - off-line

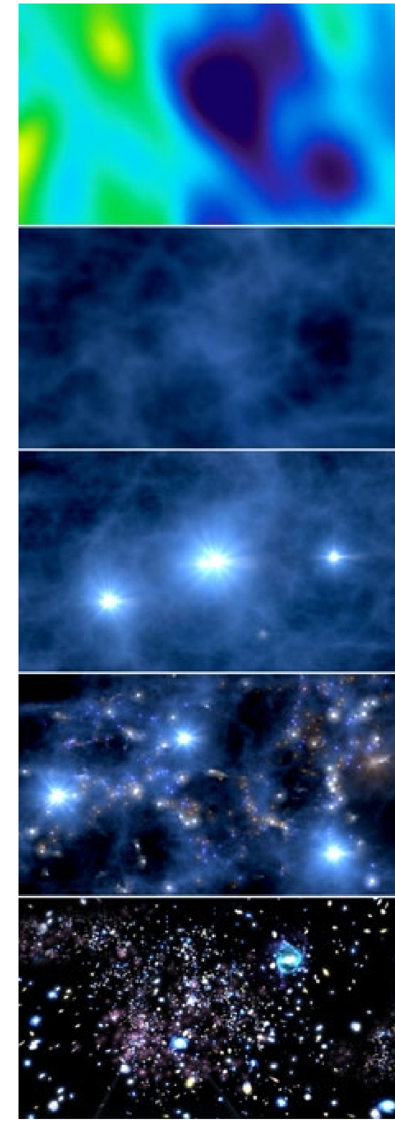


LOFAR Science

- Epoch of Re-ionization
- cosmic rays
- extragalactic surveys
- transients
- pulsars



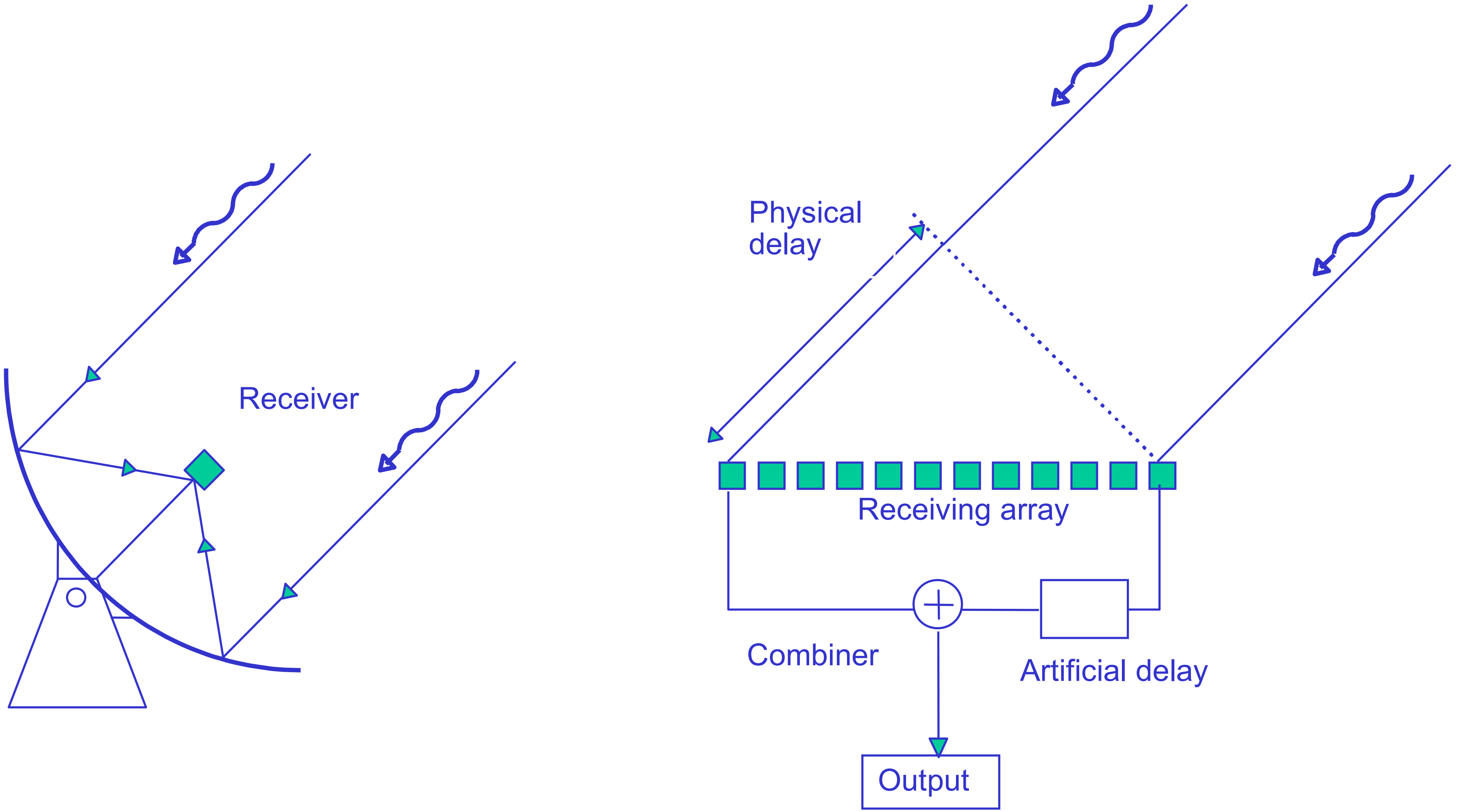




Outline

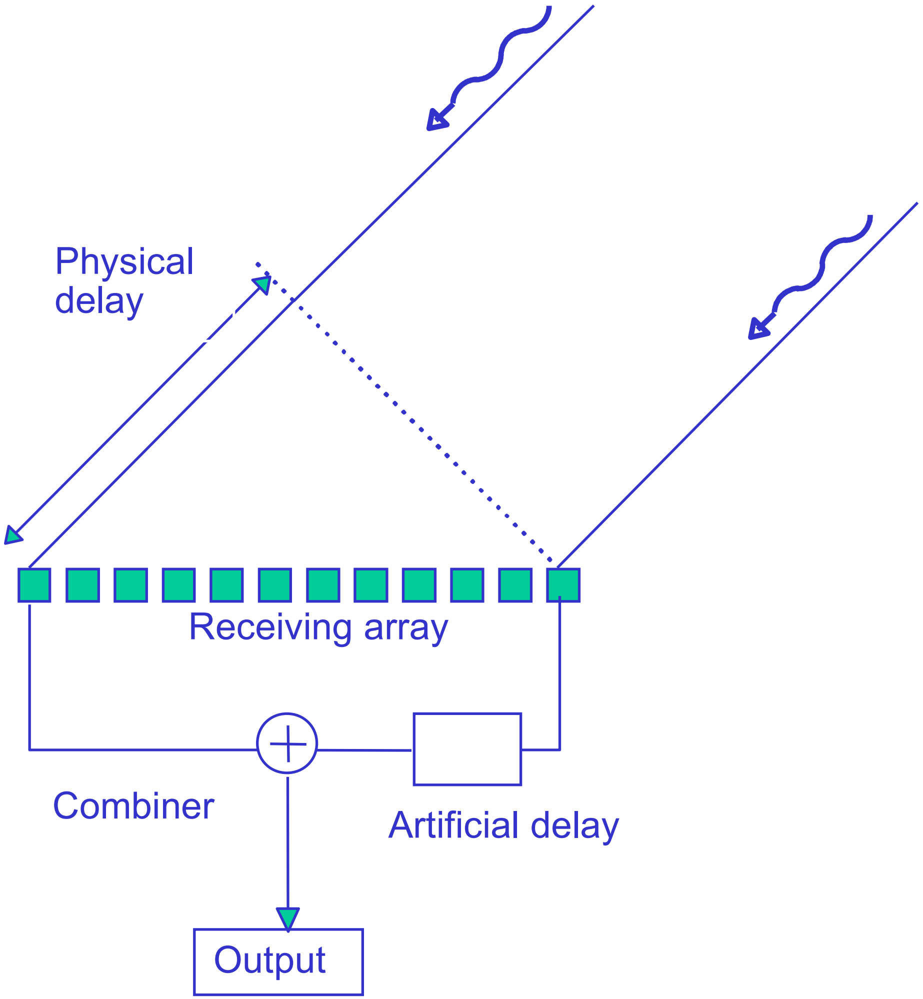
- from wave to image
  - basics
  - receivers
  - stations
  - *real-time Blue Gene/P processing*
    - *performance*
  - off-line processing
  - image

Reflectors vs. Phased Arrays



Beam Forming

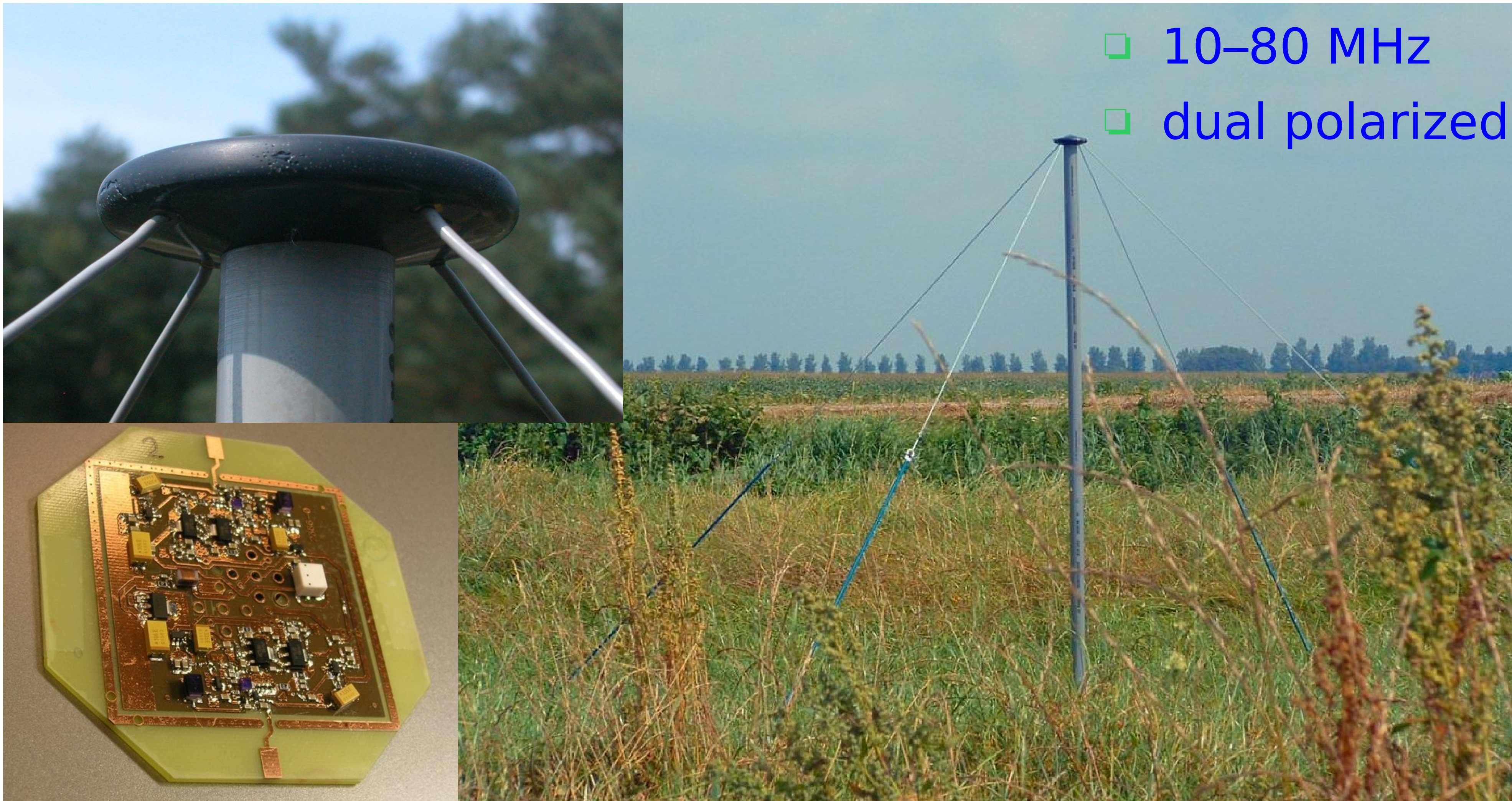
- delay determines observation direction
- beam forming = delayed addition
- diameter determines FoV
- use earth rotation



LOFAR Antennas

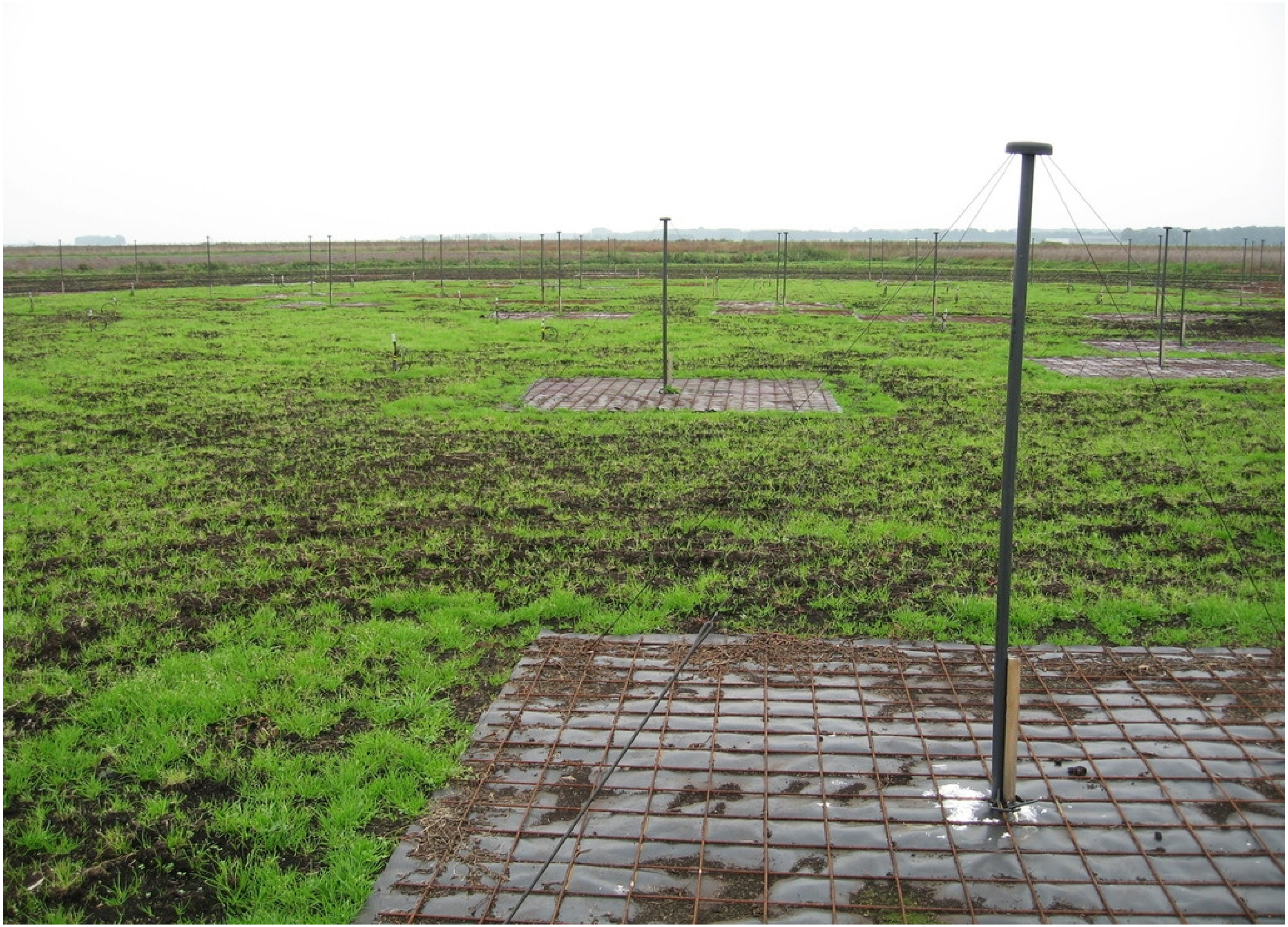
- two antenna types
  - Low-Band Antenna (10–80 MHz)
  - High-Band Antenna (110–240 MHz)
- FM radio range not covered

Low-Band Antennas



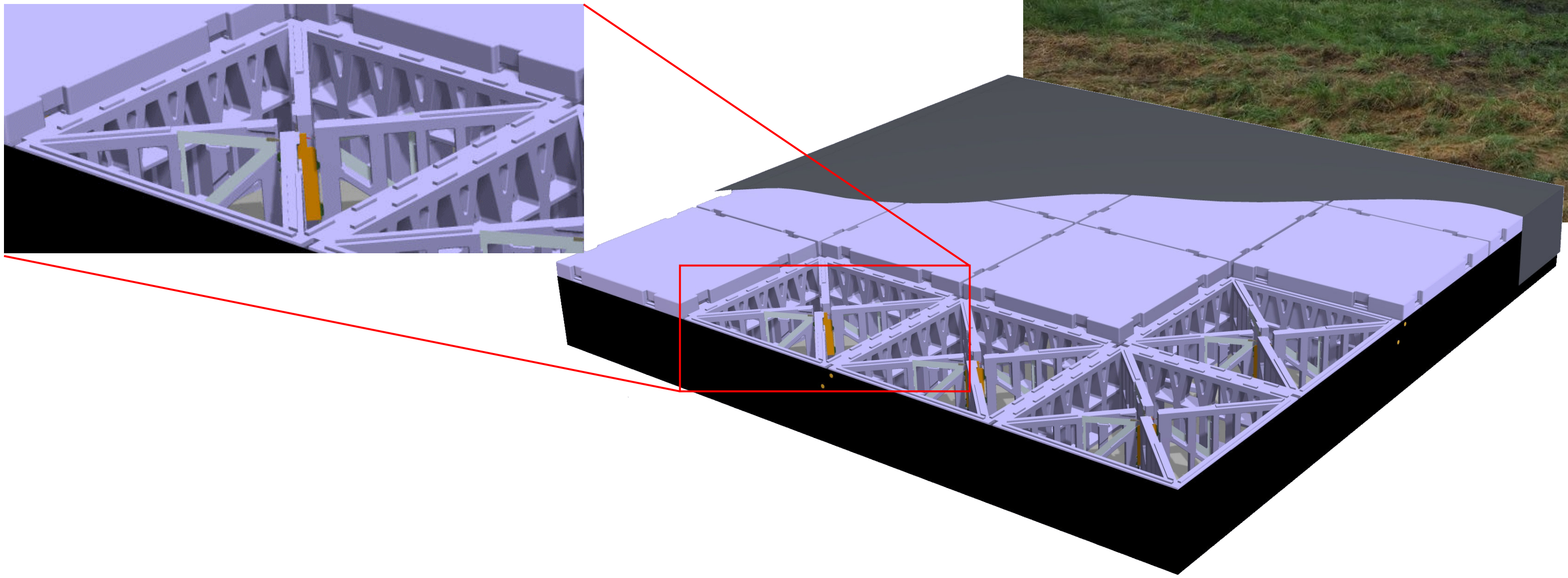
- 10–80 MHz
- dual polarized

LBA Field



HBA Tiles

- 110–240 MHz
- dual polarized
- 4x4 receivers = 1 tile
- analogue beam forming



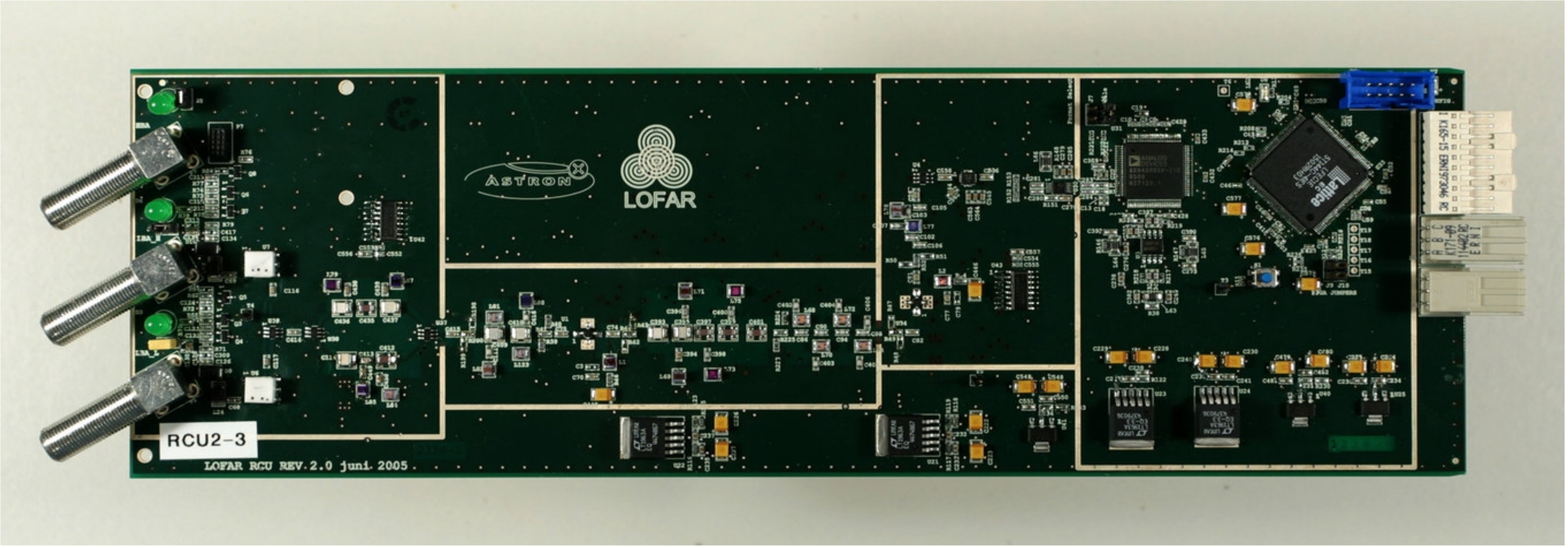
A Station

- 48–96 LBAs
- 48–96 HBA tiles

Station Cabinet

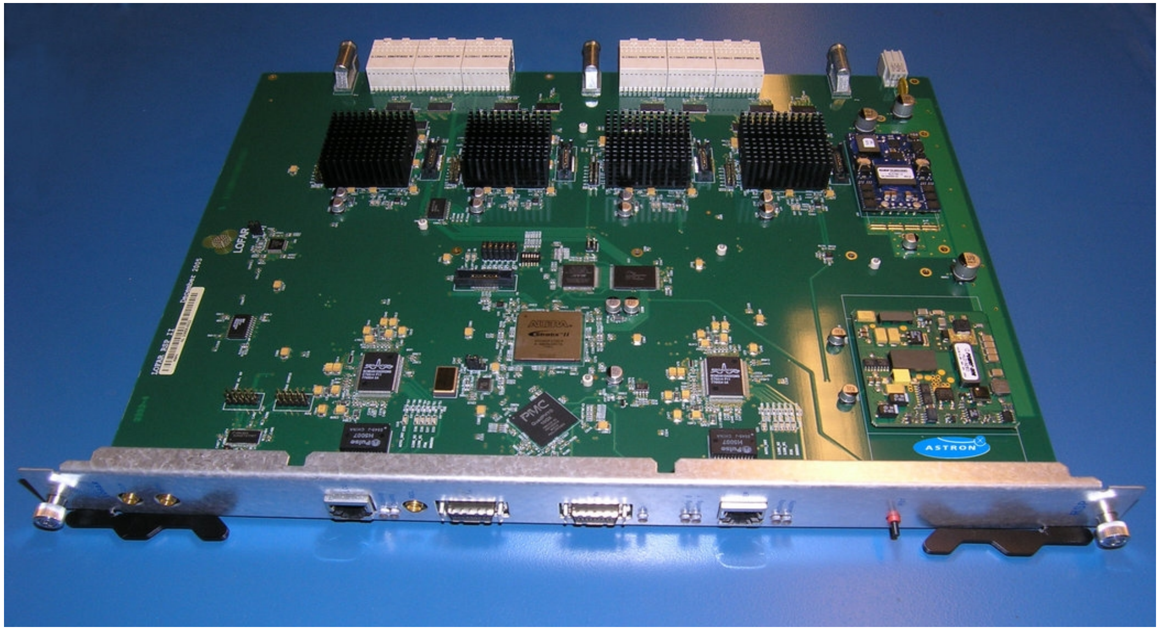
- station processing

Remote Control Unit



- 2 LBAs + 1 HBA tile
- filter
- 200 (or 160) MHz A→D conversion

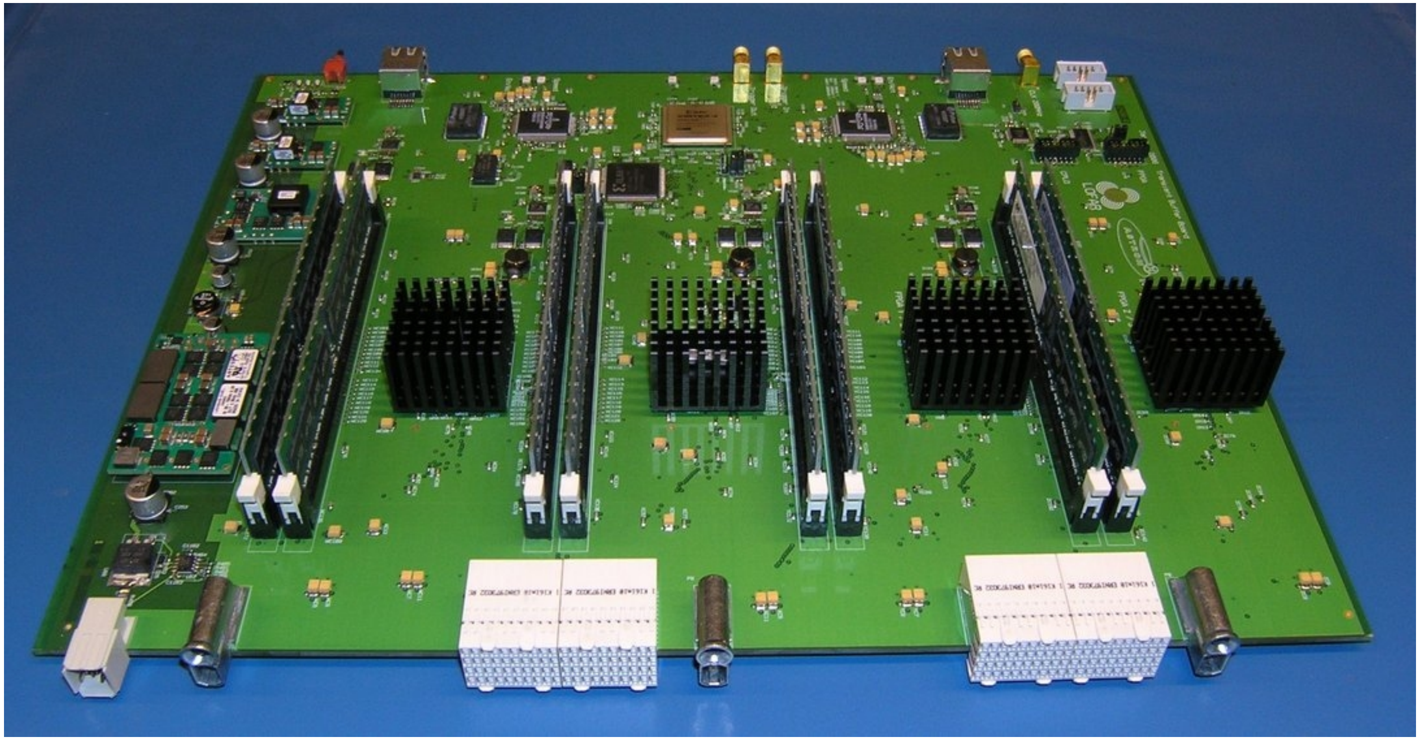
Remote Station Processing Boards



- FPGAs
- PPF: creates 512 * 195 KHz subbands
  - select up to 164 subbands
- beam form LBAs/tiles
- UDP packets over WAN to correlator

Transient Buffer Boards

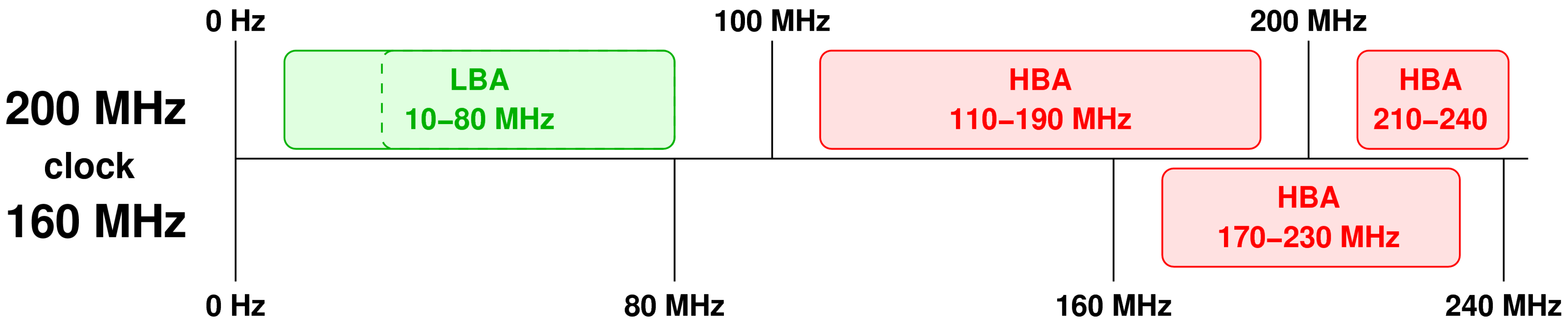
- 4 sec. raw antenna data stored in TBB
- trigger → freeze → dump → post analysis
- not possible with dishes!



Stations

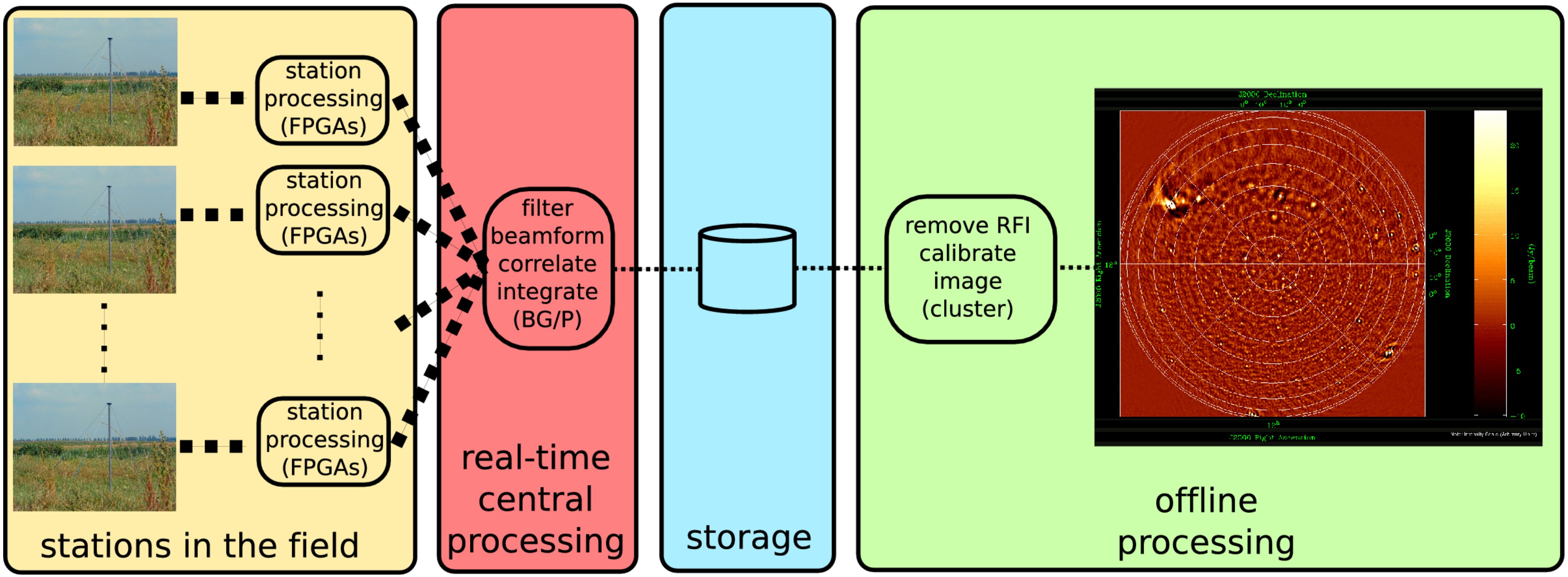
- ≤ 2009: prototypes
- building real stations *now*
  - 18–25 core
  - 18–25 remote
  - 8–20 European
- dedicated fibers to correlator

Observation Characteristics



- 2 polarizations
- 32 MHz bandwidth from 1 mode
  - select 164 * 195 KHz subbands
- up to 8 concurrent observations
  - trade bandwidth for beams

LOFAR Processing



Central Processing Pipelines

- standard imaging mode
- pulsar survey mode
- known pulsar mode
- transients mode
- very/ultra high-energy modes
- ...

Blue Gene History

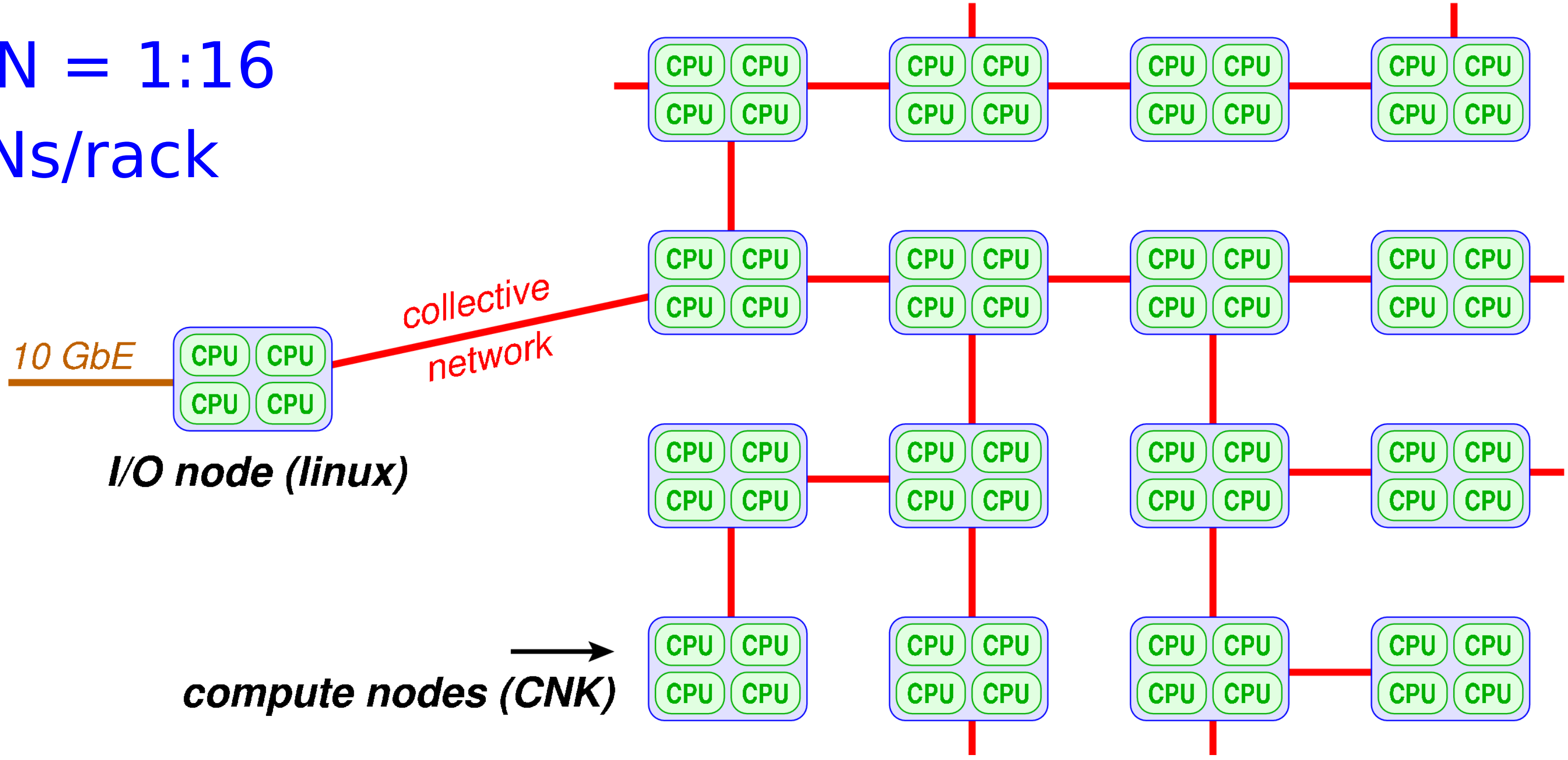
- 6 racks Blue Gene/L (2005–2008)
- 2½ rack Blue Gene/P (2008–)

The Blue Gene/P

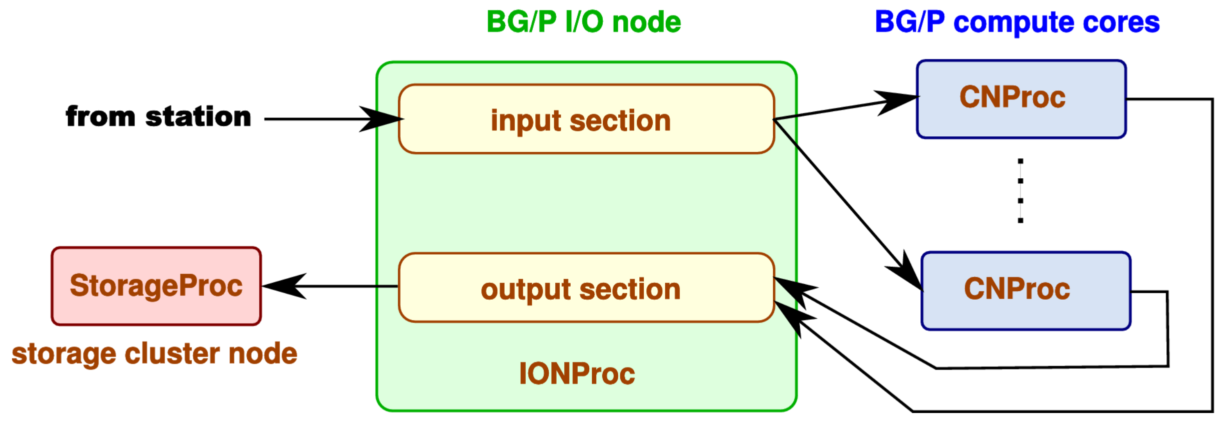
- 850 MHz PPC
  - 4 cores * 2 FPUs * 1 FMA/cycle
  - complex numbers
- 3-D torus, collective, barrier, 10 GbE, JTAG networks
- 2½ racks = 10,880 cores = 37 TFLOP/s + 160*10 Gb/s

BG/P Pset

- I/O Nodes (ION) & Compute Nodes (CN)
- ION handles I/O requests of CN
  - transparent
  - ION:CN = 1:16
  - 64 IONs/rack



The BG/P Correlator

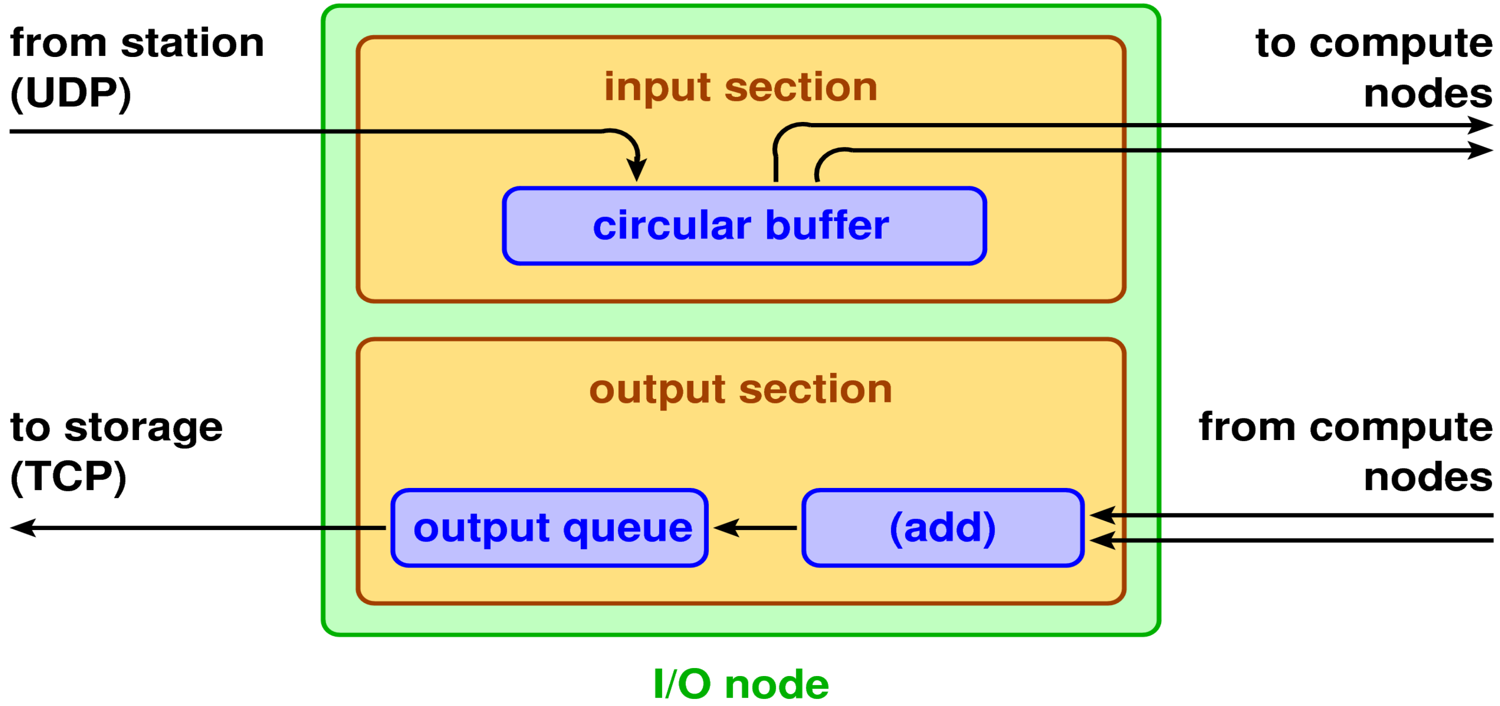


- three distributed applications/platforms
  - BG/P I/O nodes (ION)
  - BG/P compute nodes (CN)
  - external storage nodes

Application Software on I/O Node

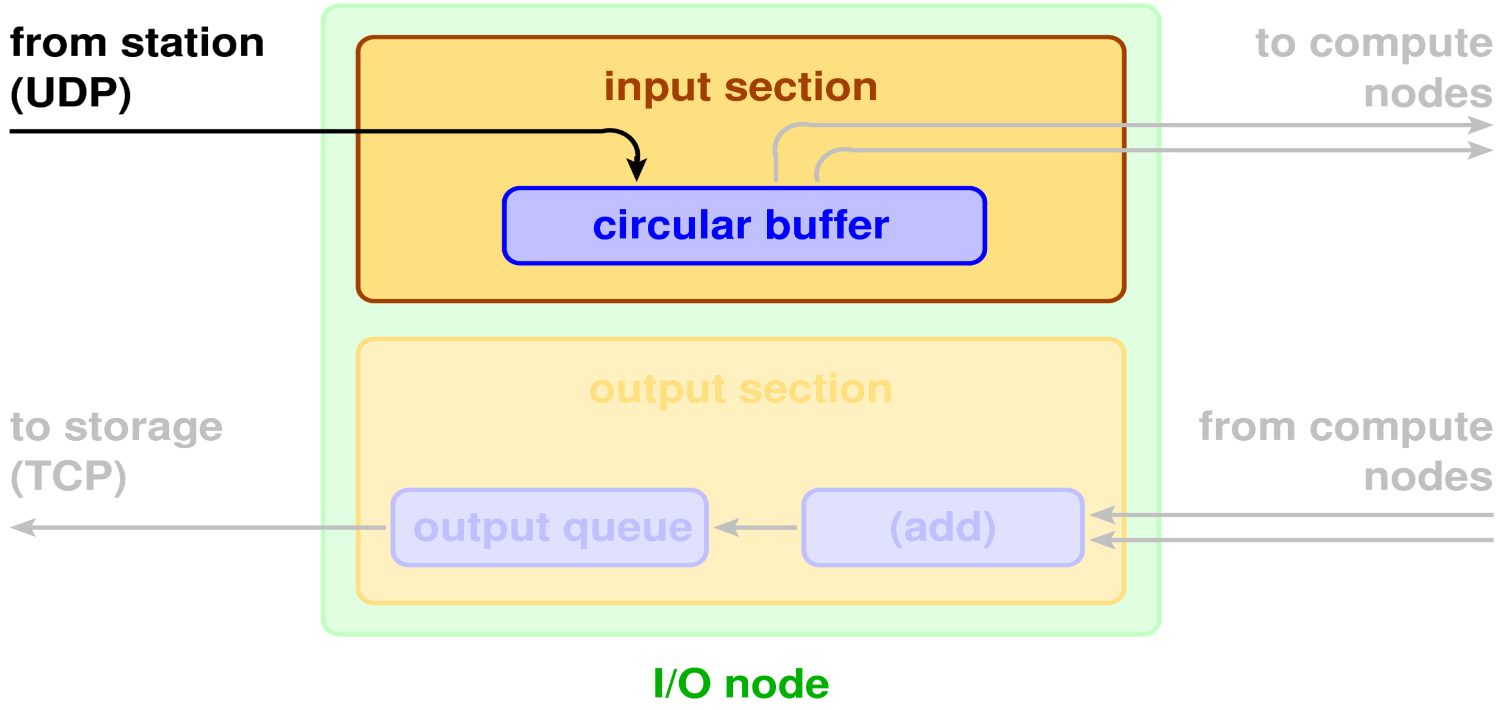
- unorthodox
- more efficient & flexible
- BG/L: saved costs; for input cluster
- BG/L: major system software changes (ZOID) *(thanks ANL!)* [PPoPP'08]
- BG/P: better support

I/O Node Processing



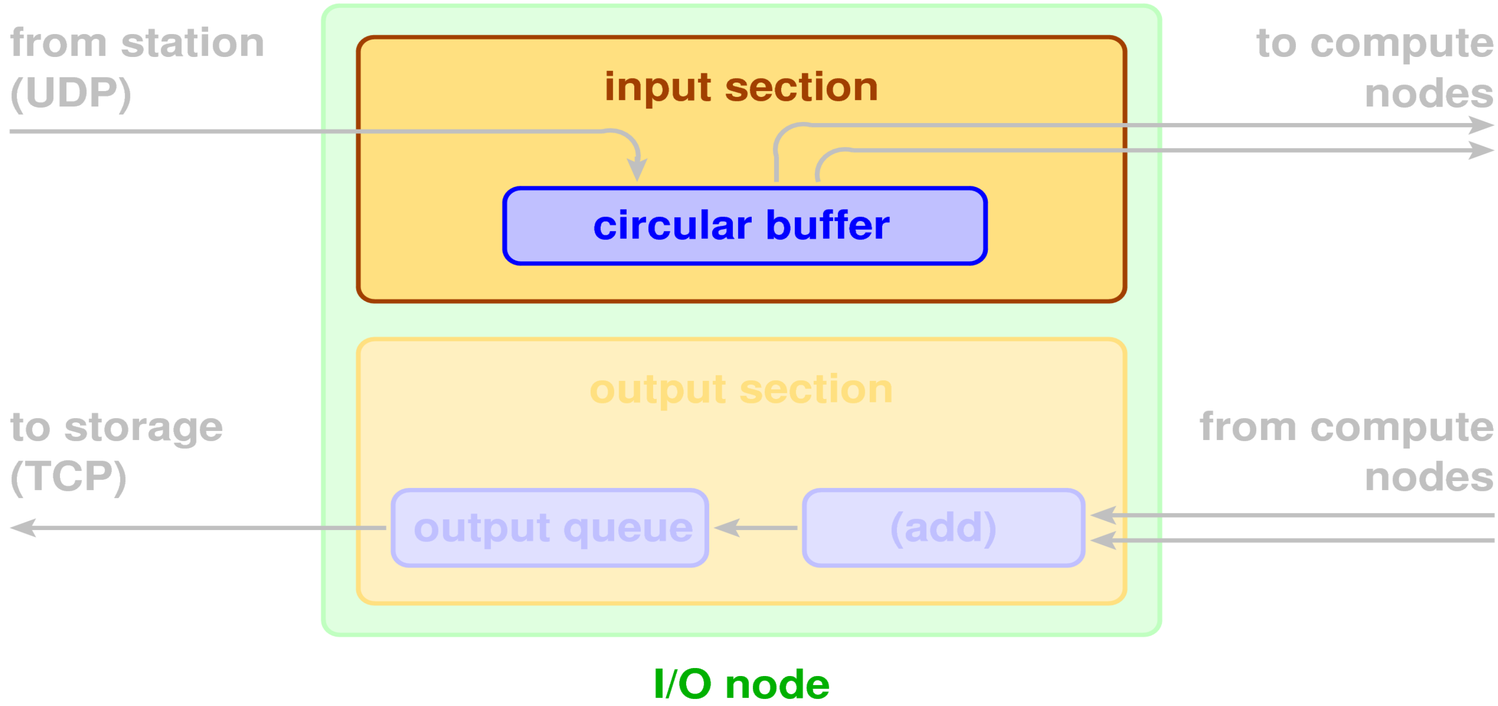
- two sections
  - input
  - output
- multi threaded

I/O Node Input Section

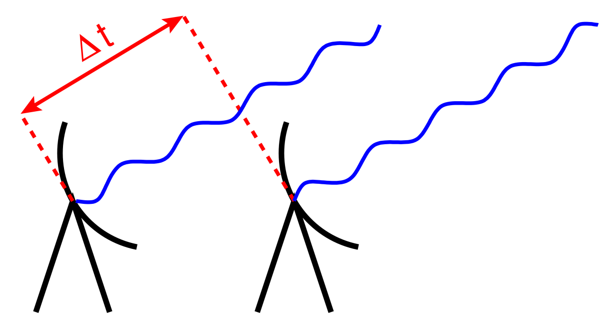


- ION receives from 1 station
  - 48,828 pkt/s
  - handles missing packets

Circular Buffer

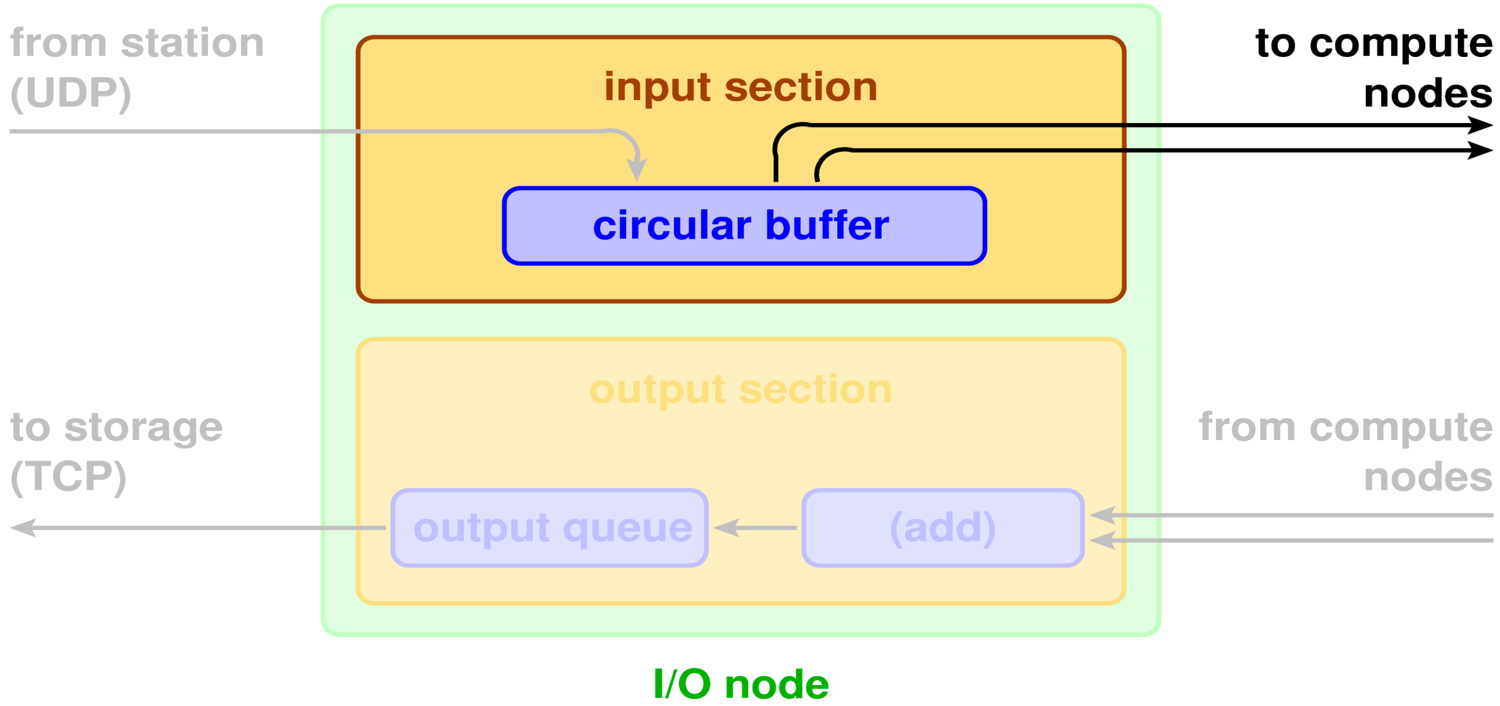


- circular buffer (~2.5 s)
  - WAN delays
  - delay stream
  - handle hiccups



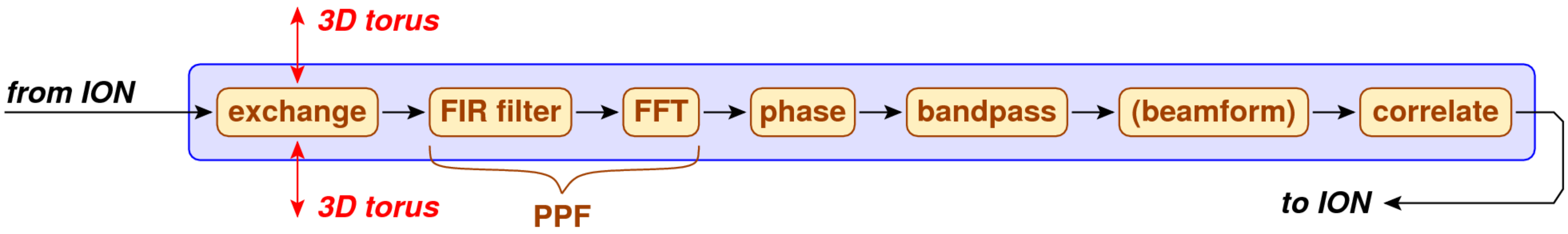
$\Delta t = 22\mu s \approx 4 * 5.12\ \mu s$ samples

I/O Node → Compute Node

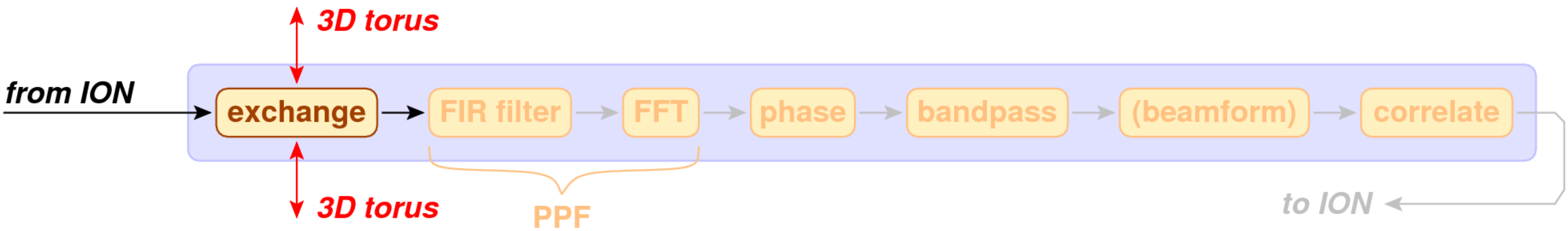


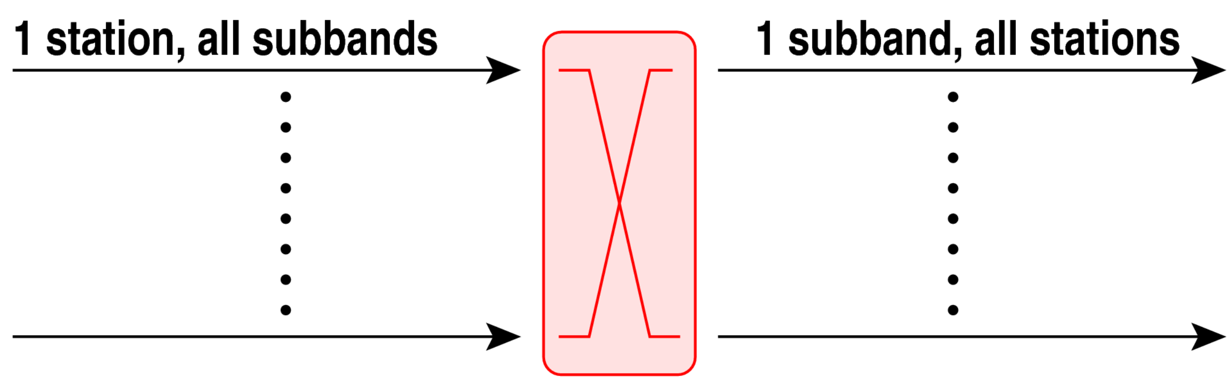
- ION sends data to CN
  - wall-clock time trigger
- chunk
  - = 196,608 samples (1.007 s), 1 subband, 2 pols, 1 station

Compute Node Processing



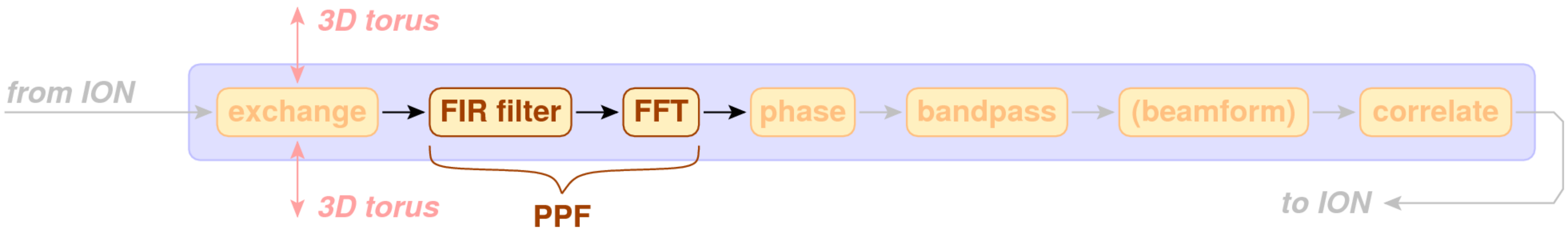
Exchange

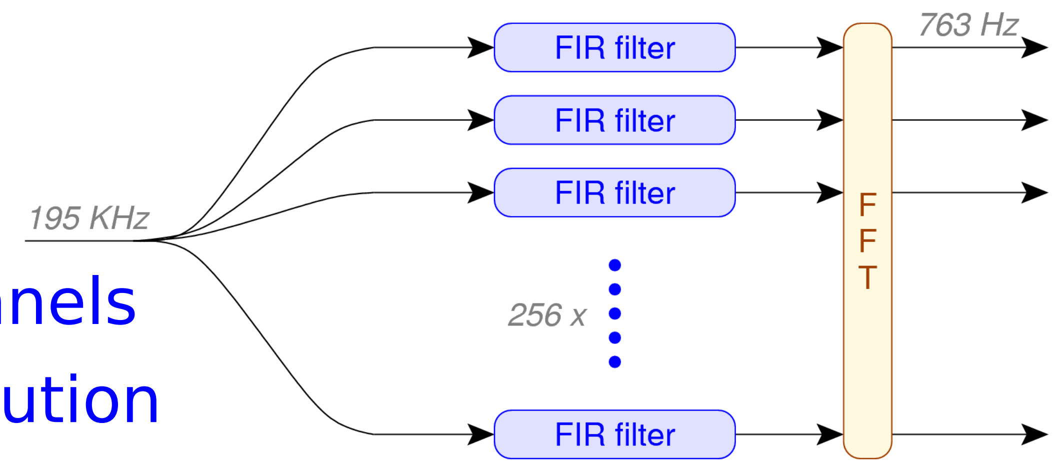




- hundreds of Gb/s
- asynchronous

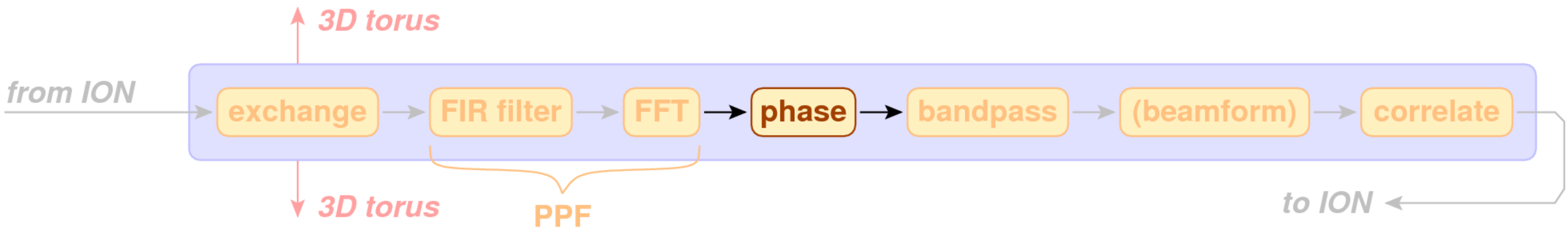
PolyPhase Filter



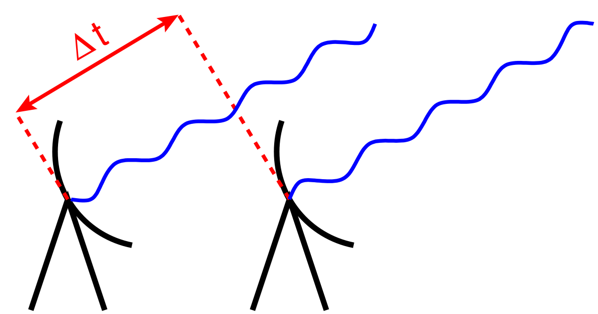


- splits subband into channels
- time vs. frequency resolution
- FIR filter + FFT
- allows narrow-band RFI removal

Phase Correction

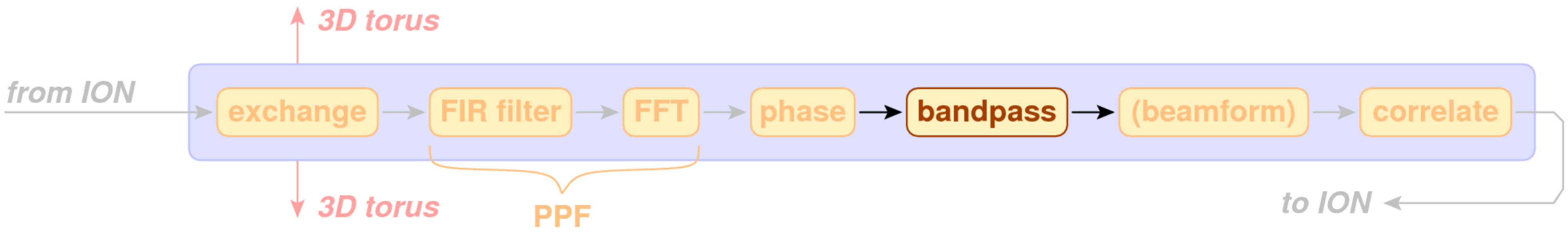


- correct observation direction
  - already shifted samples — correct rest
- interpolate

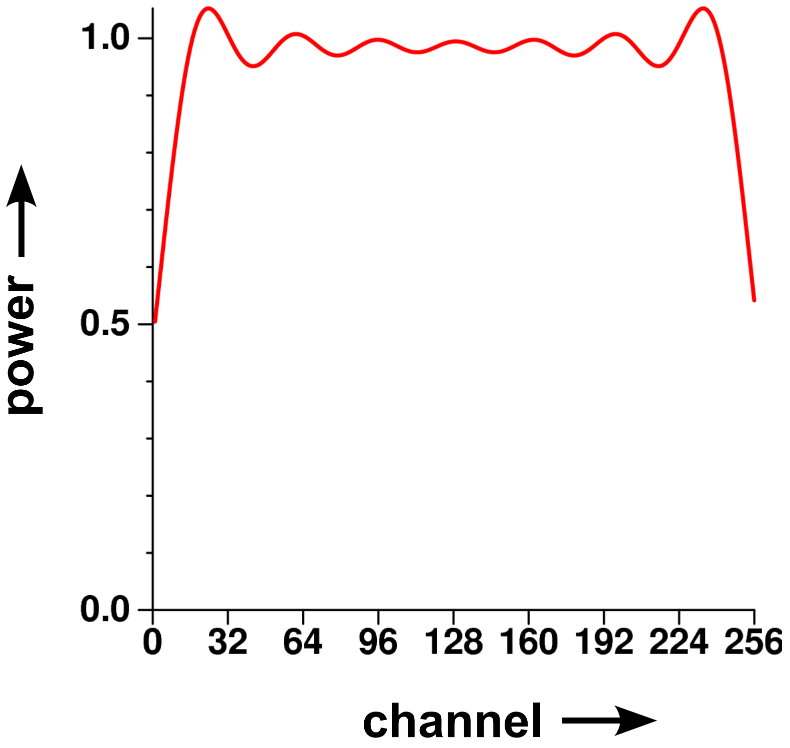


$\Delta t = 22\mu s = 4 * 5.12\ \mu s$ samples $+\ e^{-2i\pi f * 1.52}$

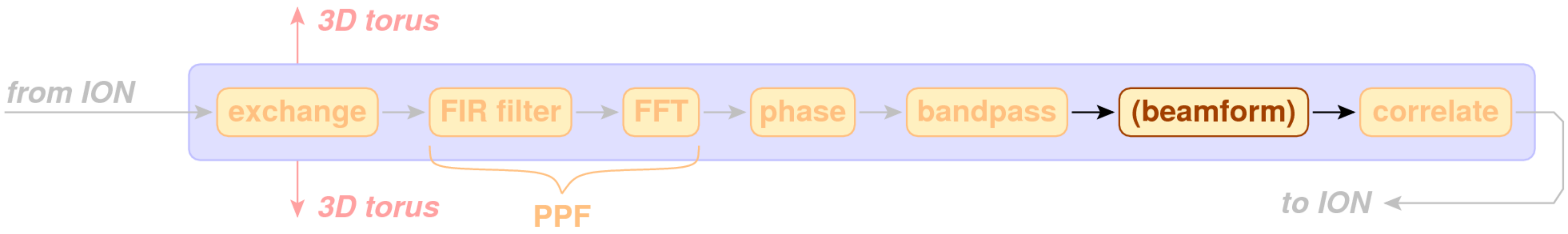
Band Pass Correction



- channel powers unequal
  - caused by station PPF
- correct

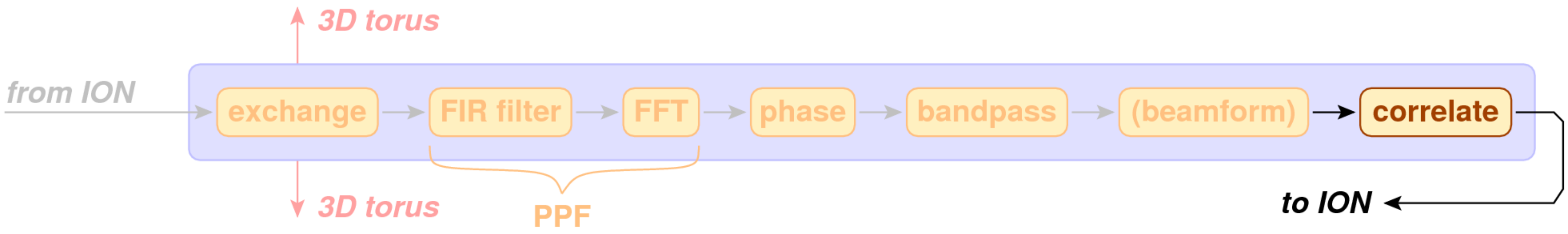


Beam Forming



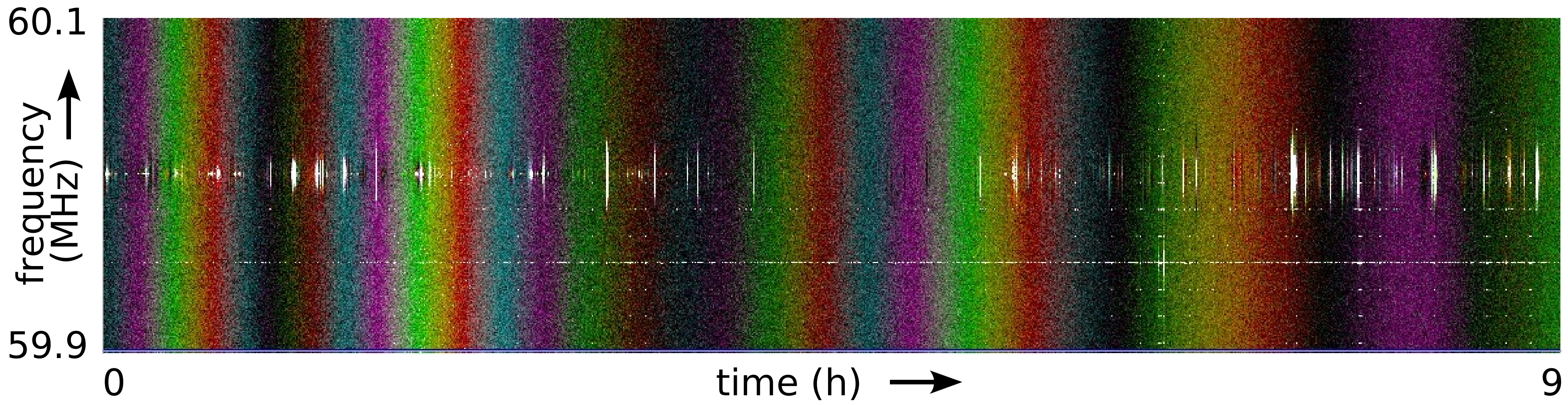
- add group of stations to form “superstation”
- optional

Correlate



- filters noise
- multiply samples of all *pairs* of stations
- integrate over time

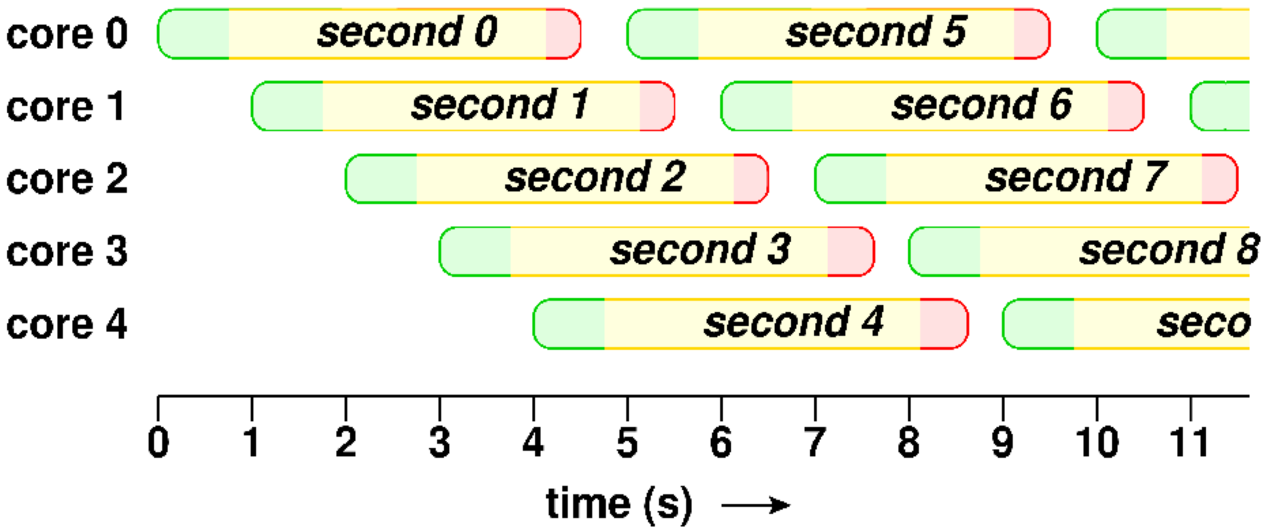
Correlator Output



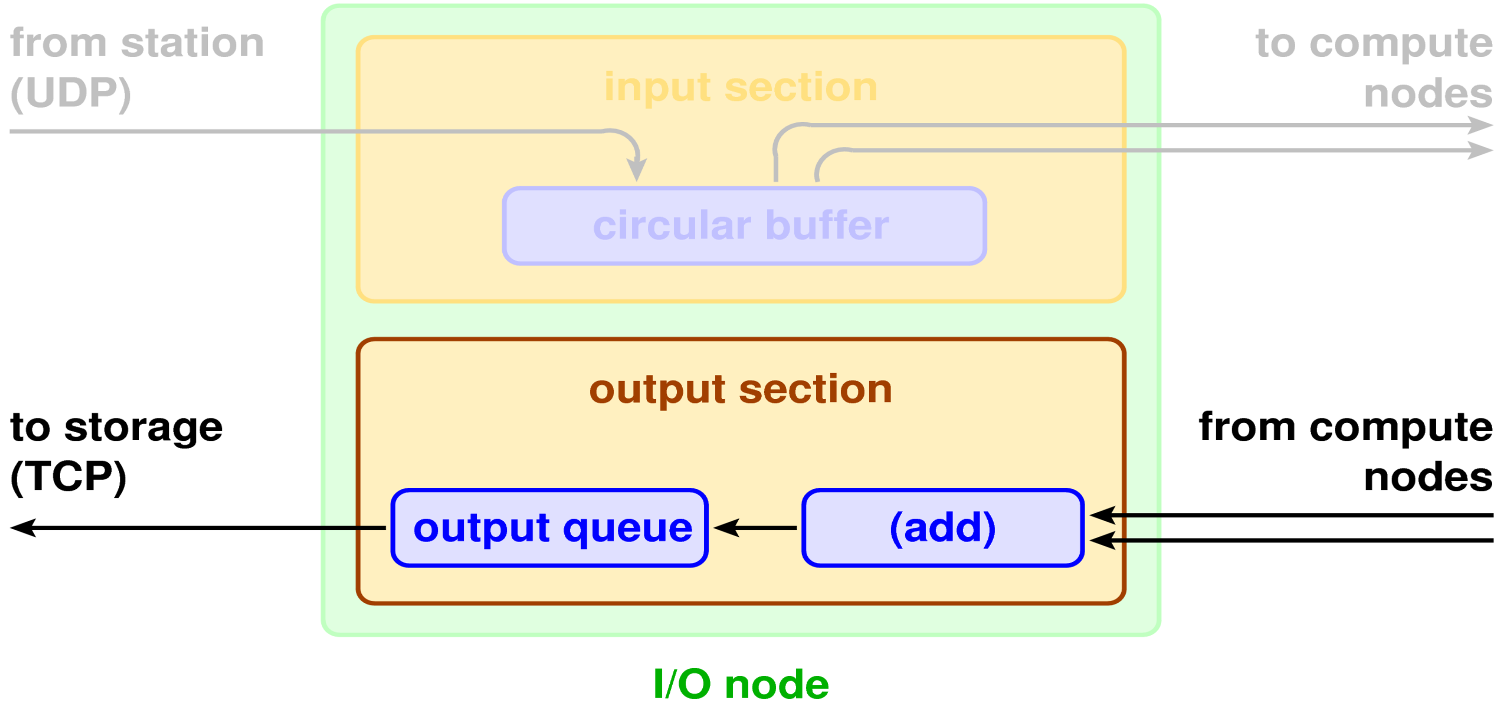
- correlations between two stations
- color = phase, intensity = power
- combined contribution of (strong) sources
- earth rotation changes phase

Work Distribution

- process subbands independently
  - stations must be combined
- chunk needs > 1 second processing time
  - round-robin distribution
  - receive, process, send, idle
- OVERLY SIMPLIFIED!

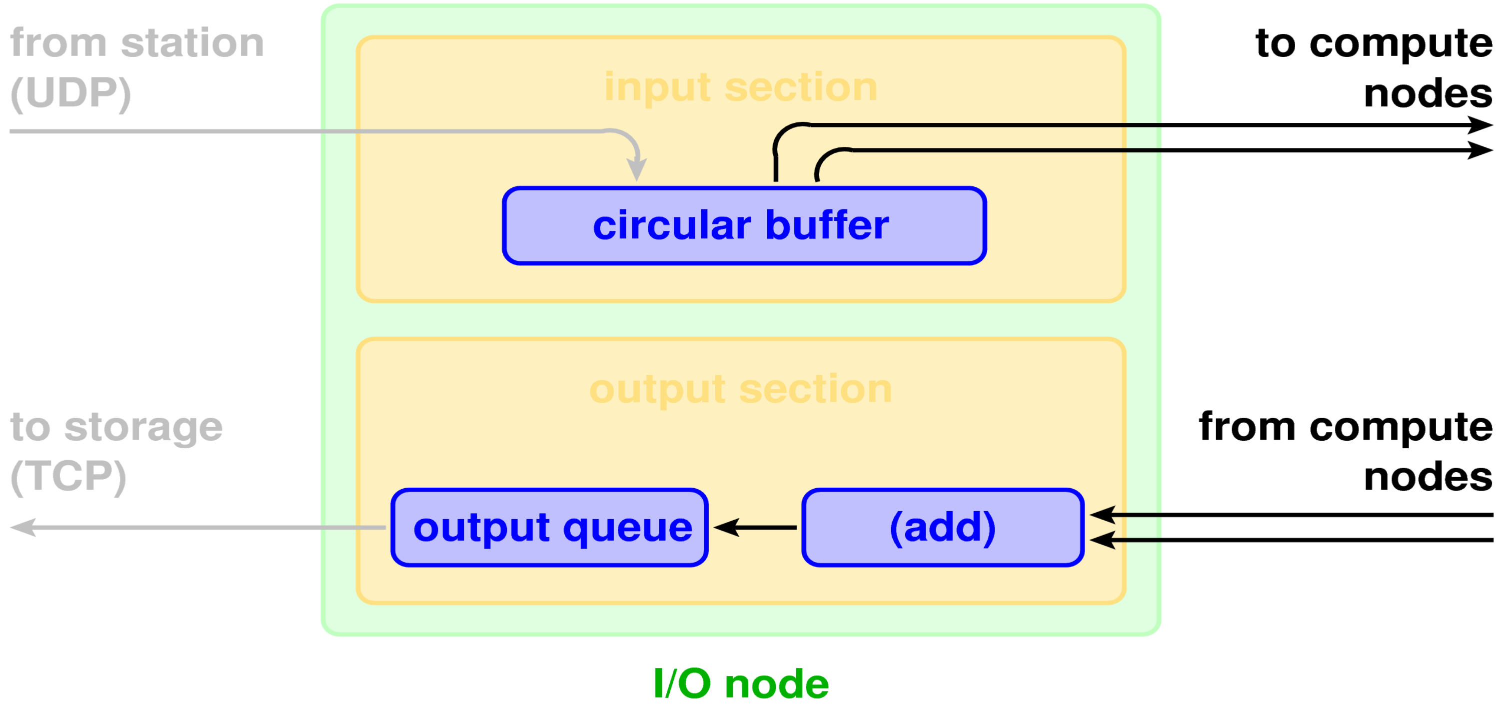


I/O Node Output Section



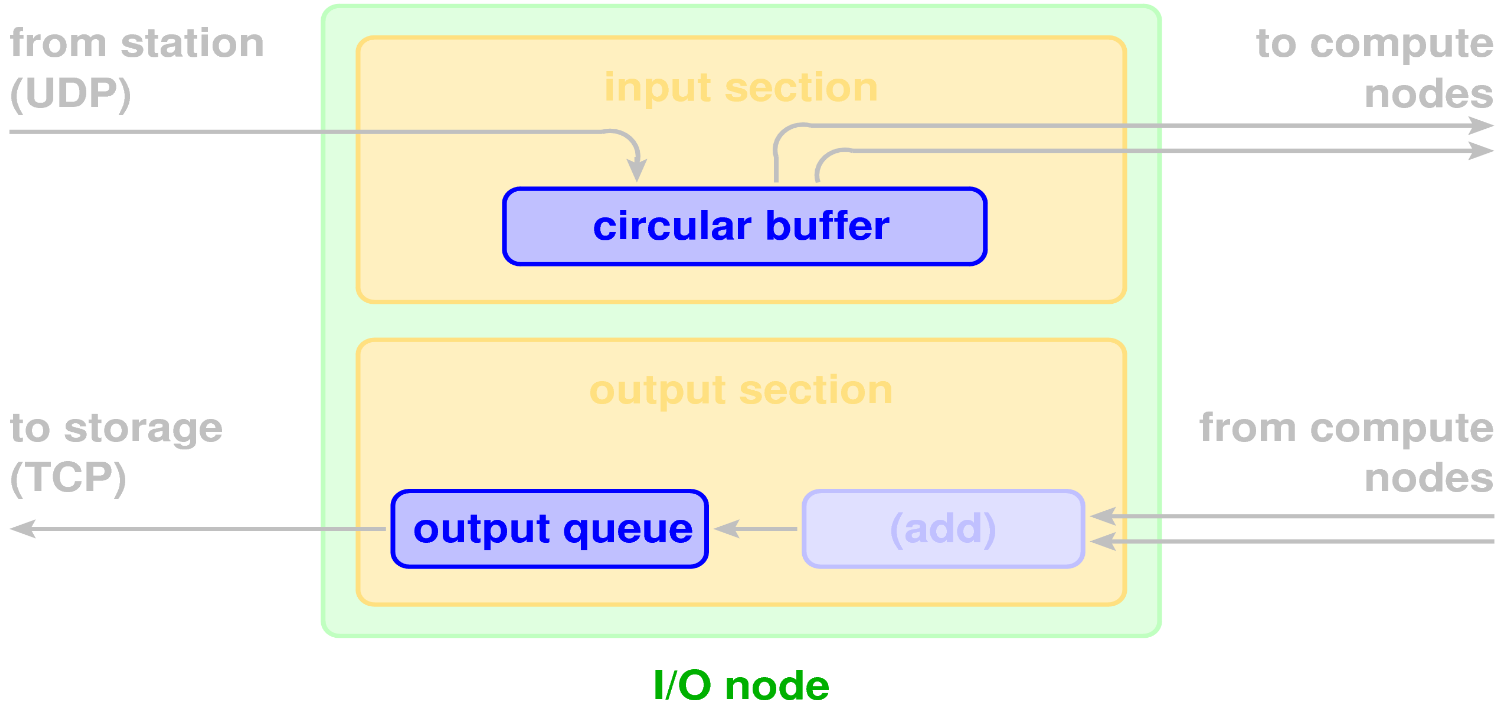
- (adds correlations)
- best-effort queue
  - ensures real-time continuation of correlator

I/O Node Real-Time Scheduling



- use Linux RT scheduler

I/O Node Memory



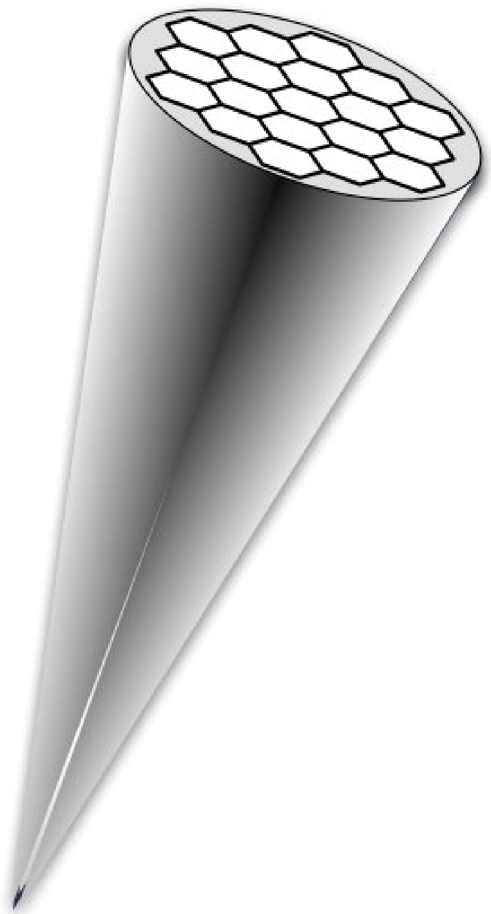
- PPC 450: software TLB-miss handler [P2S2'09]
  - Linux: slows down applications by 40%−300%
- modified kernel to provide 6 * 256 MiB “fast” pages *(thanks ANL!)*

Storage

- correlations saved on disk
  - external cluster
  - ~1 PB
  - post-processed within week

Pulsar Pipelines

- find & observe pulsars
- beam form instead of correlate
  - 5 pipeline flavors
- functional; needs optimizations
- correlate & beam form concurrently

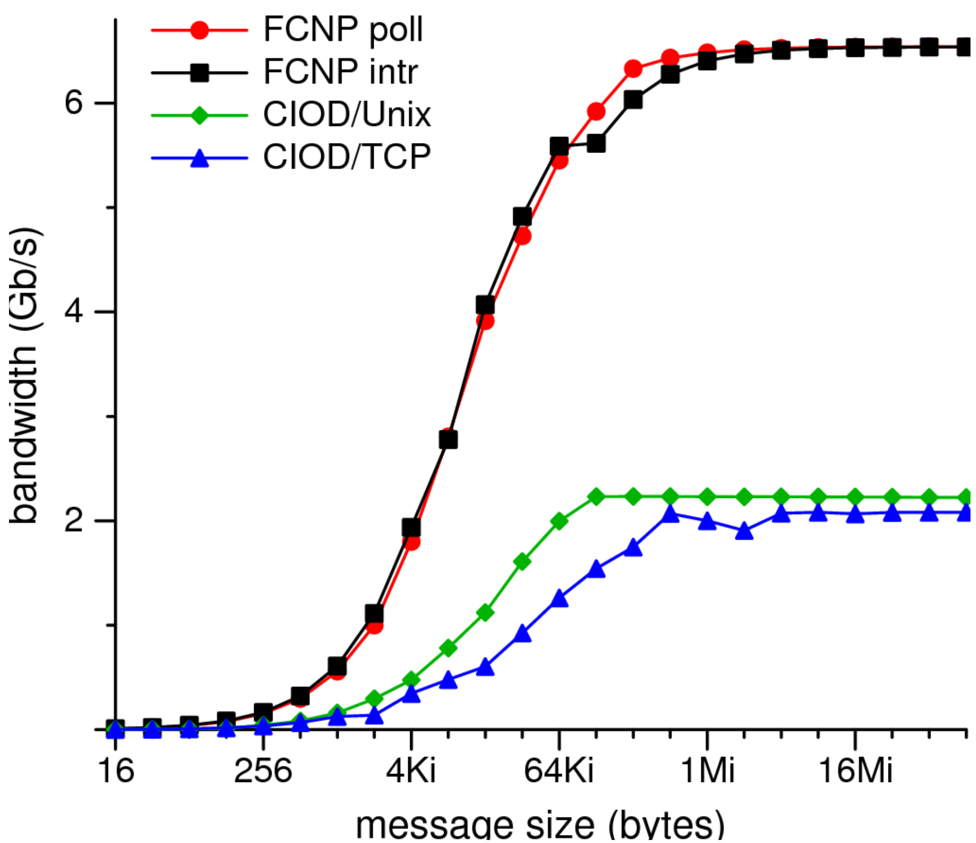


Communication

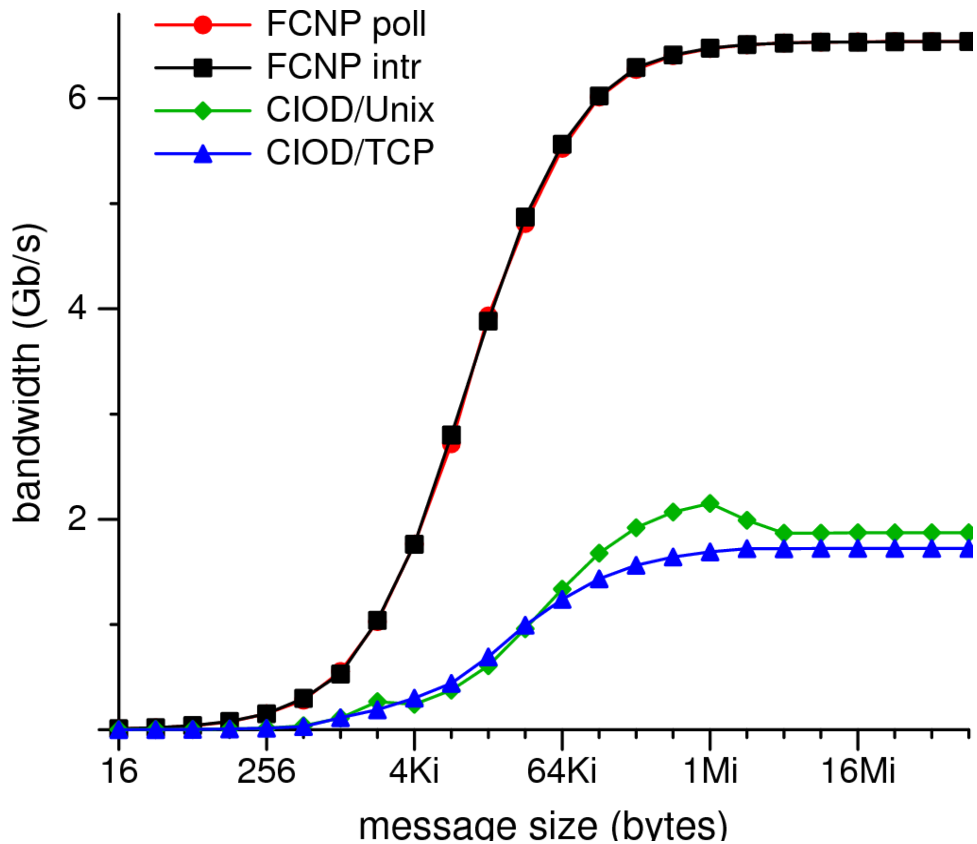
Fast Collective Network Protocol

- ION ⇔ CN bandwidth insufficient
  - socket overhead
  - core hardly keeps up with network
- new ION ⇔ CN protocol [PDPTA'09]
  - low overhead
    - user space
  - simultaneous send & receive
  - uses free virtual channel *(thanks IBM!)*
  - supports interrupts *(thanks IBM!)*

FCNP Performance



ION → CN



CN → ION

- approaches link speed

Correlator Performance

Optimizations

- correlator, beam former, FIR filter, FFT written in assembly
  - goal: 4 FLOPS/cycle
    - minimize memory accesses
    - use L2 prefetch units
    - influence cache behavior
    - concurrent loads/stores & FPU ops
    - hide load & FPU latencies
  - ~10x faster than C++

FPU Efficiency

	GFLOP	time (s)	efficiency
FIR	1.61	0.553	86%
FFT	0.812	0.553	43%
Correlate	12.9	3.96	96%

- one chunk, 64 stations
- 256-point FFT: 8262 ops (< $5n \log n$)

How Fast Can We Go?

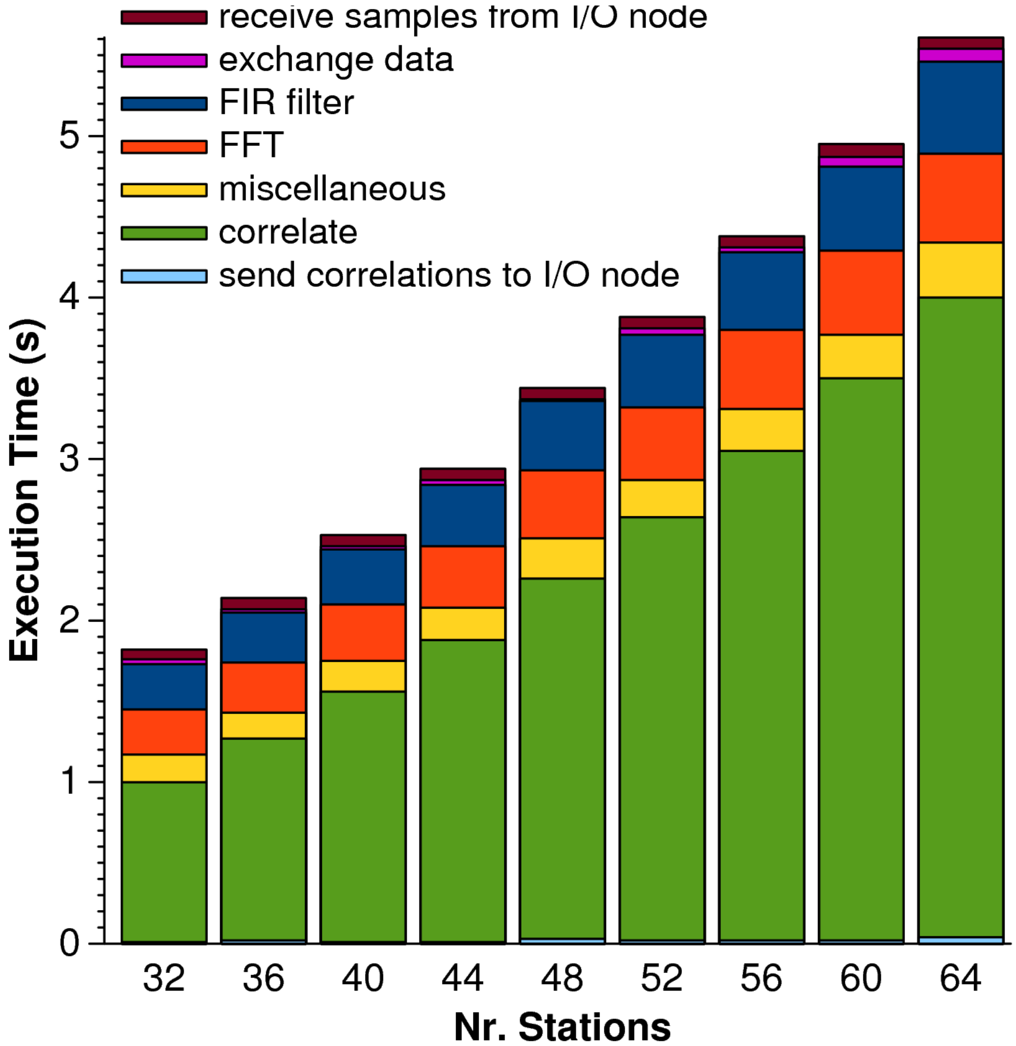
	required	possible
station	32 MHz ~ 2.05 Gb/s	48.4 MHz ~ 3.1 Gb/s
WAN	2.05 Gb/s	10 Gb/s
correlator	32 MHz ~ 2.05 Gb/s	???

- goal:
  - process 50% more data ...
  - ... using 40% of the hardware

Correlator Performance

- test setup
  - 1 rack generates data
  - 1 rack correlates
  - ½ rack ~~“stores”~~ data
- realistic simulation
- up to 64 stations

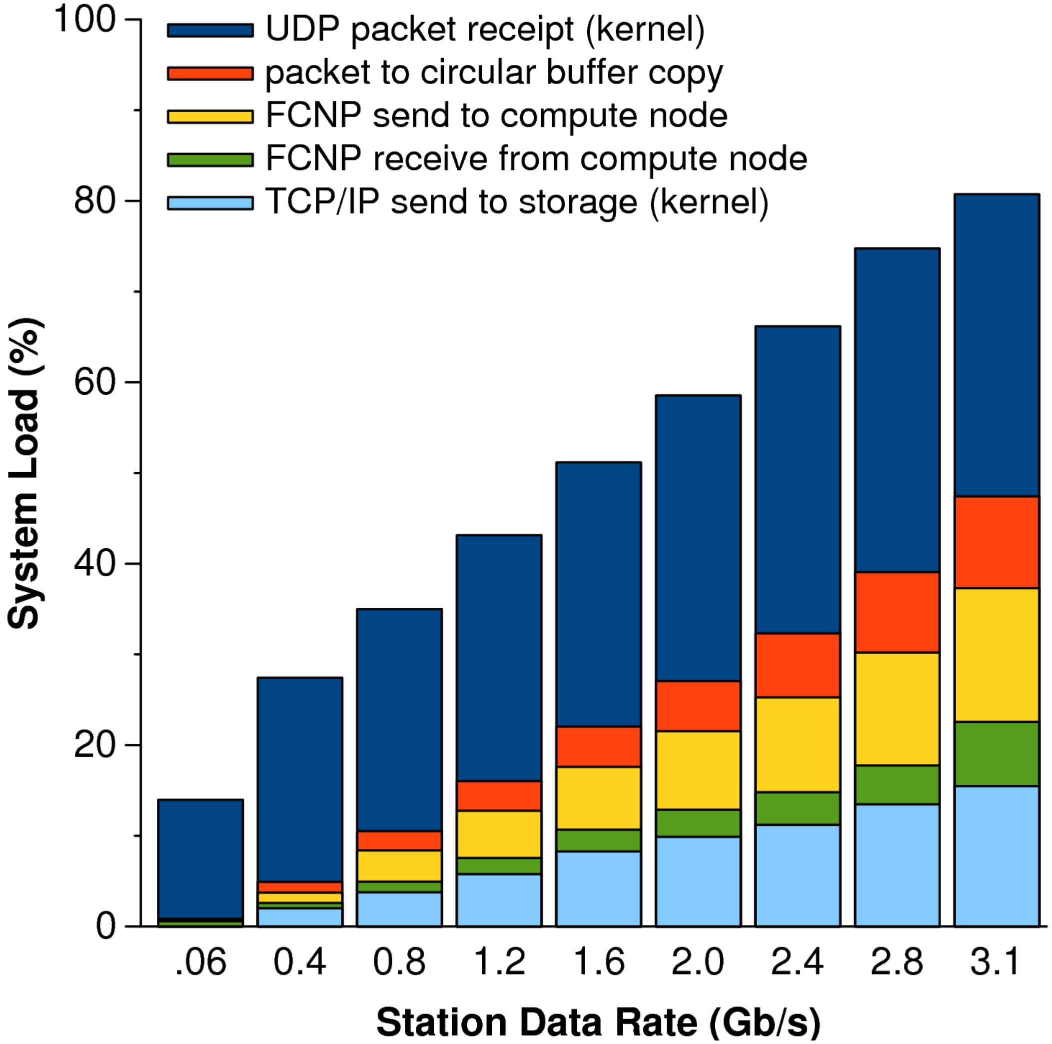
Compute Node Scaling



- 1 chunk, ≤64 stations
- correlate: $O(n^2)$

I/O Node Scaling

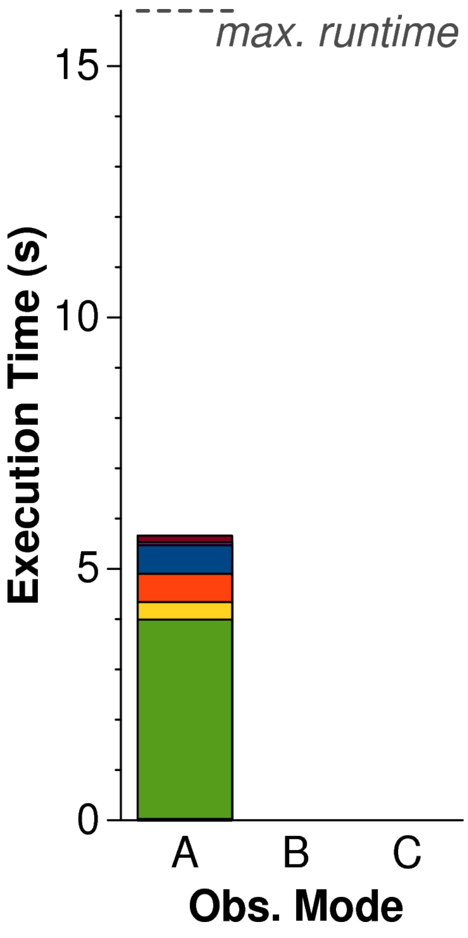
- increase station bandwidth
- ≤3.1 Gb/s in; ≤1.2 Gb/s out
- IP stack expensive
- >84% load: data loss



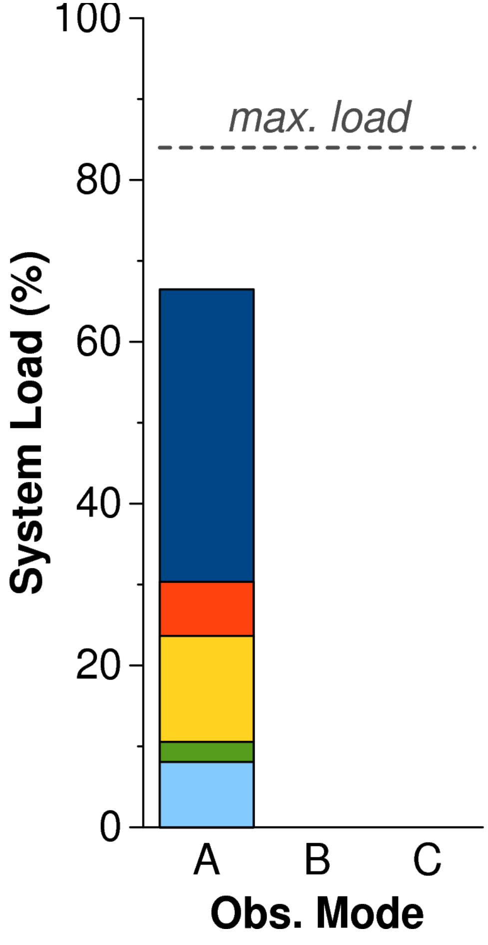
Observation Mode A

observation mode	A
#stations	64
#bits/sample	16
#subbands	248
ION I/O (Gb/s)	3.1+0.58
CPU load CN	35%
CPU load ION	67%

- standard mode
  - 50% more subbands



CN



ION

Three (Future) Station Modes

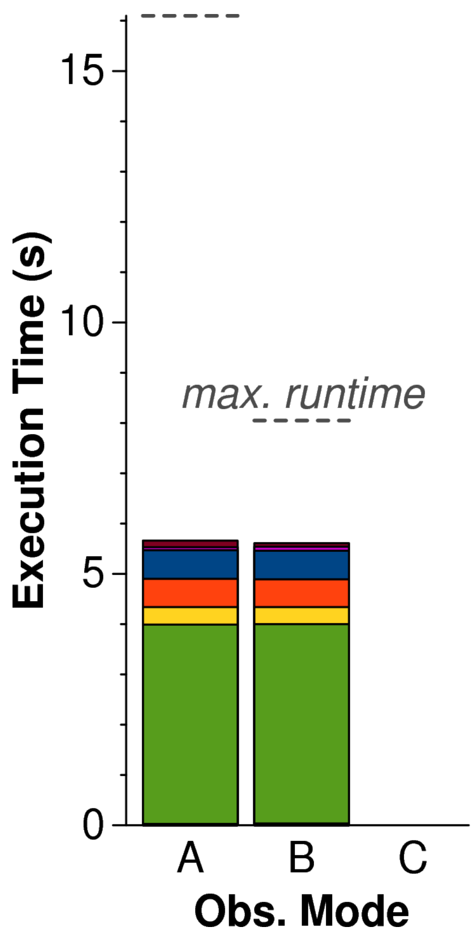
mode	bits/sample	#subbands	Gb/s
A	16	248	3.1
B	8	496	3.1
C	4	992	3.1

- trade accuracy for subbands
- station data rate unaffected
  - correlator: 2x #subbands ⇒ 2x work; 2x output!

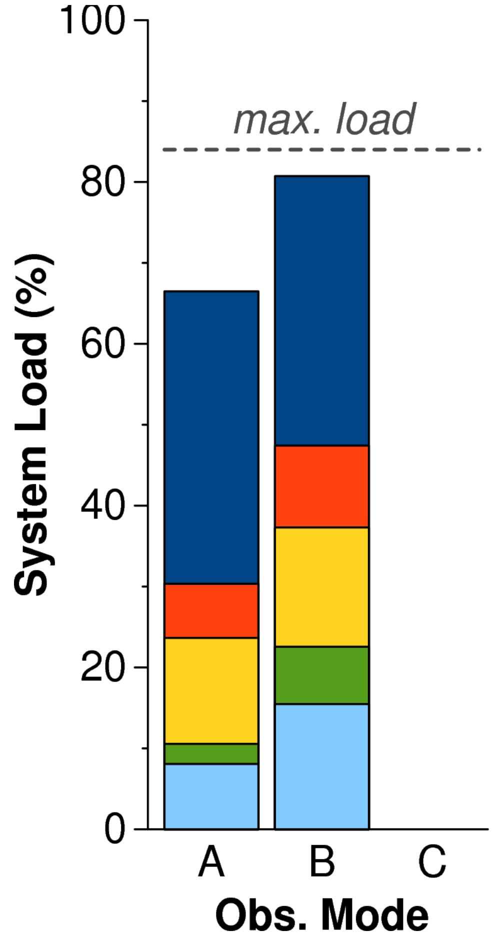
Observation Mode B

observation mode	B
#stations	64
#bits/sample	8
#subbands	496
ION I/O (Gb/s)	3.1+1.2
CPU load CN	70%
CPU load ION	81%

- halved bits/sample
- doubled #subbands
- 275 Gb/s



CN

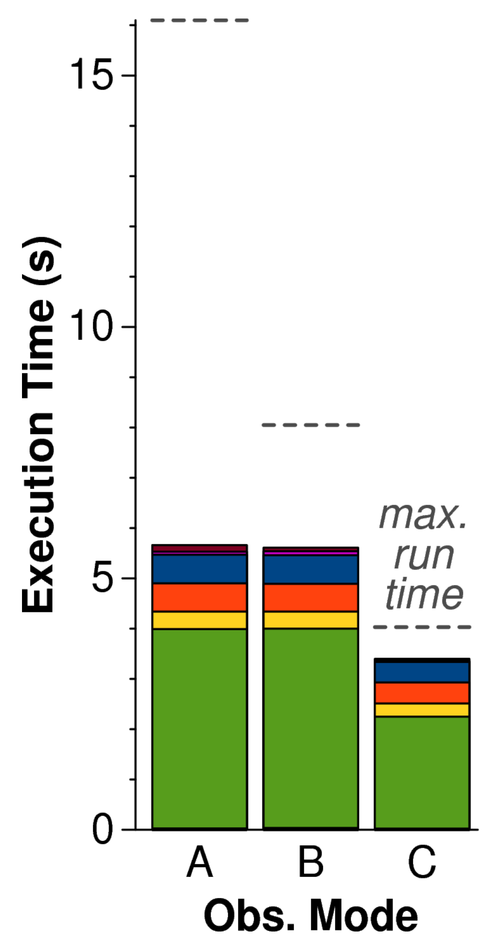


ION

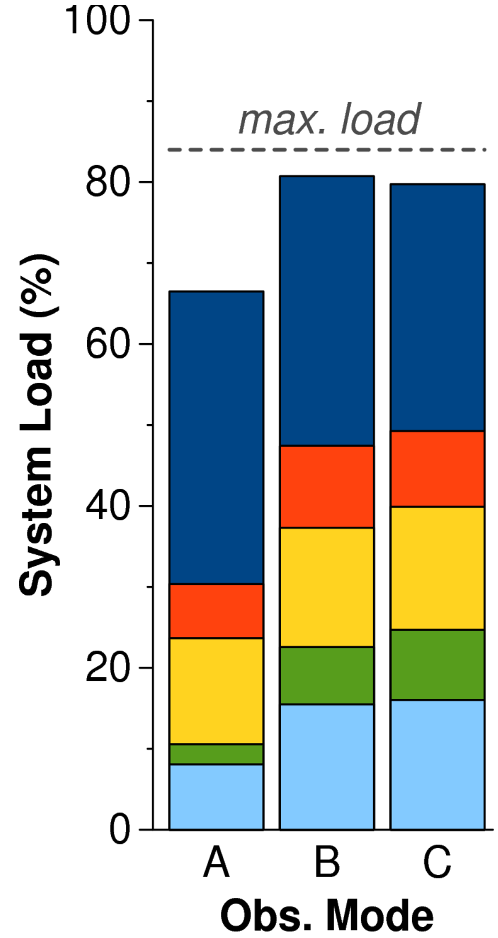
Observation Mode C

observation mode	C
#stations	48
#bits/sample	4
#subbands	992
ION I/O (Gb/s)	3.1+1.3
CPU load CN	85%
CPU load ION	80%

- Epoch of Reionization
- reduced #stations
- >9.3 GFLOP/s



CN



ION

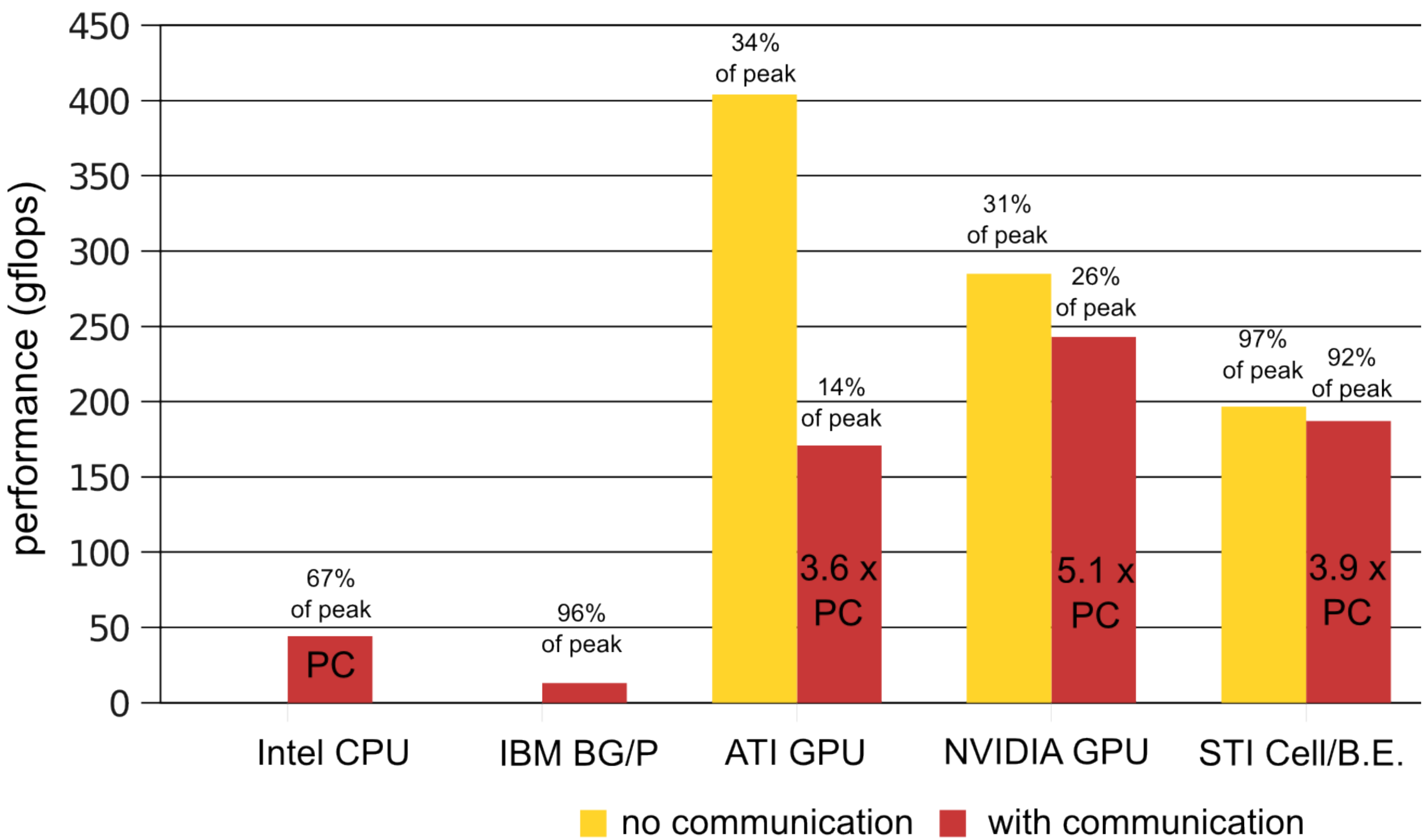
Performance Conclusions

- can process all foreseeable modes
  - at 50% more bandwidth
  - using 1 rack only!
- changed the specs!

BG/P: The Right Choice?

- compared correlator performance of BG/P, Cell BE, GTX 280, RV770, Core i7 [ICS'09]
- written in assembly
  - compiler quality unimportant

Many-Core Comparison



Percentages are the fraction of the
theoretical peak performance **for that architecture**

Many-Core Comparison (2)

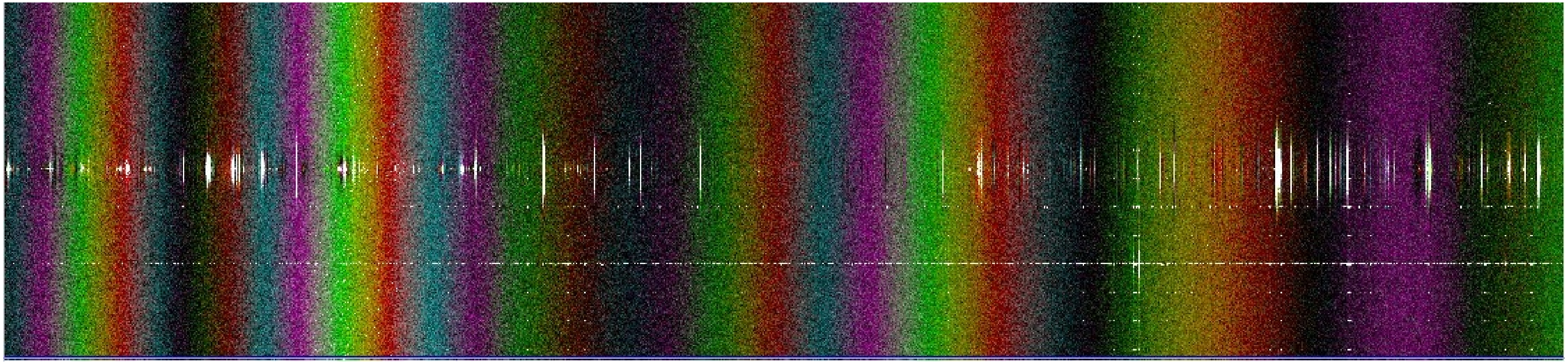
Architecture	Intel Core i7	IBM BG/P	ATI 4870	NVIDIA C1060	STI Cell
measured gflops	48	13.1	171	243	187
achieved efficiency	67%	96%	14%	26%	92%
measured bandwidth (GB/s)	19	6.6	47	94	50
bandwidth efficiency	73%	48%	41%	93%	192%
achieved gflops/Watt	**0.37**	**0.54**	**1.07**	**1.00**	**3.74**

- Cell BE wins, due to software-managed cache
- GPUs are I/O bound
- BG/P: built-in interconnect; densely packed

Off-Line Processing

- flagging
- self calibration
- imaging

Flagging



- invalidate RFI
  - mostly narrow band
- several algorithms

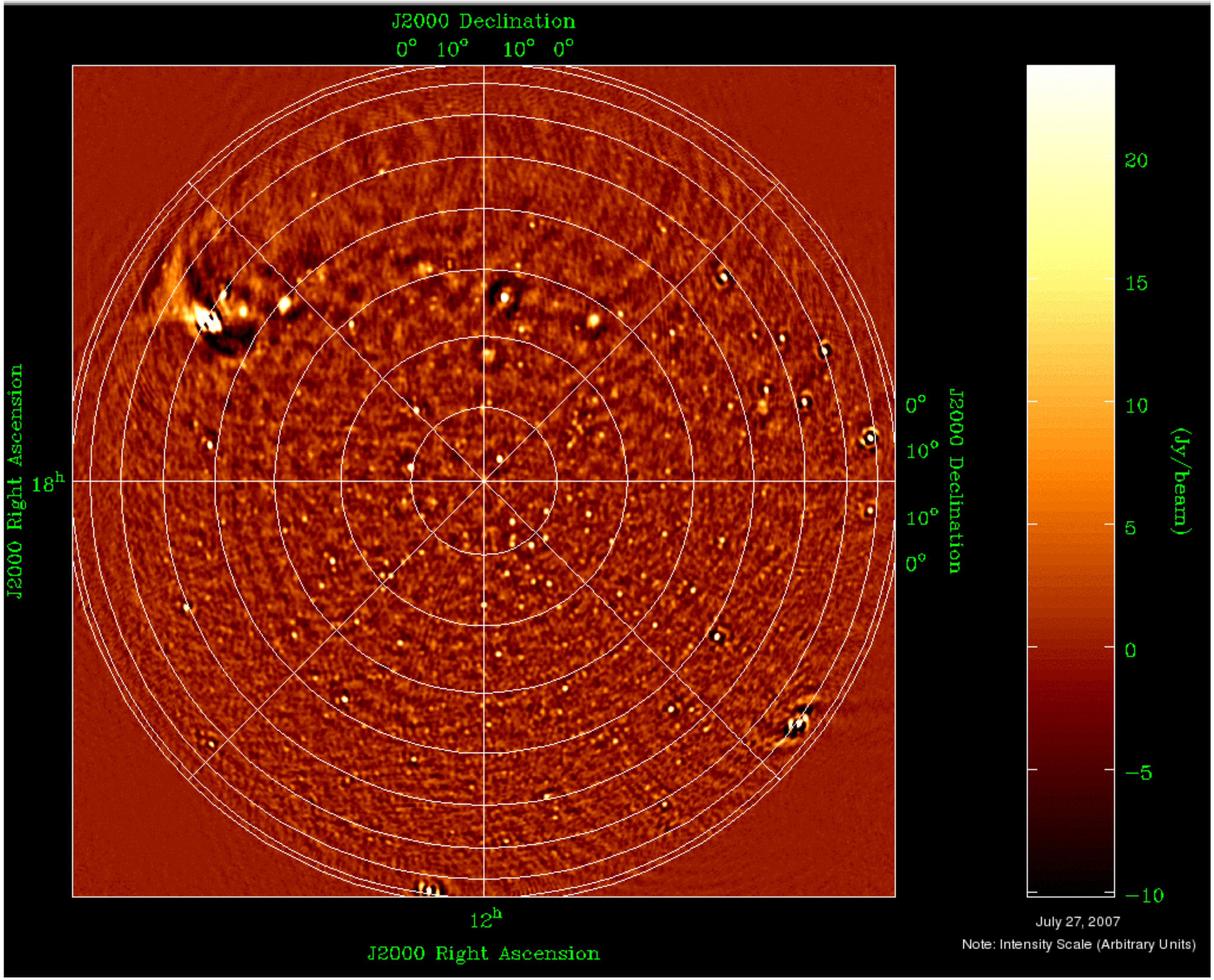
Self-Calibration

- newly developed algorithm
- correct instrumental, environmental errors & sky parameters
- Global Sky Model
  - pos, flux, pol of O(100,000,000) sky objects
  - continuously refined
- subtract bright sources
- compare predicted & measured data
- solve
- need another supercomputer ...

Imaging

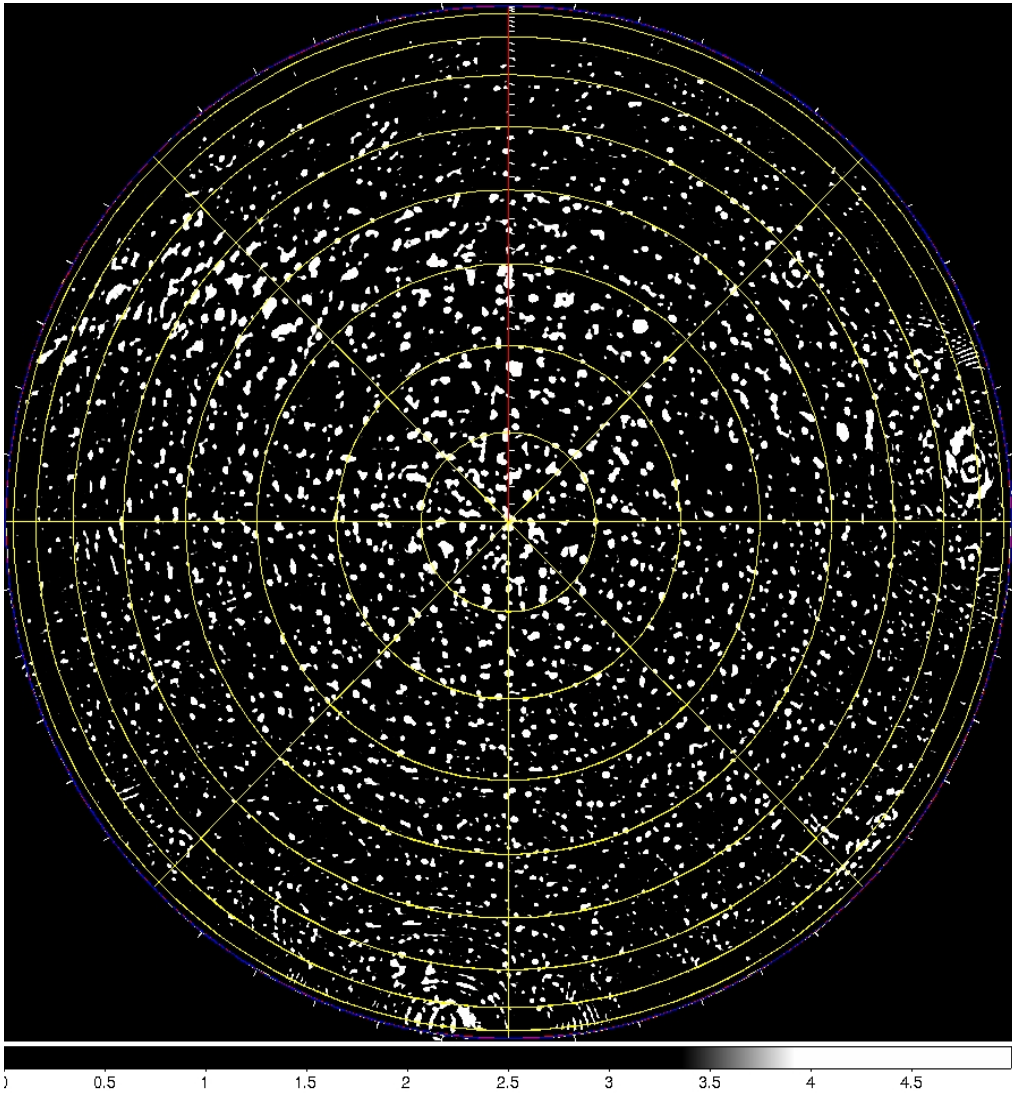
- Fourier transform (U,V) plane → (X,Y) image
- several algorithms being considered
  - special attention to GPU, Cell BE, etc.

An All-Sky Image

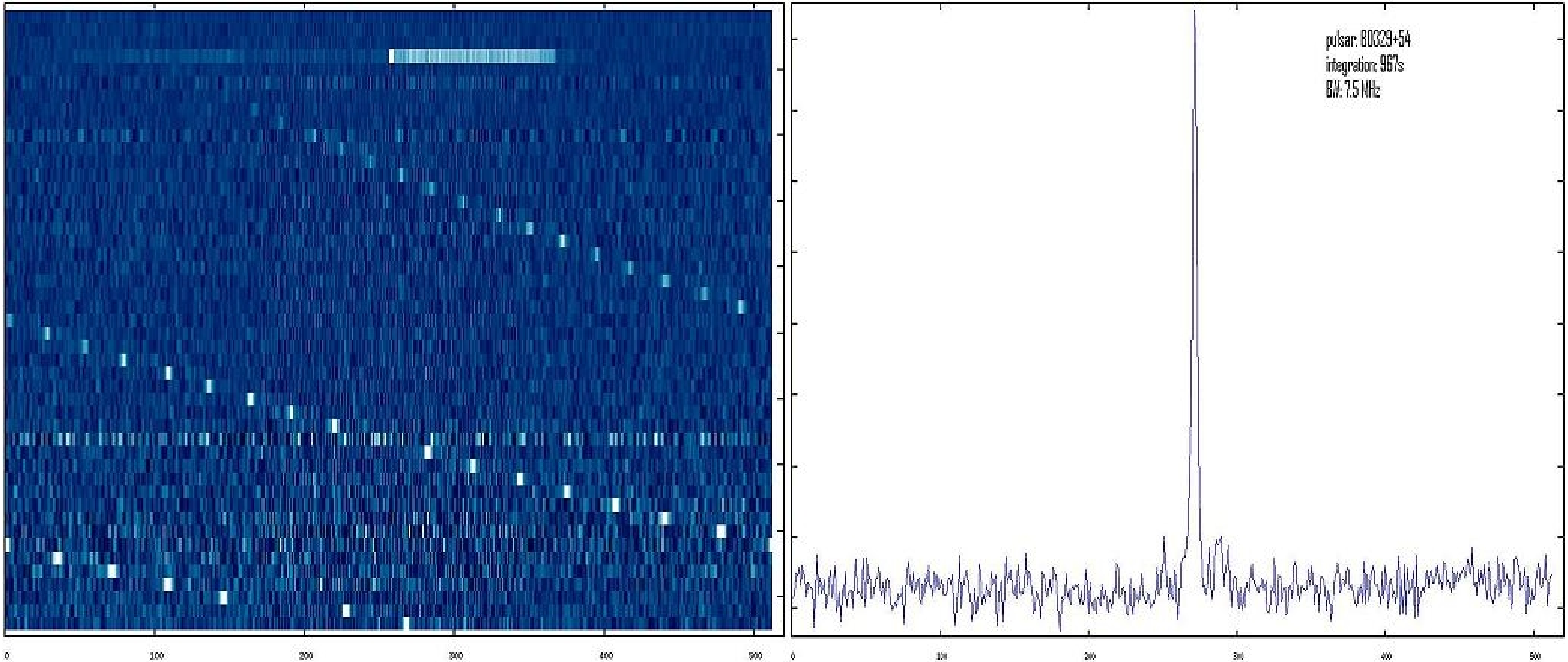


And Another One

- ~1,000 sources
- 1:20,000 dynamic range
  - resolution limited



Pulsar



pulsar: B0329+54
integration: 967s
BW: 7.5 MHz
0
100
200
300
400
500

Conclusions

- LOFAR promises interesting, new science
- Blue Gene/P:
  - very high computational performance
  - very high bandwidth
- bandwidth increase makes LOFAR 50% more efficient

Acknowledgments

ASTRON: Chris Broekema, Martin Gels, Jan David Mol, Rob van Nieuwpoort

ANL: Kamil Iskra, Kazutomo Yoshii

IBM: Bruce Elmegreen, Todd Inglett, Tom Liebsch, Andrew Taufener

References

- John W. Romein, P. Chris Broekema, Jan David Mol, and Rob V. van Nieuwpoort, Processing Real-Time LOFAR Telescope Data on a Blue Gene/P SuperComputer, Under review
- Kazutomo Yoshii, Kamil Iskra, P. Chris Broekema, H. Naik, and Pete Beckman, Characterizing the Performance of Big Memory on Blue Gene Linux, International Workshop on Parallel Programming Models and System Software for High-End Computing (P2S2'09), Vienna, Austria, September, 2009
- John W. Romein, FCNP: Fast I/O on the Blue Gene/P, Parallel and Distributed Processing Techniques and Applications (PDPTA'09), Las Vegas, NV, July, 2009
- Rob V. van Nieuwpoort and John W. Romein, Using Many-Core Hardware to Correlate Radio Astronomy Signals, ACM International Conference on SuperComputing (ICS'09), New York, NY, June, 2009
- Kamil Iskra, John W. Romein, Kazutomo Yoshii, and Pete Beckman, ZOID: I/O-Forwarding Infrastructure for Peta-Scale Architectures, ACM Symposium on Principles and Paradigms of Parallel Programming (PPoPP'08), Salt Lake City, NV, February, 2008
- John W. Romein, P. Chris Broekema, Ellen van Meijeren, Kjeld van der Schaaf, and Walther H. Zwart, Astronomical Real-Time Signal Processing on a Blue Gene/L SuperComputer, ACM Symposium on Parallel Algorithms and Architectures (SPAA'06), Cambridge, MA, July, 2006