LOFAR CEP Design & Performance

Chris Broekema

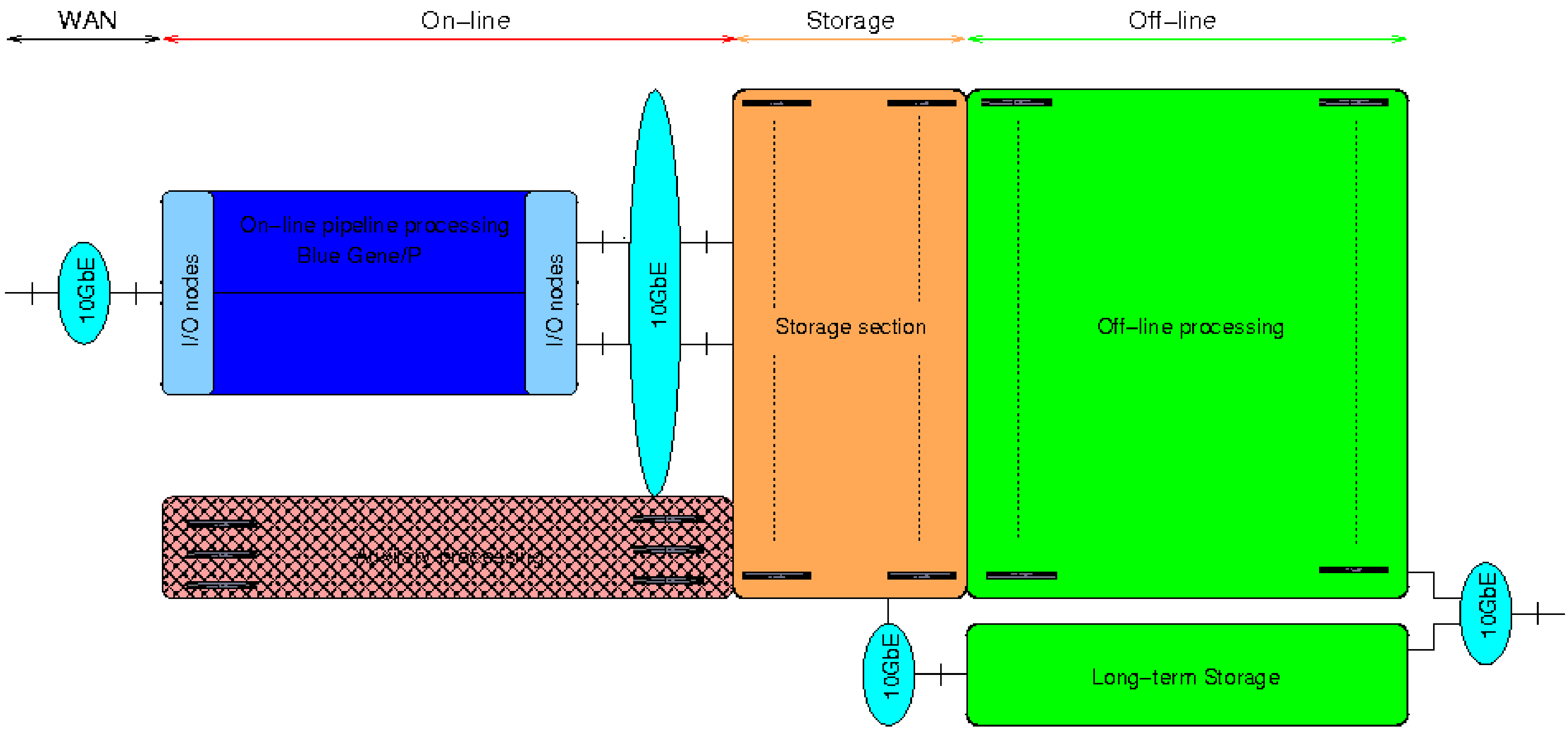
ASTRON

ASTRON LOFAR NWO

Outline

- The LOFAR Central Processor
  - Top level design
  - Current hardware
- Current status and recent results
  - Standard imaging mode
  - Tied-array beamforming (pulsar mode)
- The offline processor
  - Performance requirements
  - Design
- Summary

The LOFAR central processor

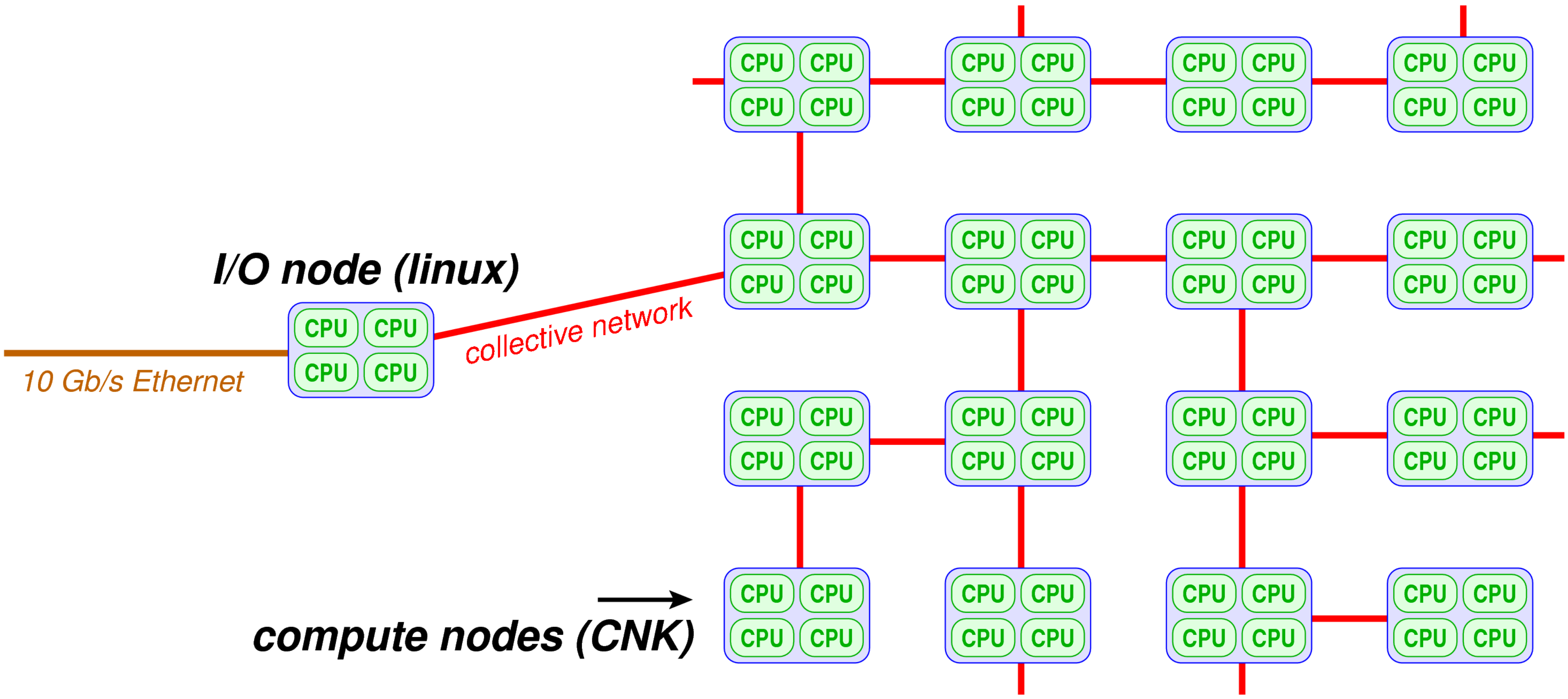


ASTRON LOFAR NWO

Central processor

- 3 rack IBM Blue Gene/P
  - #75 in the Top 500 (11-2008)
  - Peak performance 41.8 TFlops
    - (actually 44.4 TFlops, including I/O nodes)
  - 13056 PowerPC cores @ 850 MHz
    - Quad core system-on-chip CPUs
    - Double FPU
    - exceptional complex number support
  - ~6 TiB memory
  - 192 10GbE links
  - Several dedicated internal networks (torus, tree)

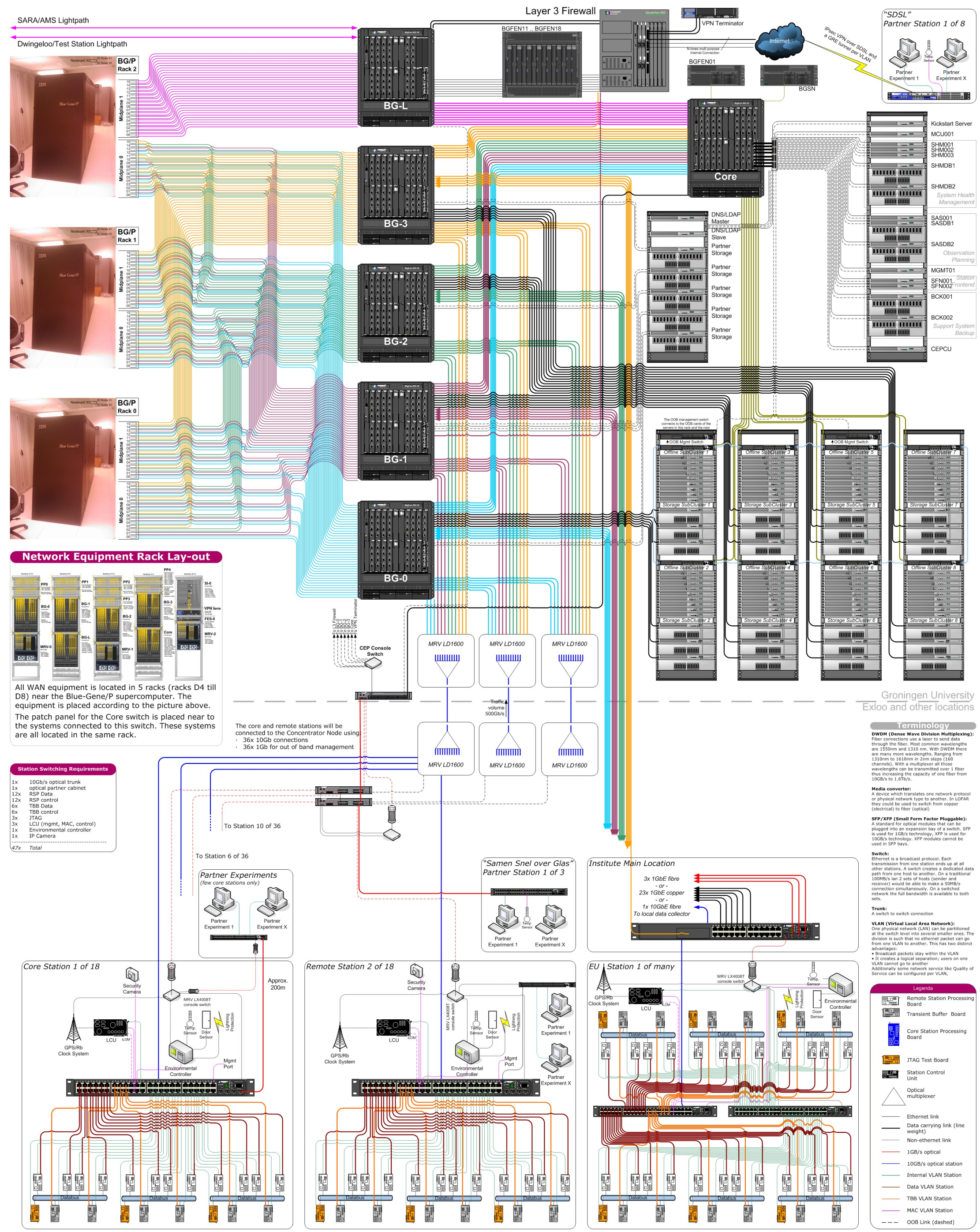
Blue Gene/P pset



Central processor

- 6 Foundry BigIron RX16 switching frames
  - 1 core, 4 leafs and 1 infrastructure
- ~350 10 GbE ports
  - 192 BG/P, ~70 stations, ~70 uplink, ~10 science
- ~300 GbE ports
- Dataflow optimized network design
  - keep dataflow within one switching frame
  - Bandwidth between frames limited (~50 Gbps)

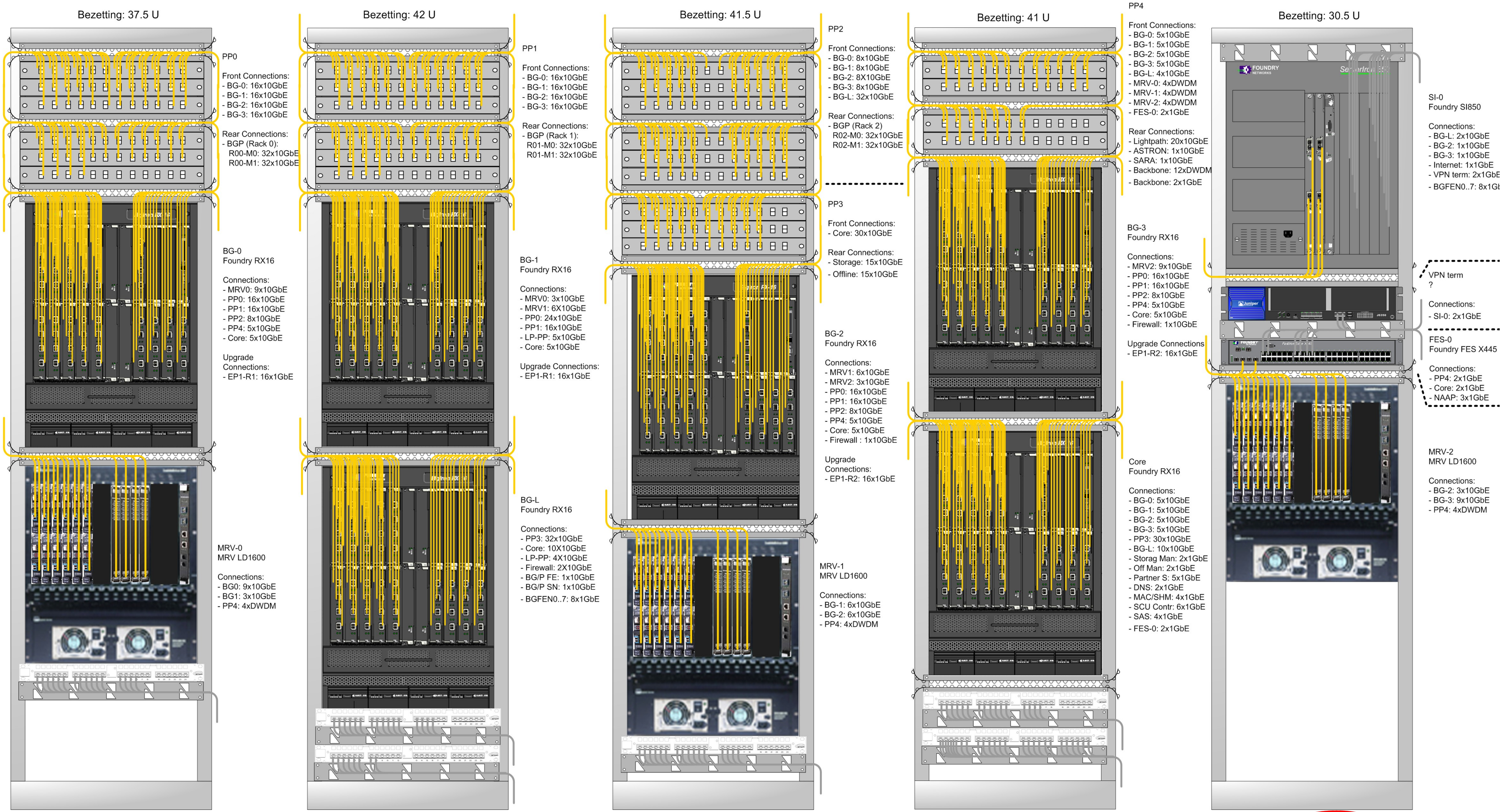
LOFAR
Network Topology



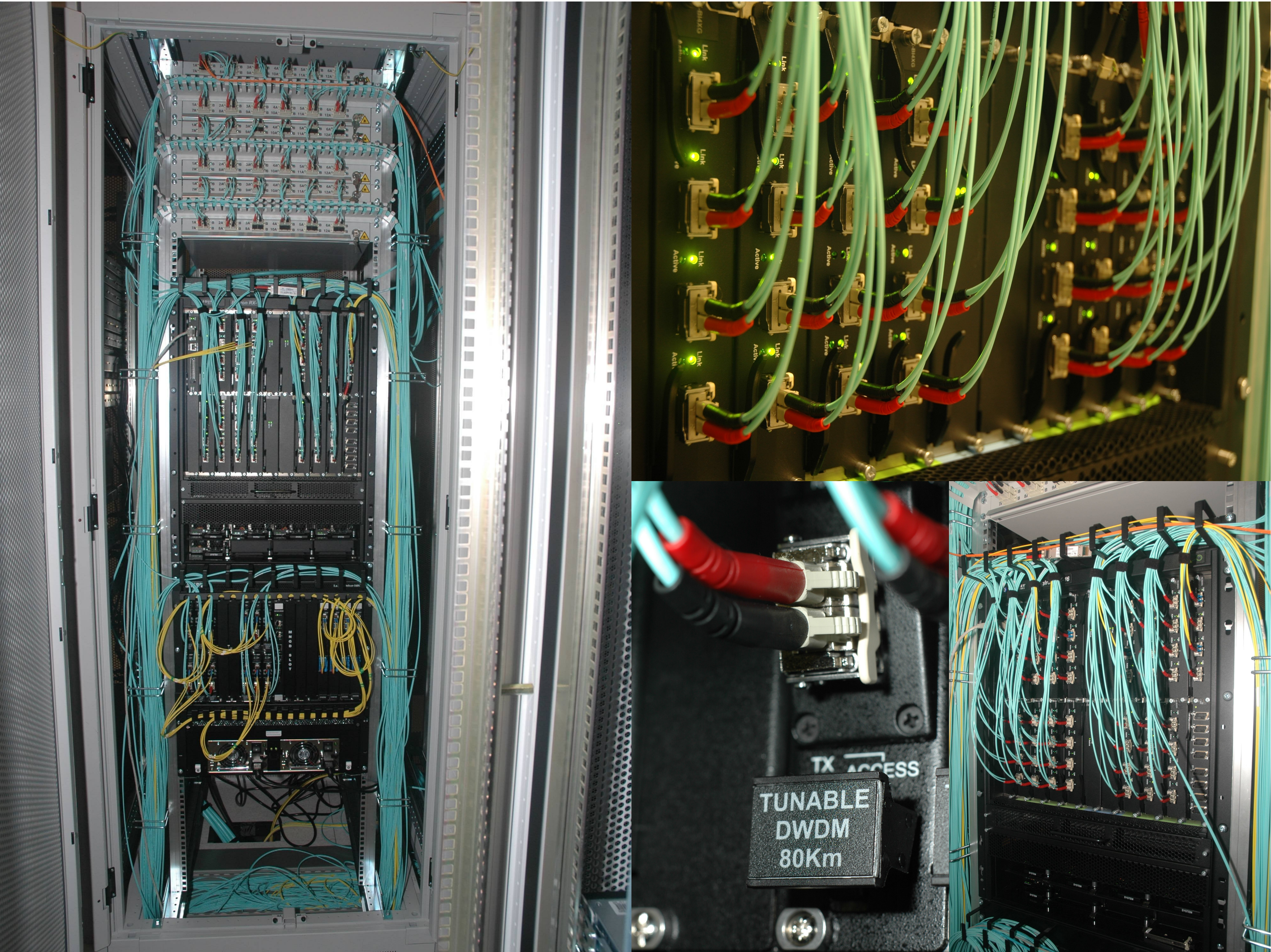
Version: October 30, 2008

Peter Maat <maat@astron.nl>
Klaas Stuurwold <stuurwold@astron.nl>
Roel Gloudemans <gloudemans@astron.nl>

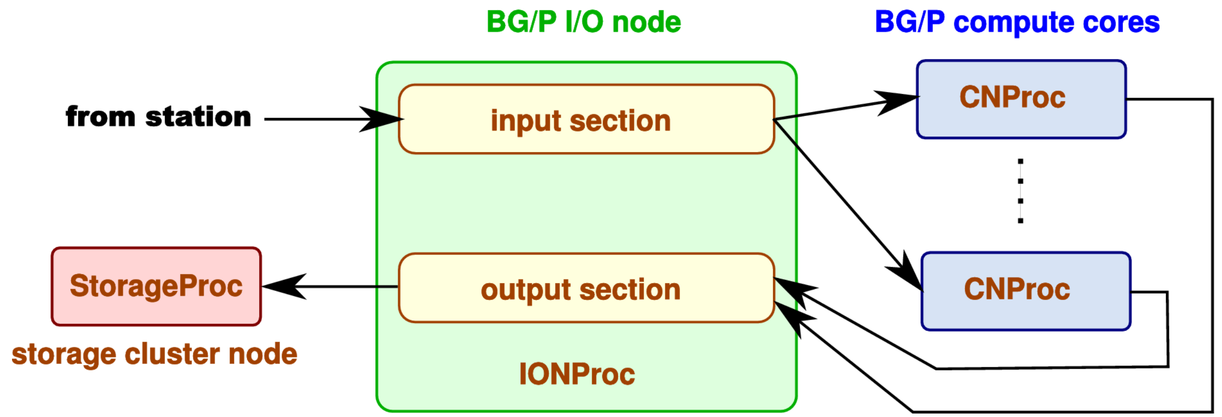
Network Equipment Rack Lay-out



ASTRON LOFAR NWO

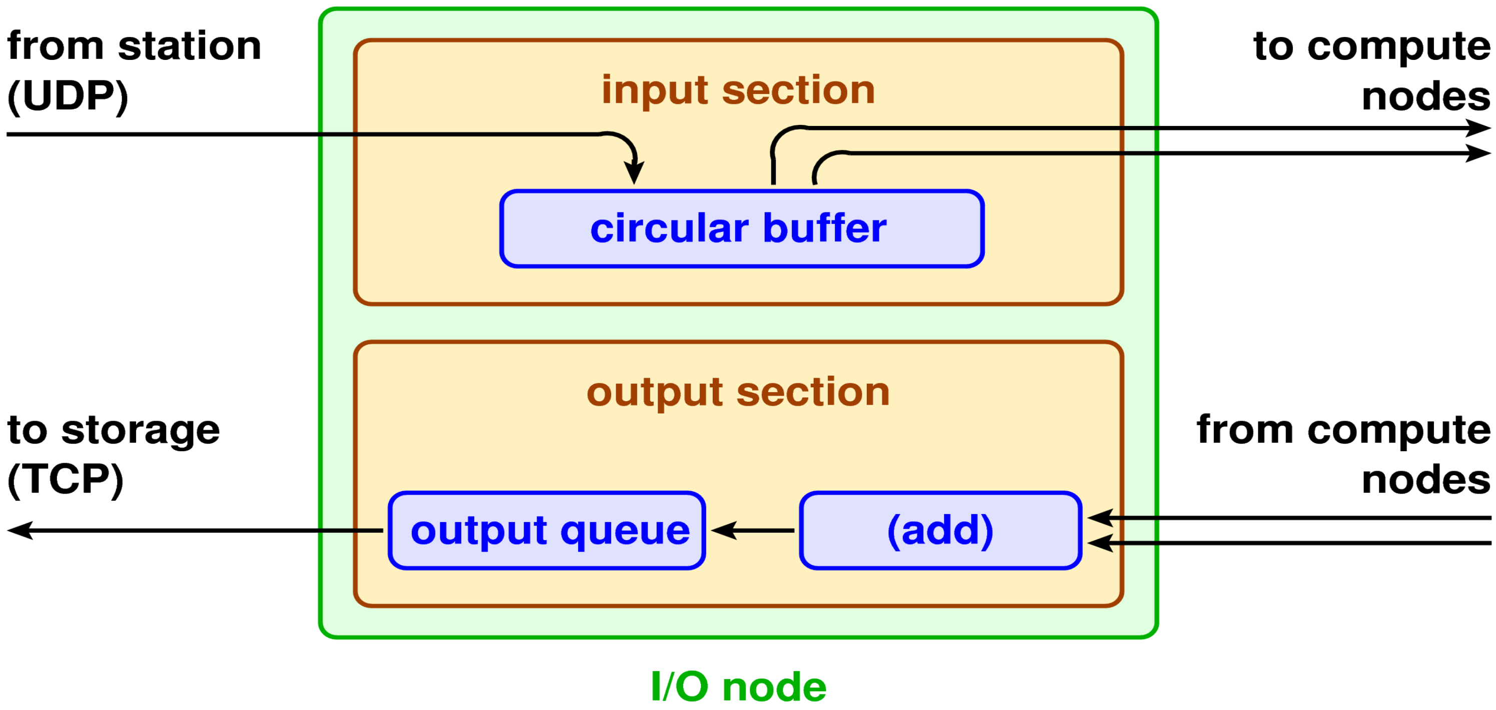

Link
Active
TX ACCESS
TUNABLE
DWDM
80Km

The Blue Gene/P Correlator



- three distributed applications/platforms
  - BG/P I/O nodes
  - BG/P compute nodes
  - external storage nodes

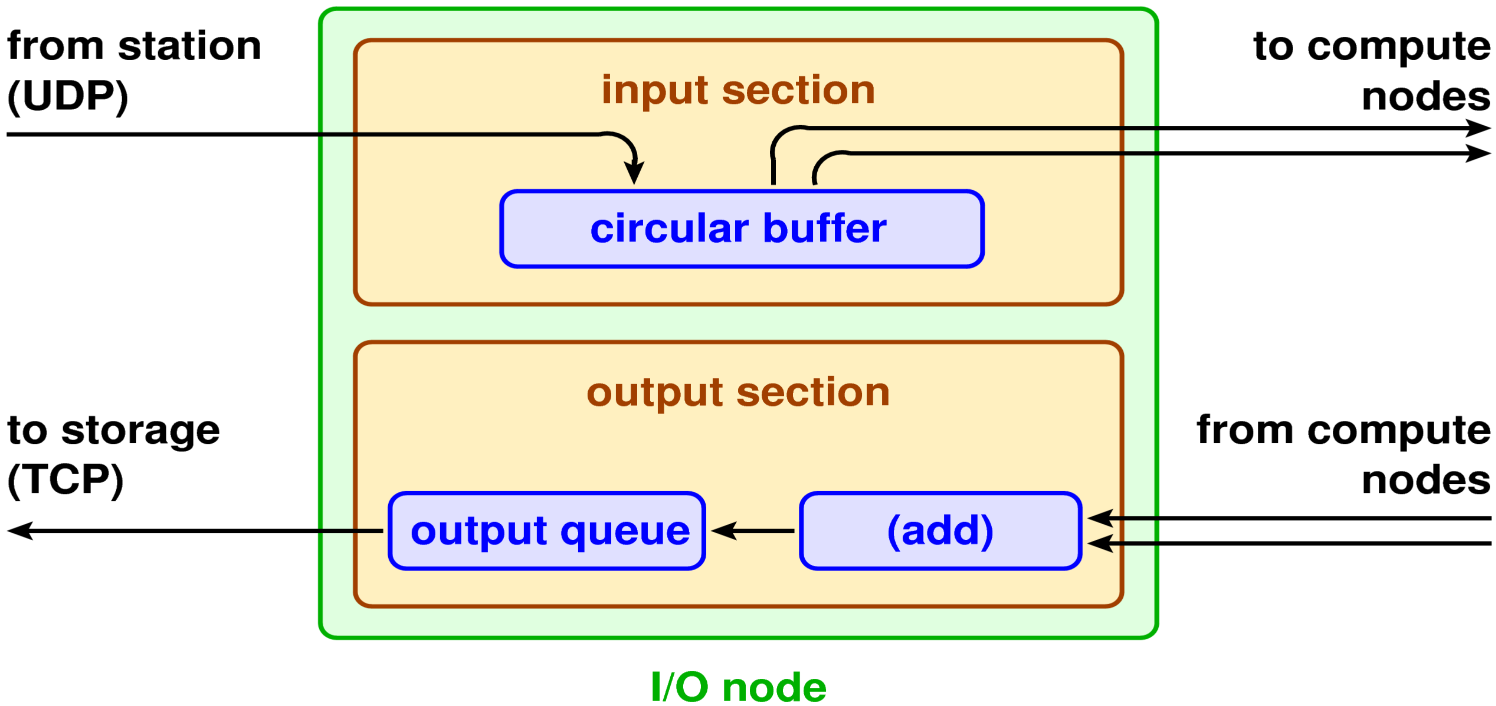
I/O node processing



- application on I/O nodes
  - more efficient & flexible
  - BG/L: saved costs for input cluster
  - BG/L: major system software changes (ZOID) [PPoPP'08]

ASTRON LOFAR NWO

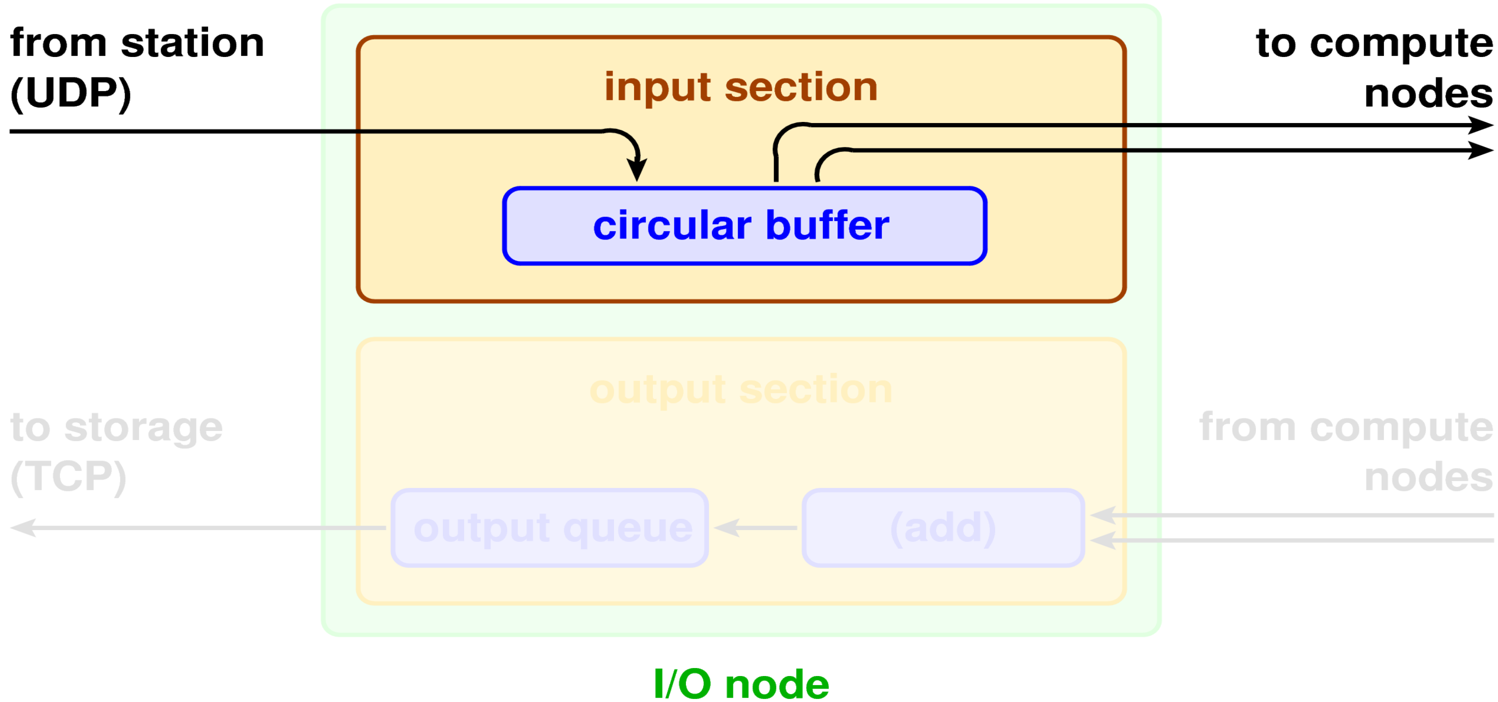
I/O node processing



- Two sections
  - Input section
  - Output section
- Heavily threaded & optimized

ASTRON LOFAR NWO

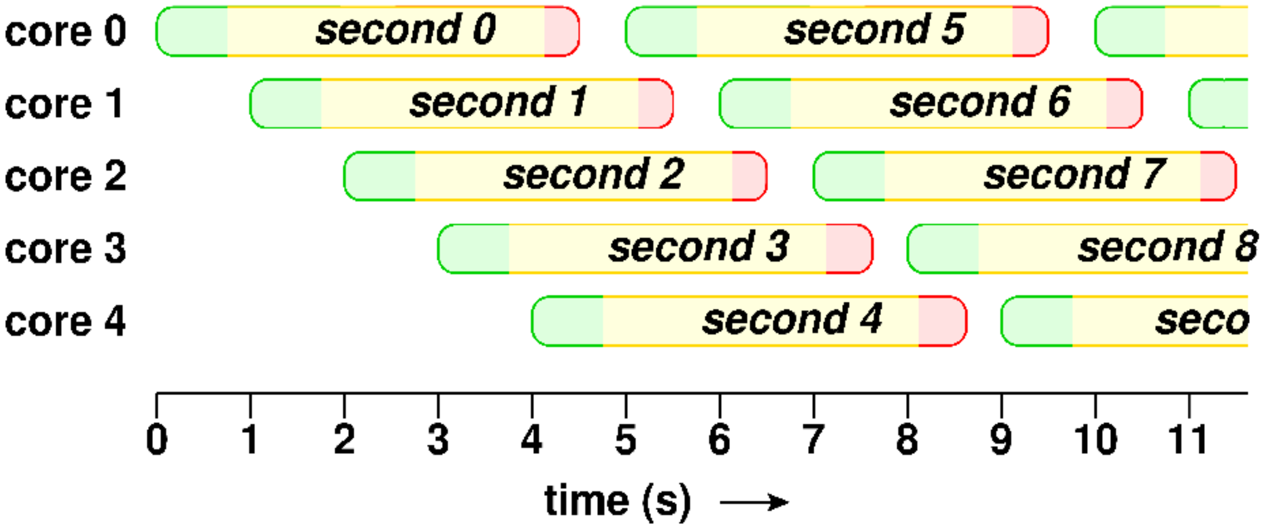
I/O node input section



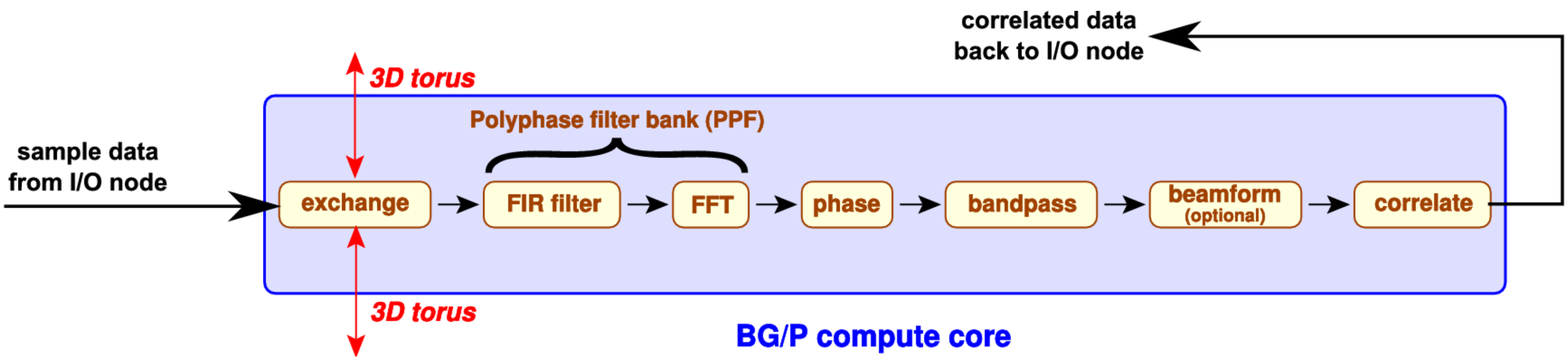
- one station per I/O node
- 48,828 pkt/s
- circular buffer (~2.5 s):
  - WAN delays
  - set observation direction
  - handle hiccups
- handles missing data
- wall-clock trigger

Work distribution

- *O*(100) independent data chunks
  - 1 second, 1 subband, all stations
  - needs > 1 second processing time
- distribute round robin over cores
  - receive, process, send, idle

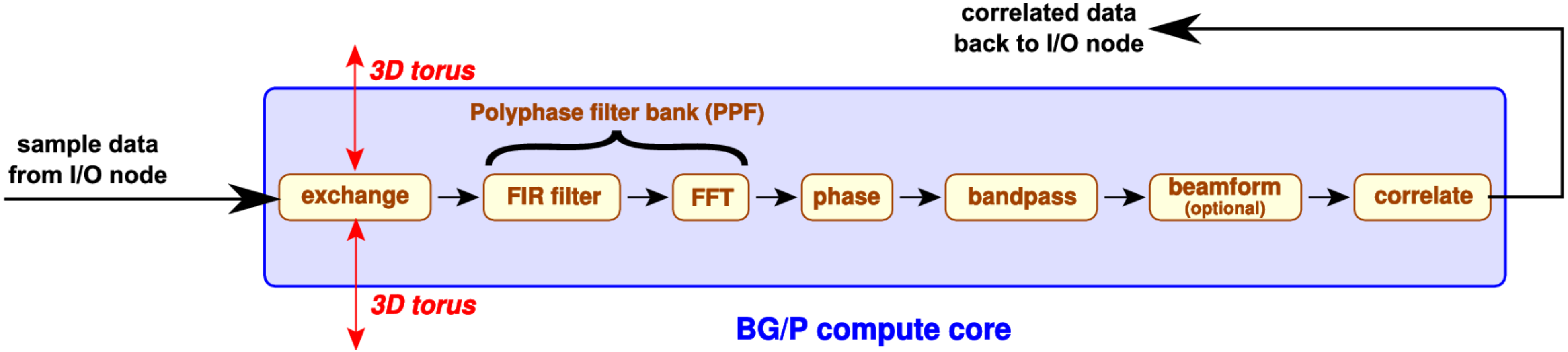


Compute node processing (1)



- Exchange (transpose)
  - All subbands; 1 station → all stations 1 subband
  - asynchronous
- Polyphase filter creates channels
- Phase correction to point accurately

Compute node processing (2)

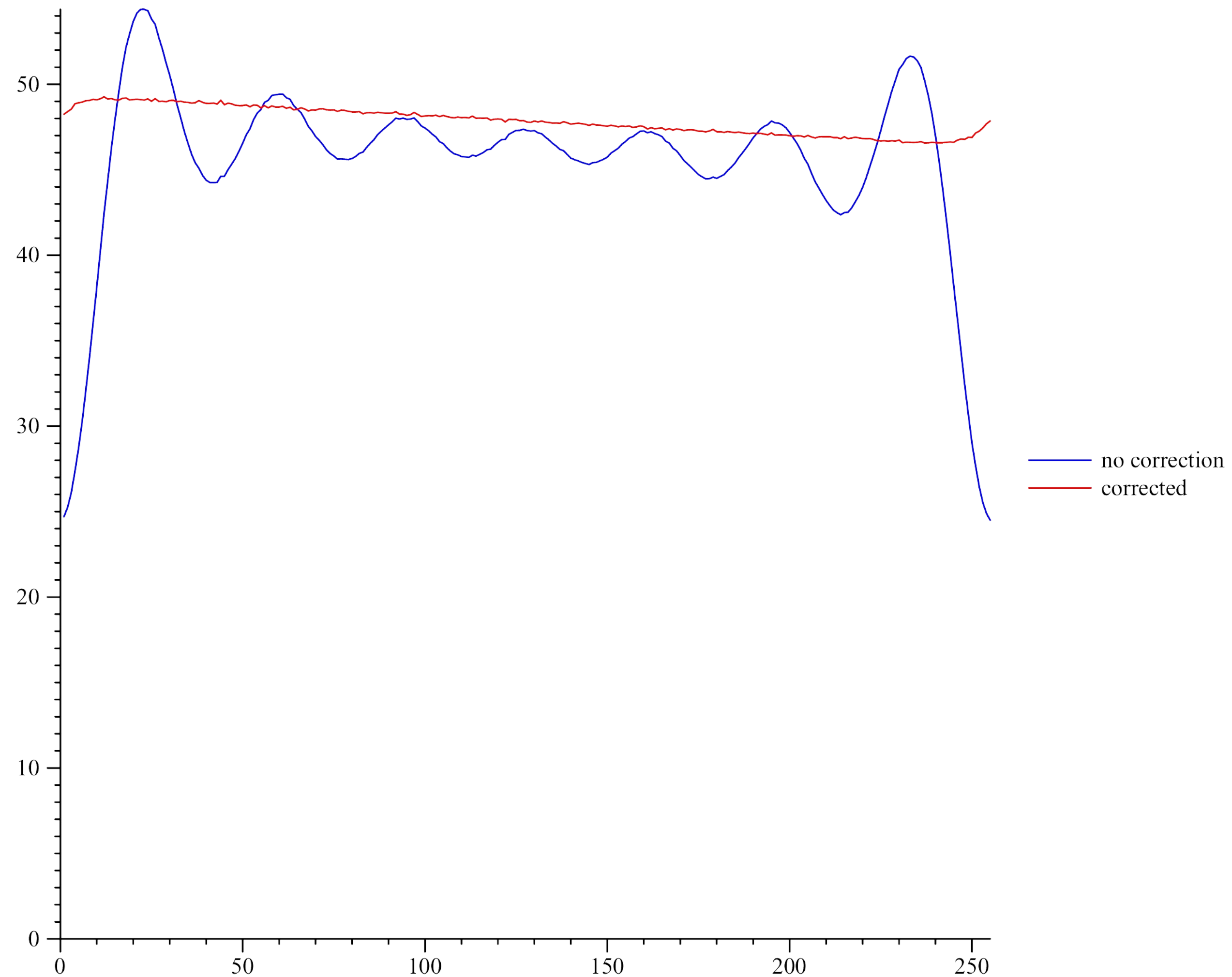


- Correct station-introduced bandbass
- Beam form (add) to create “Super Station” (optional)
- Correlate station samples pair-wise

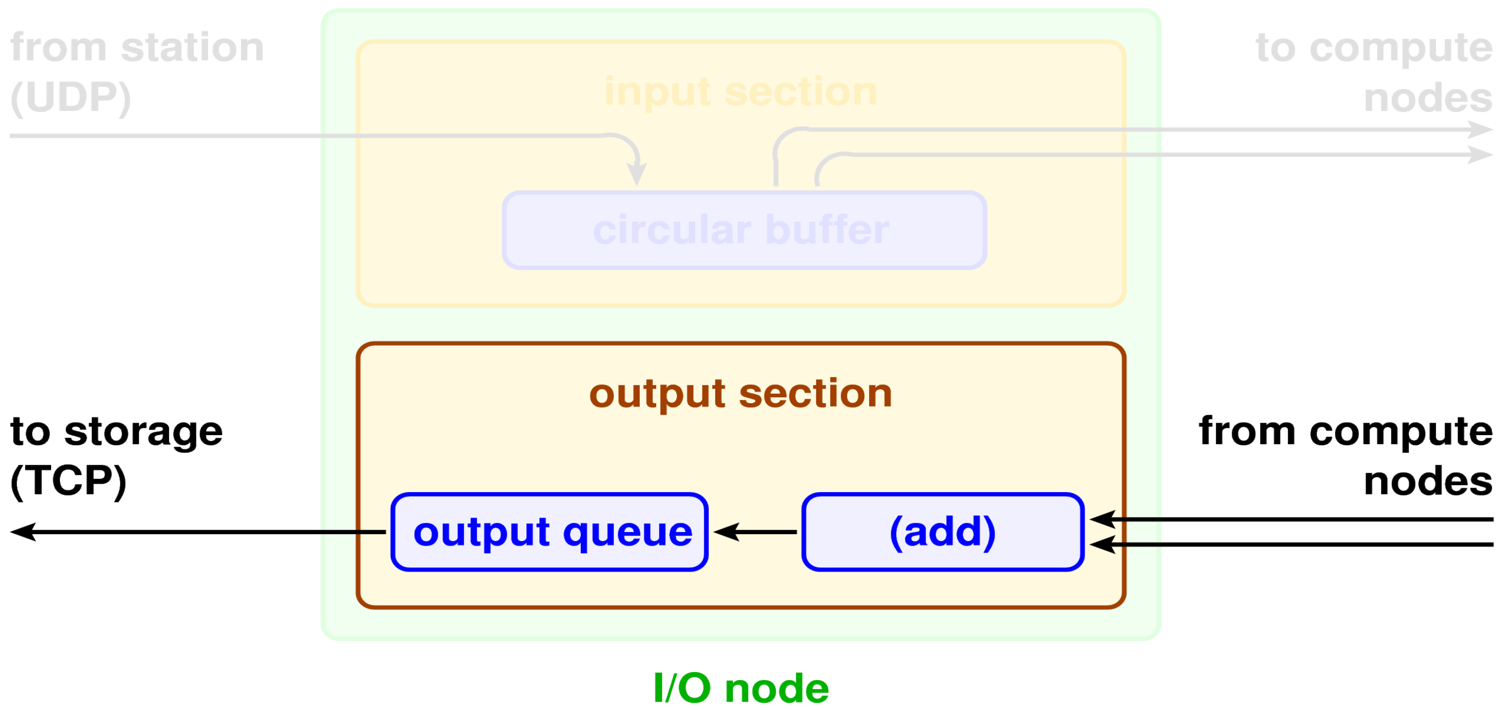
ASTRON LOFAR NWO

Bandbass correction

- 2 single dipoles
- 58.6 MHz
- 30 minute integration

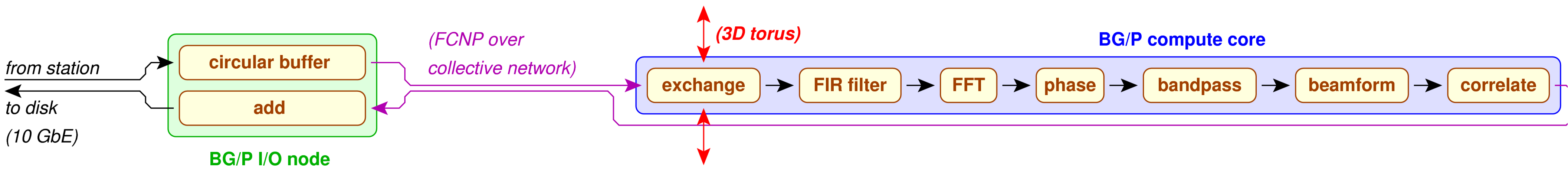


I/O node output section



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

Std imaging mode performance



- 1 rack BG/P used as correlator
- 1 rack BG/P generates simulated station data
  - Up to 64 stations @ 3.1 Gbps each
- ½ rack BG/P receives (and dumps) visibilities

ASTRON LOFAR NWO

Std imaging mode performance

observation mode	A	B	C
#stations	64	64	48
#subbands	248	496	992
#bits/sample	16	8	4
obs. bandwidth (MHz * #beams)	48.4	96.9	194
input bandwidth (Gb/s)	64 * 3.1	64 * 3.1	48 * 3.1
output bandwidth (Gb/s)	62 * 0.58	62 * 1.2	62 * 1.3
CPU load compute nodes	35%	70%	85%
CPU load I/O nodes	67%	81%	80%
data loss	~ .0001%	~ 0.01%	~ 0.01%

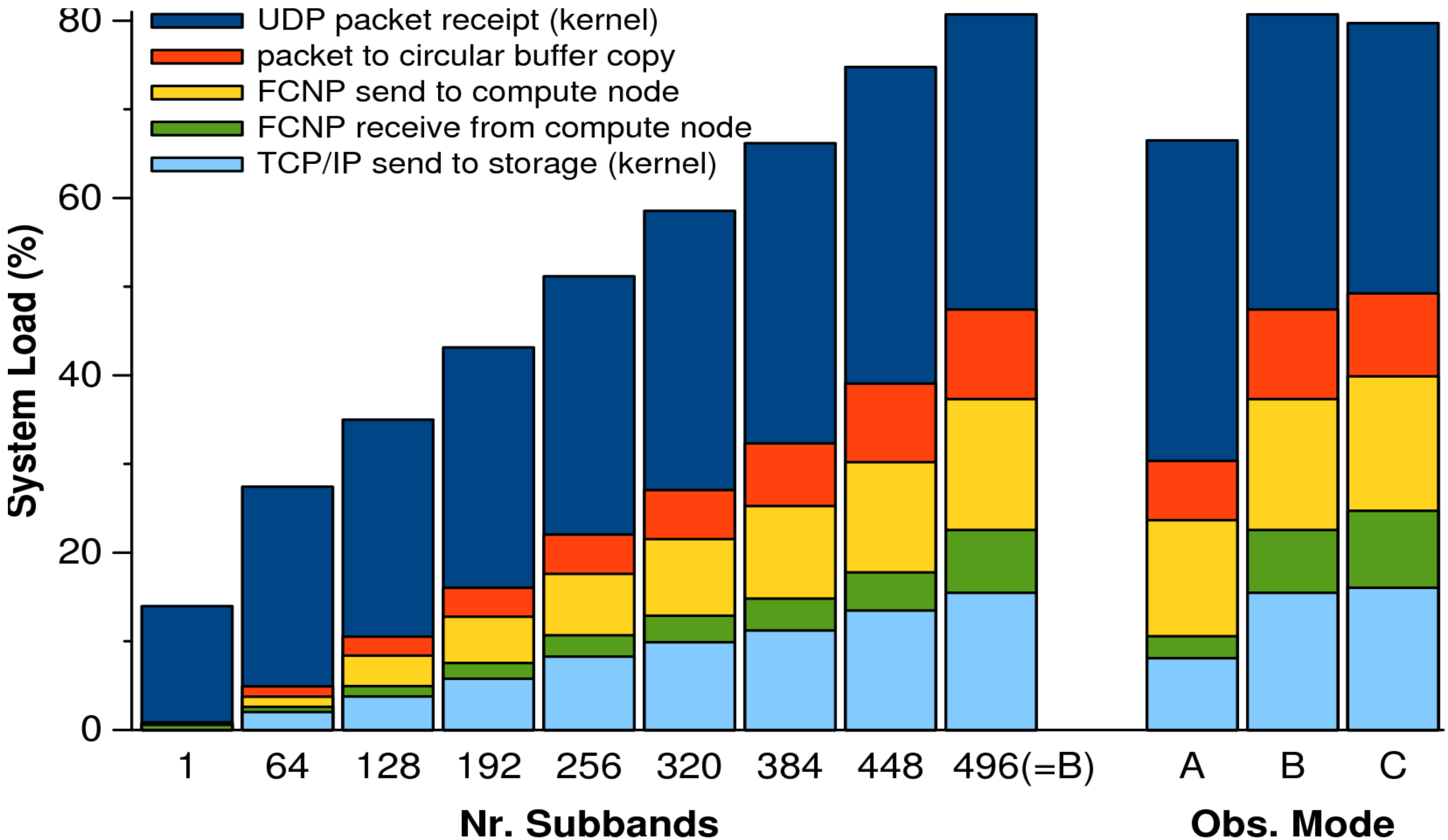
ASTRON LOFAR NWO

Std imaging mode performance

- This is representative for full LOFAR
  - Up to 64 stations
- In two new observation modes (8 bit & 4 bit)
- At 150% of the specified bandwidth
- With half the designed resources
- Without significant data loss
- EoR mode can be done on 1 rack BG/P
  - (6 Racks BG/L originally)

Std imaging mode performance

Blue Gene/P I/O node load

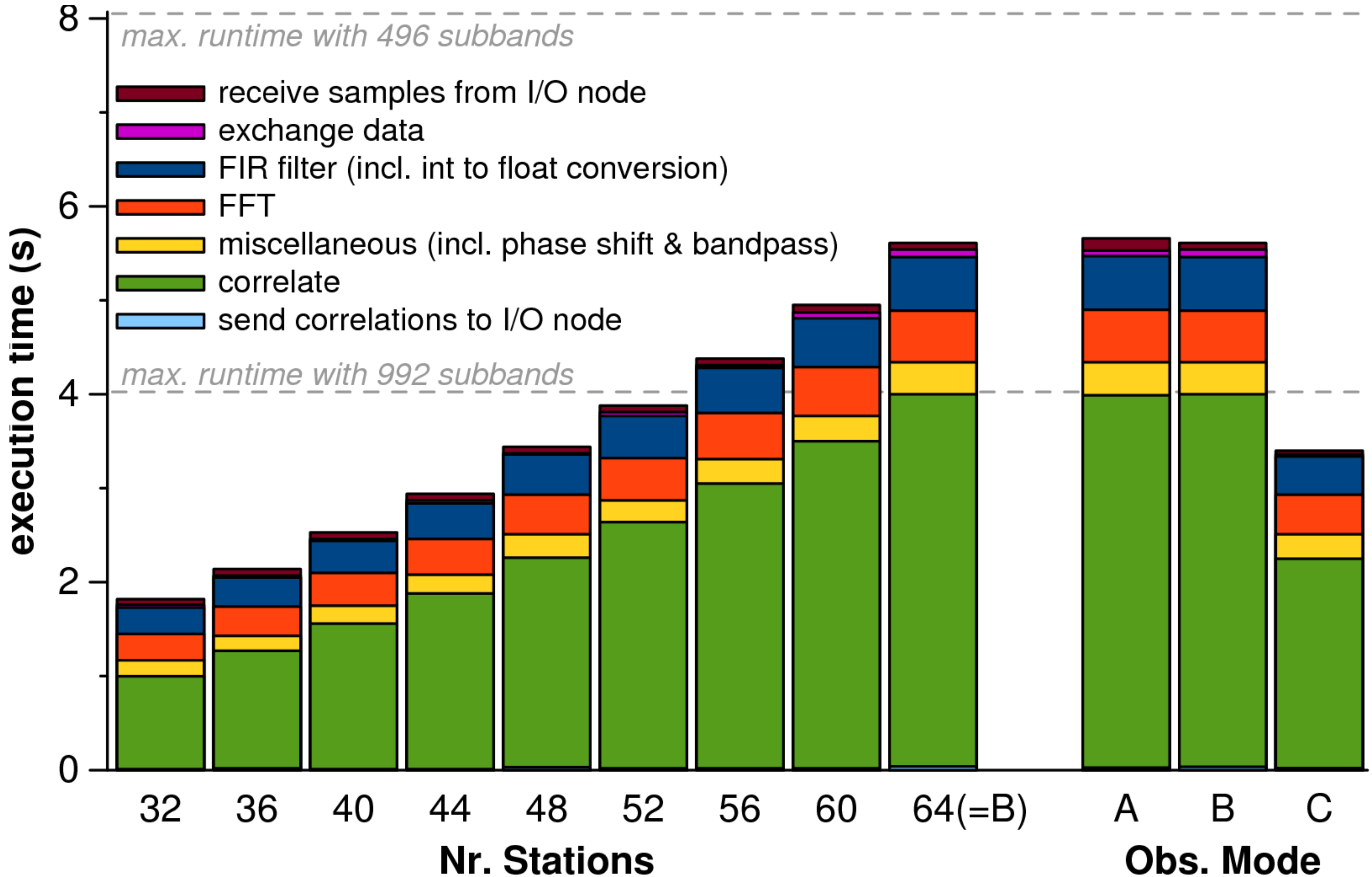


Std imaging mode performance
Blue Gene/P I/O optimizations

- Heavily modified I/O node Linux kernel
  - Avoid TLB misses
  - Optimize network stack buffer sizes
- Low overhead protocol to Compute nodes
- Optimum scheduling of threads in application
  - Use Linux real-time threads
- Use of assembler where appropriate

Std imaging mode performance

Blue Gene/P Compute node load



ASTRON LOFAR NWO

Std imaging mode performance
Blue Gene/P Compute node optimizations

- Heavy use of assembler in hot spots
  - Correlator (96% of peak FPU performance)
  - FIR filter (86% of peak FPU performance)
  - FFT (43% of peak FPU performance)
- Rewrite transpose to use DMA engine
  - Uses asynchronous send/recv instead of MPI_Alltoallv()
  - Hides transpose time completely
- Low overhead protocol to (and from) I/O nodes

ASTRON LOFAR NWO

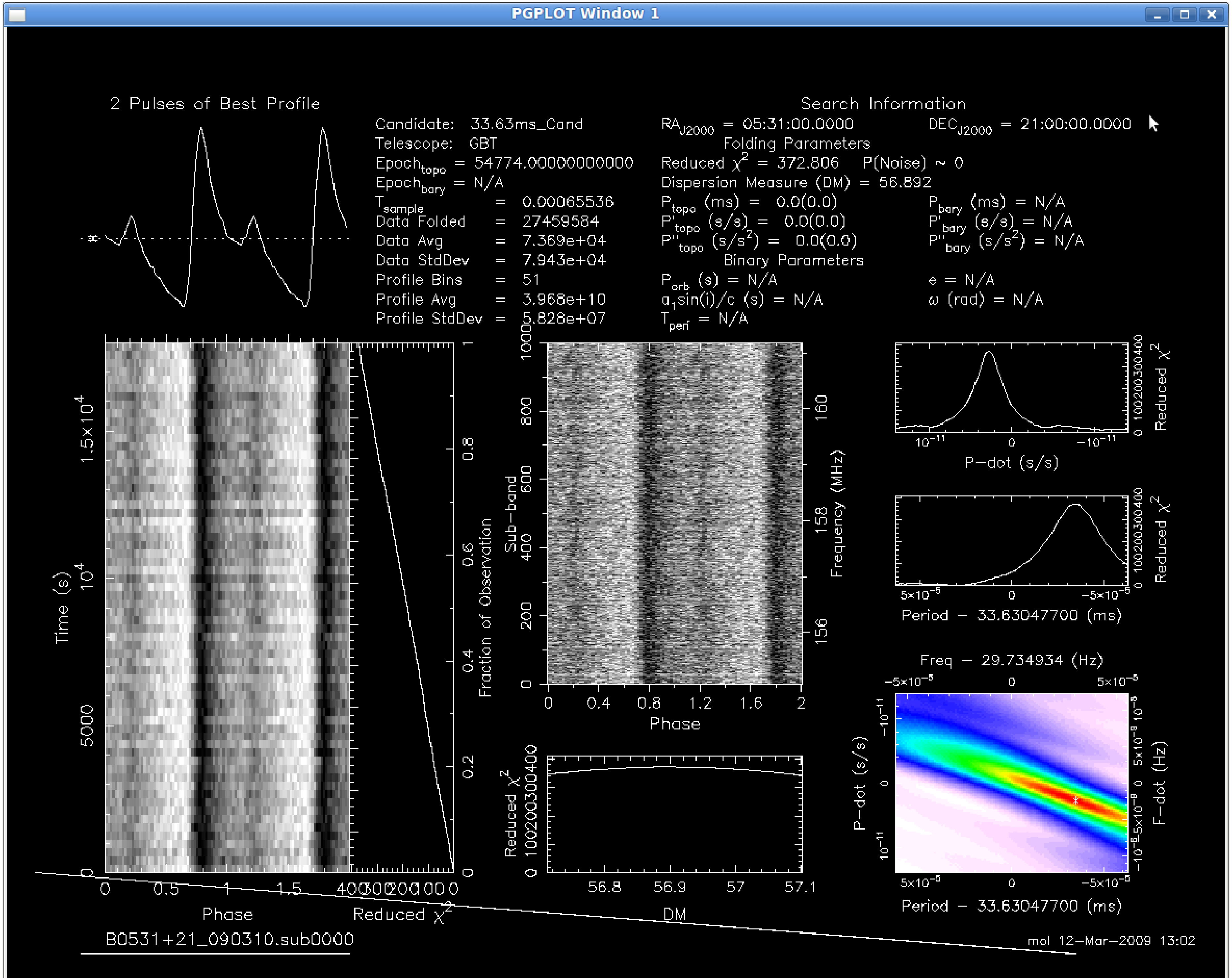
Tied-Array beamforming

- Reference implementation available
  - Real-time
  - Capable of creating multiple close beams
- Complex voltages
- Stokes I
- Stokes I, Q, U, V
- Incoherent
- Online integration over time

ASTRON LOFAR NWO

Tied-Array beamforming

- Runs correctly and stable
  - Several successful pulsar observations done
  - Multiple pencil beams, up to 64 hours
- Still way too slow
  - Can only form a few beams in real-time
  - Reference C++ implementation written for clarity
  - Optimized (assembler) version available
    - unverified, not finished
    - But observed to run ~30 times faster


PGPLOT Window 1
2 Pulses of Best Profile
Candidate: 33.63ms_Cand
Telescope: GBT
Epoch_topo = 54774.00000000000
Epoch_bary = N/A
T_sample = 0.00065536
Data Folded = 27459584
Data Avg = 7.369e+04
Data StdDev = 7.943e+04
Profile Bins = 51
Profile Avg = 3.968e+10
Profile StdDev = 5.828e+07
Search Information
RA_J2000 = 05:31:00.0000
DEC_J2000 = 21:00:00.0000
Folding Parameters
Reduced χ2 = 372.806 P(Noise) ~ 0
Dispersion Measure (DM) = 56.892
P_topo (ms) = 0.0(0.0)
P_bary (ms) = N/A
P'_topo (s/s) = 0.0(0.0)
P'_bary (s/s) = N/A
P''_topo (s/s2) = 0.0(0.0)
P''_bary (s/s2) = N/A
Binary Parameters
P_orb (s) = N/A
e = N/A
a1sin(i)/c (s) = N/A
ω (rad) = N/A
T_peri = N/A
Time (s)
Phase
Reduced χ2
Fraction of Observation
Sub-band
Frequency (MHz)
P-dot (s/s)
Period − 33.63047700 (ms)
Freq − 29.734934 (Hz)
F-dot (Hz)
DM
B0531+21_090310.sub0000
mol 12-Mar-2009 13:02

The offline processor Phase 1 specification

- Temporary storage
  - ~500 TB
  - ~15 Gbps input (continuous)
  - ~30 Gbps output (burst)
- Compute cluster
  - Flagging, calibration, imaging, source finding
  - ~5 TFlops
  - Needs to keep up with the correlator
    (although not necessarily in real-time)

Phase 1 hardware (1)

- 24 storage nodes
  - 2 Quad-core low-power Intel Xeon CPUs
  - 16 GiB main memory
  - 24 x 1TB disks each → ~20 TB usable capacity
  - 4 GbE interfaces
- 72 compute nodes
  - 2 Quad-core low-power Intel Xeon CPUs
  - 16 GiB main memory
  - 1 TB local storage (2x 500 GB in RAID-0)
  - 2 GbE interfaces

Phase 1 hardware (2)

- 8 GbE data switches
  - One for each sub-cluster
  - 20 Gbps uplink to Core infrastructure
- 2 frontend nodes
  - 2 Quad-core low-power Intel Xeon CPUs
  - 16 GiB main memory
  - ~2 TB storage capacity in RAID-5

Bandwidth optimized sub-clusters

- Offline cluster does mostly batch processing
- Inter node communication is limited
- Huge data volumes
  - Communication needs to be limited
  - Necessary communication needs to be optimized
  - Cache locally to avoid unnecessary transport
- Divide cluster resources into 8 sub-clusters with optimum connectivity

Subcluster configuration (INTERNAL USE ONLY)

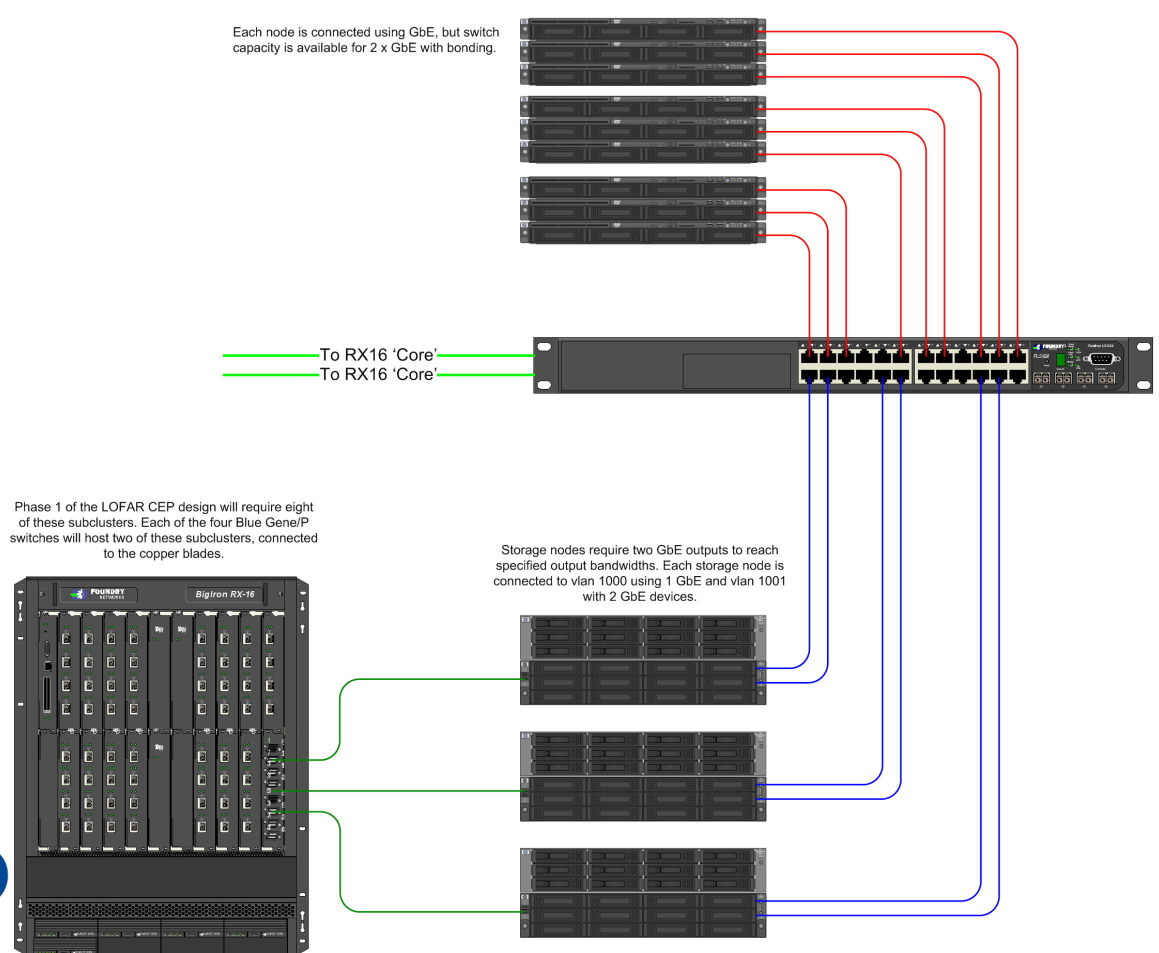
Legend

- 10 GbE uplink (green)
- GbE BG/P to storage (vlan 1000) (dark green)
- GbE storage out (vlan 1001) (blue)
- GbE offline cluster (vlan 1001) (red)

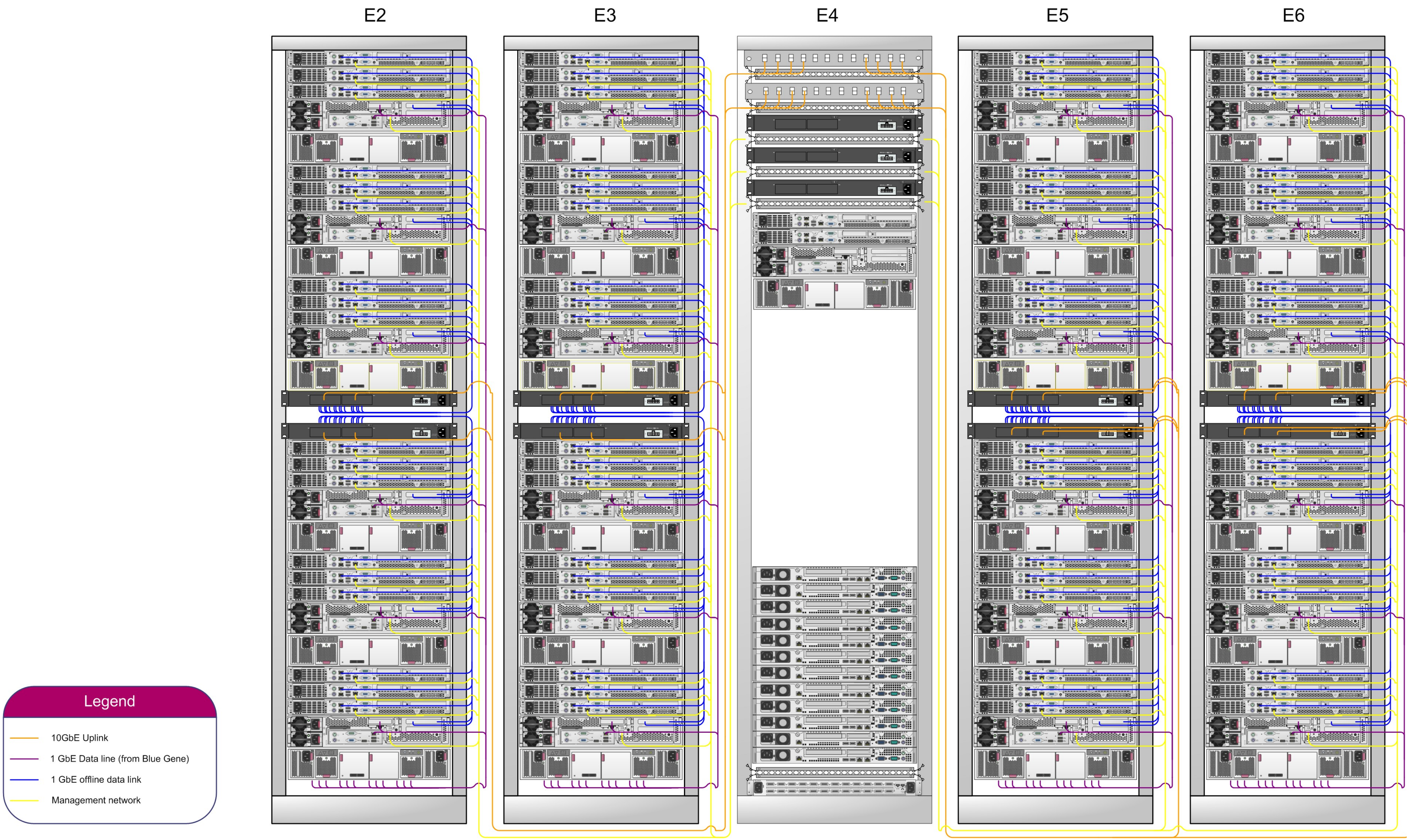
Phase 1 LOFAR CEP Hardware:

8 24 port switches with 2 10 GbE uplinks each
24 Storage nodes, each with ~24 TB disks
72 Computational nodes

Total storage capacity:	~480 TB
Total input bandwidth:	~24 Gbps
Total output bandwidth:	~48 Gbps



LOFAR Phase 1 cluster
Rear view



Phase 1 hardware

- Delivery scheduled this week
- Installation until beginning of June
- Commissioning as soon as possible
    - Operating system, infrastructure & applications
    - Staggered roll-out per subcluster
    - Subclusters may be temporarily reassigned
    - Currently available cluster OS migration

Phase 2 hardware

- Q4 2009 – Q1 2010
- Storage component grows to 2 PB
  - Input b/w ~50 Gbps (sustained)
  - Output b/w ~100 Gbps (burst)
- Offline processing cluster
  - At least ~10 TFlops
  - May not be enough

ASTRON LOFAR NWO

Summary & Conclusions (1)

- The LOFAR central processor is ready
  - We can handle full LOFAR in std imaging mode
  - At 150% of designed bandwidth
  - Tied-array beamforming is coming along nicely
- The phase 1 offline processor to be built shortly
- Phase 2 specifications to be defined next
  - Using LOFAR-20 experiences
  - Probably dominated by calibration

Summary & Conclusions (2)

- Getting to this point required specialists
  - Linux kernel hacking on BG/P I/O nodes
  - Assembler kernels for computational hotspots
  - Detailed hardware design optimized for application
- Computation cannot be separated from I/O
  - Network → node
  - Memory → cache or CPU
  - Cache → CPU
  - Many-core architectures complicate this problem

Acknowledgements

John W. Romein

Jan David Mol

Rob van Nieuwpoort

Argonne National Lab:

Kazutomo Yoshii

Kamil Iskra

IBM:

Bruce Elmegreen

Todd Inglet

Tom Leibsch

Andrew Taufener

ASTRON

LOFAR

NWO

Relevant publications

- 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