ASTRON

Netherlands Institute for Radio Astronomy

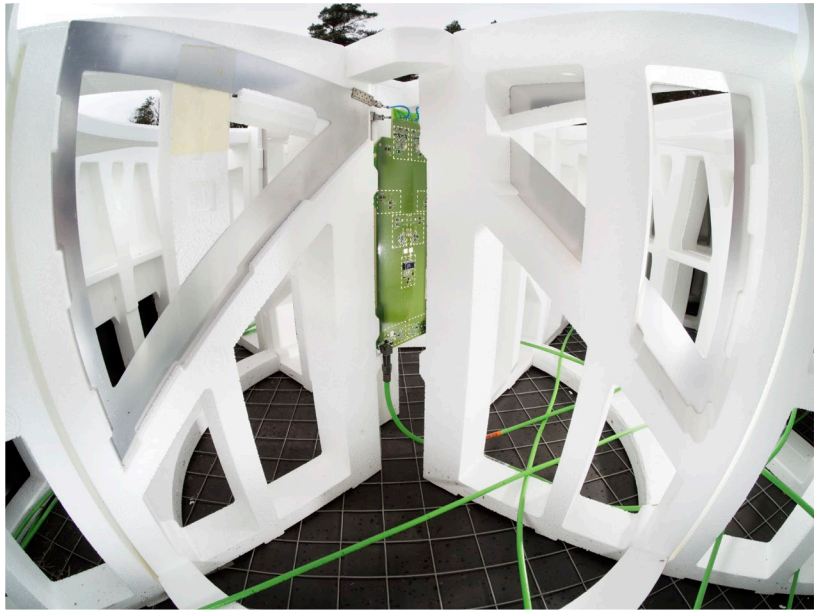
LOFAR: A decade of correlating in a software telescope

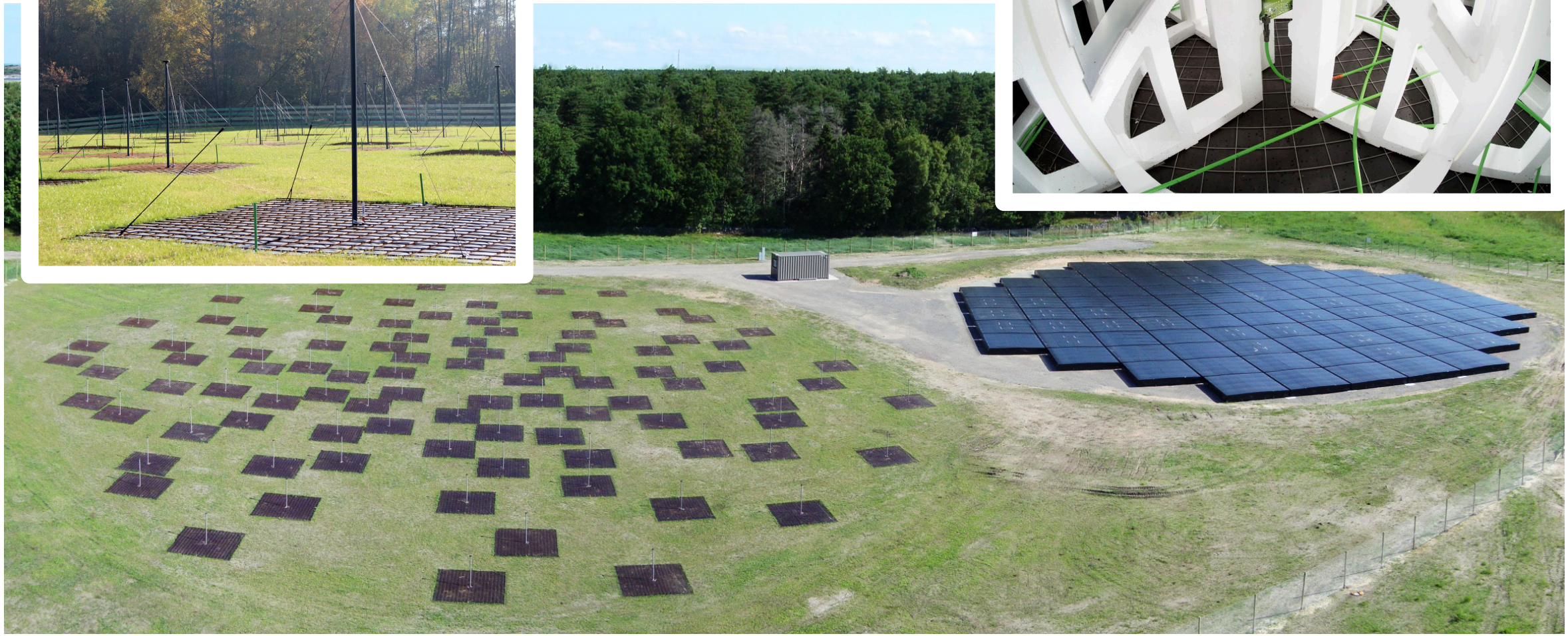
Jan David Mol,

The 7th Int'l VLBI Technology Workshop, Krabi, Thailand, 2018

ASTRON
Netherlands Institute for Radio Astronomy

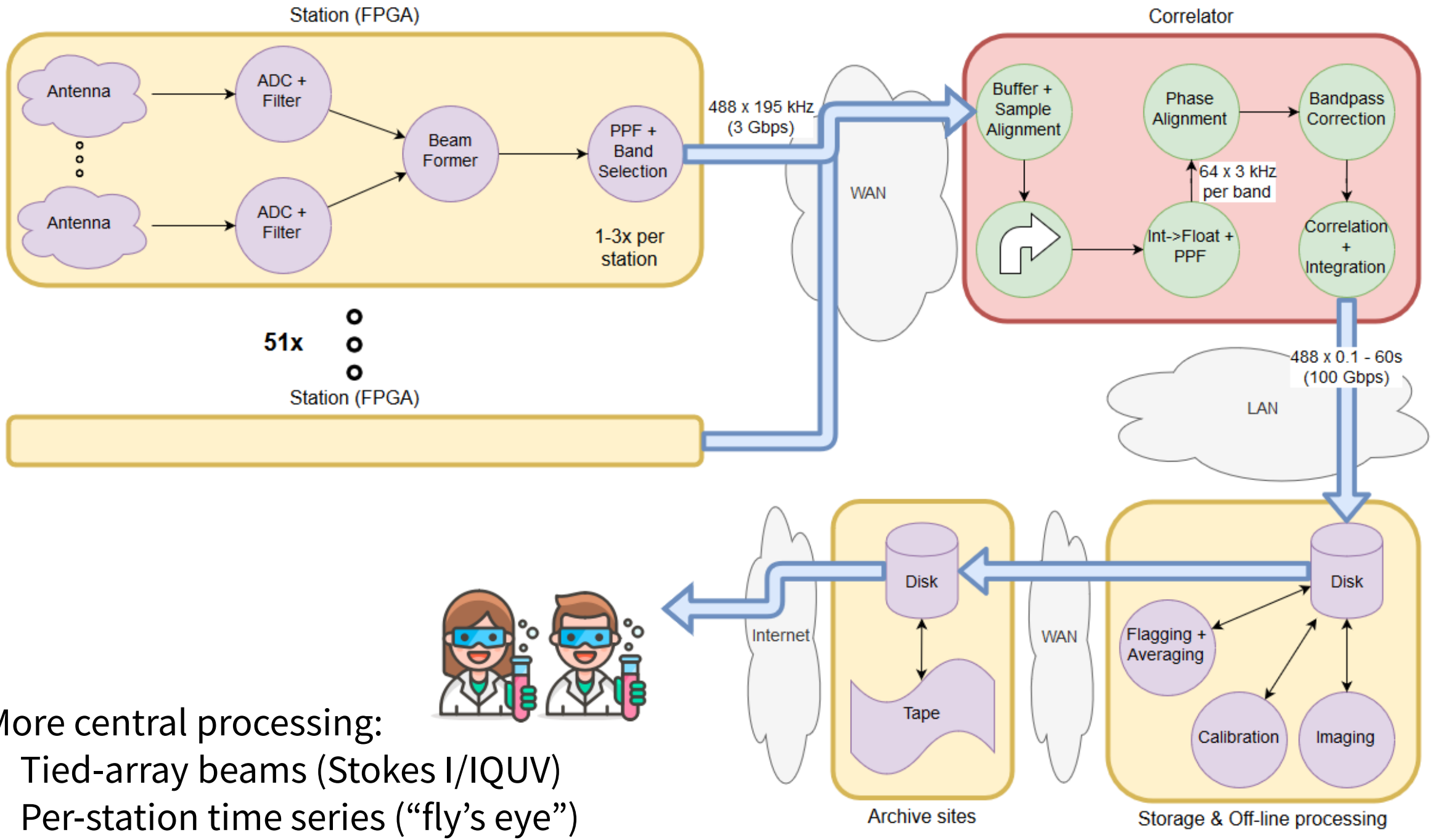
- 48 - 96 antennas, 2 pol
- 10 – 240 MHz
- 200 MHz 8-bit complex
- Beamformed on FPGA






2011 © AEROPHOTO EELDE

- Baselines:
  - ≤ 2000 km
  - 38 Dutch stations, 13 Intl
- FoV:
  - 300 deg^2 @ 15 MHz
  - 1.3 deg^2 @ 240 MHz
- Resolution:
  - ≤ 3.3 arcsec @ 15 MHz
  - ≤ 0.2 arcsec @ 240 MHz
- Clock:
  - Shared @ 24 stations
  - GPS + Rubidium
  - (White Rabbit)
  - Central phase alignment
- Network:
  - ≤ 20 ms latency
  - 3 – 6 (9) Gbps/station



More central processing:

- Tied-array beams (Stokes I/IQUV)
- Per-station time series (“fly’s eye”)

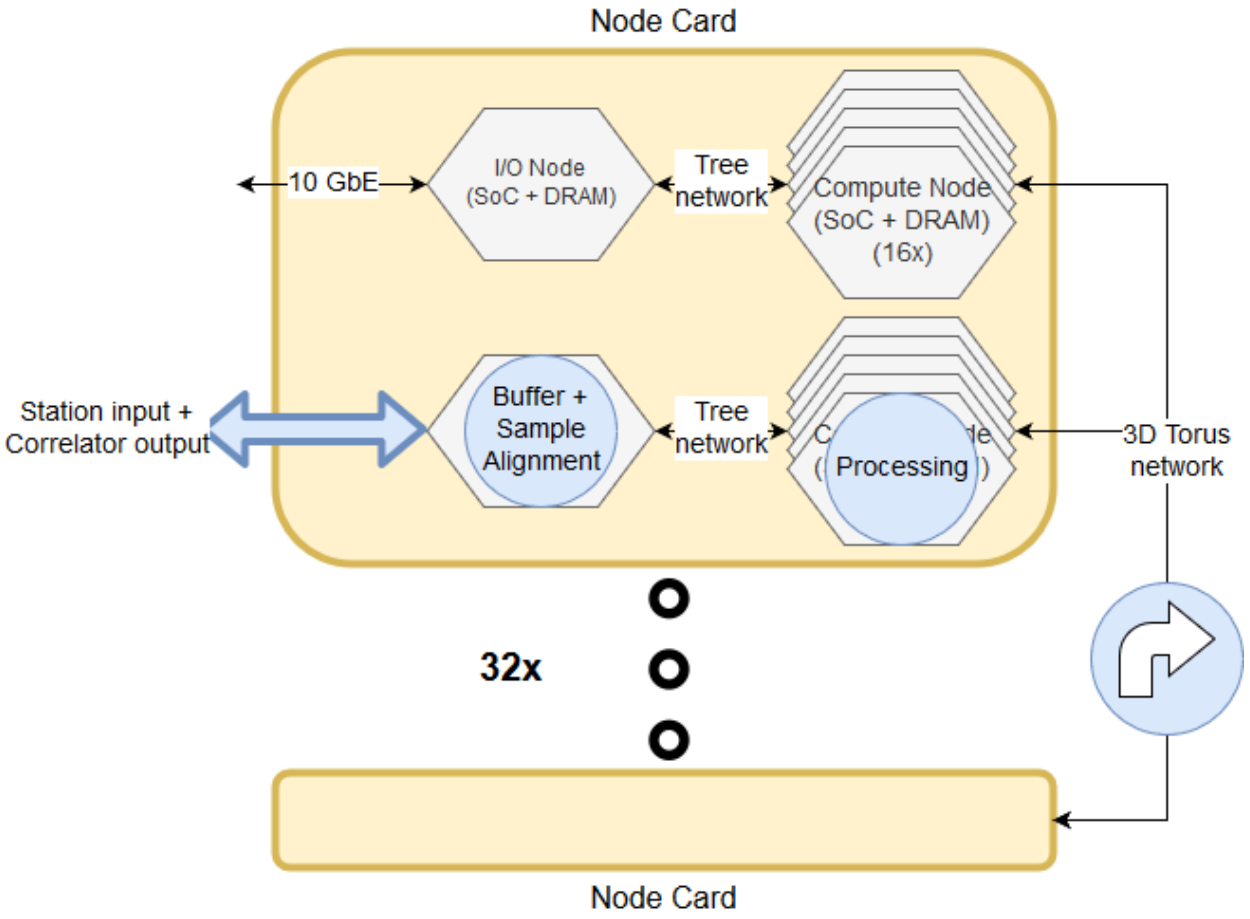
Early days: IBM BlueGene

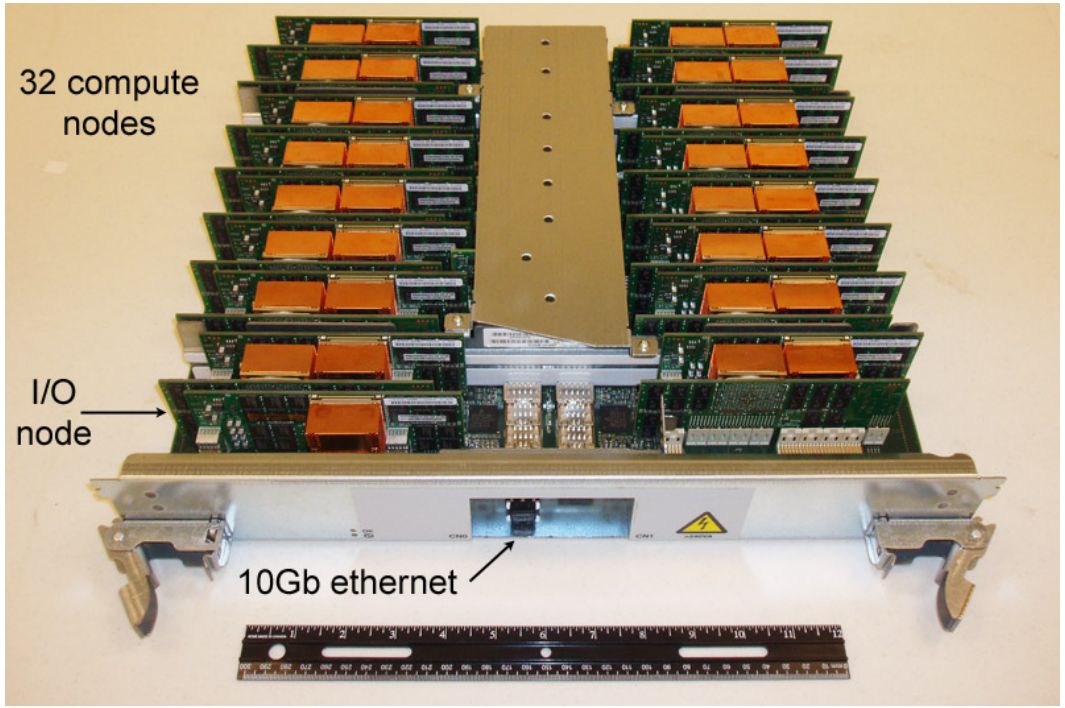
- 2005: IBM BlueGene/L
  - 1 rack + 2 head nodes
  - 1024x PPC440 @ 700 MHz
- 2008: IBM BlueGene/P
  - 1 rack + 2 head nodes
  - 1024x PPC450 @ 850 MHz
- Dual FP pipeline (“Double Hummer”)
  - 1 FLOP/byte
- 1 thread/core
- Fast interconnects
  - 3D Torus + Tree
- I/O: 64x 10GbE

BlueGene/L

BlueGene/P

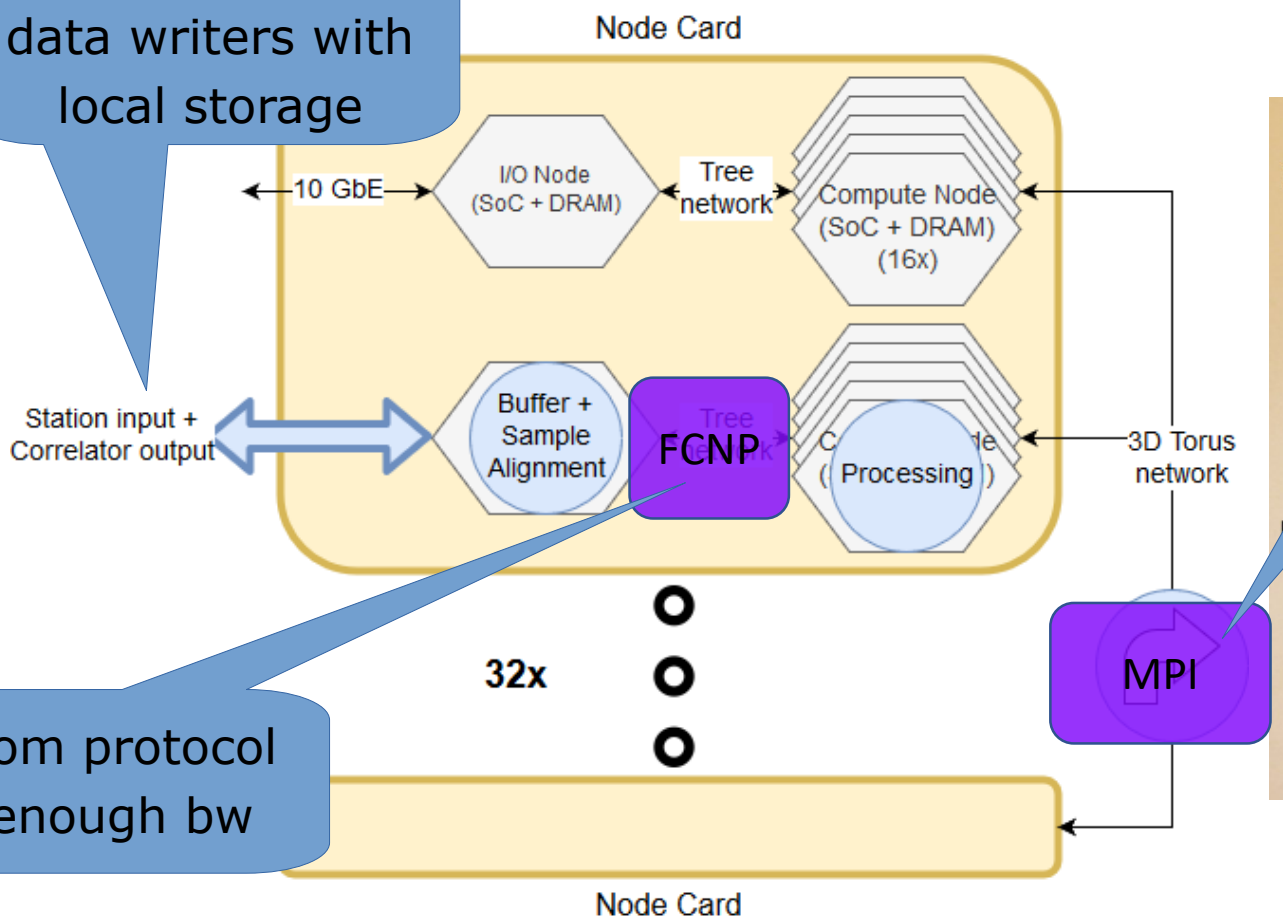
BlueGene software architecture



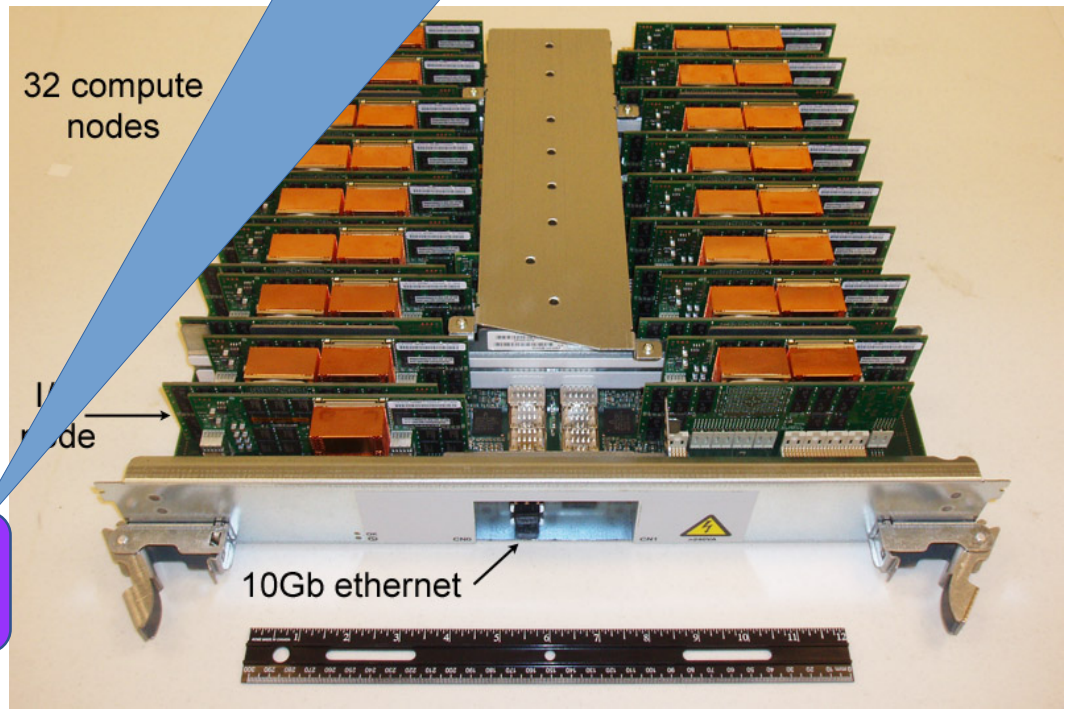


BlueGene parallelisation

Output to 100 data writers with local storage



- Point-to-point
- Avoid link overload

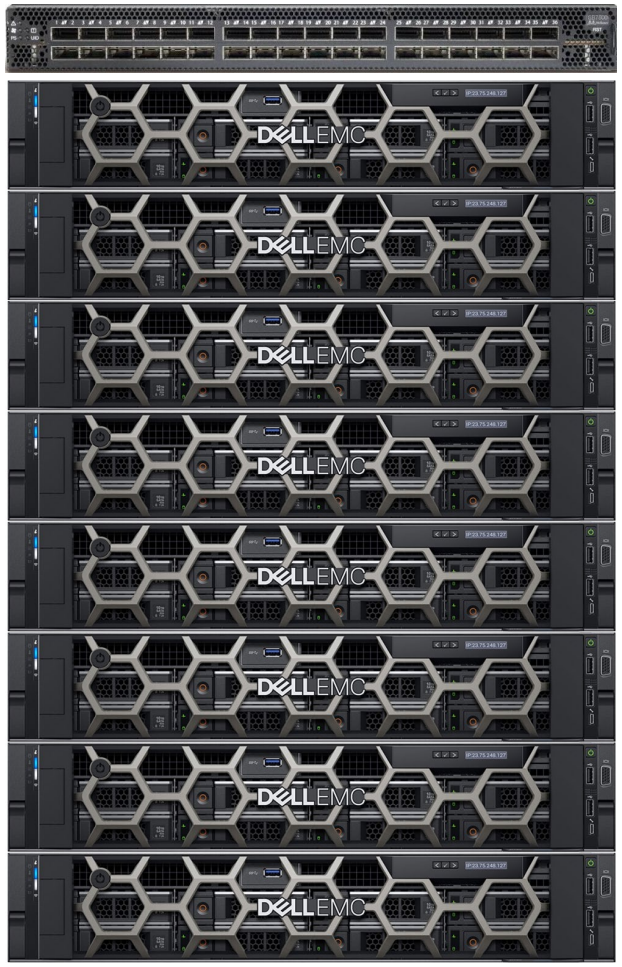


Custom protocol for enough bw

Now: GPU servers (“COBALT”)

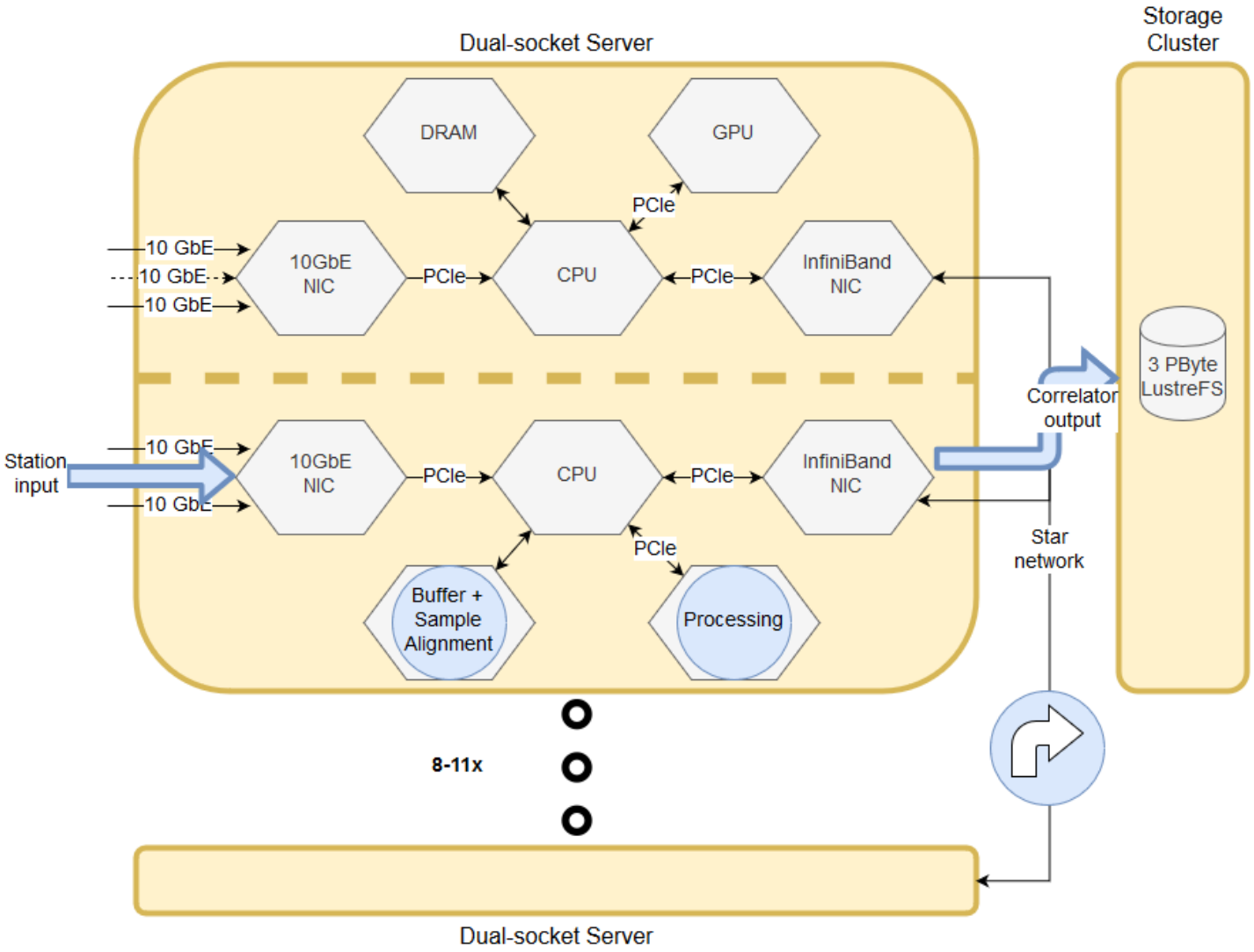
- 2013: COBALT1
  - 8x Dell PowerEdge T640 (Xeon E5-2660)
  - 16x NVIDIA Tesla K10
  - 16x FDR Infiniband (56 Gbps)
  - I/O: 32x 10GbE
- Dec 2018: COBALT2
  - 1x Dell PowerEdge R440 (head node)
  - 11x Dell PowerEdge R740 (Xeon Gold 6140)
  - 22x NVIDIA Tesla V100
  - 23x EDR InfiniBand (100 Gbps)
  - I/O: 80x 10GbE
- COTS
- Balanced PCIe layout

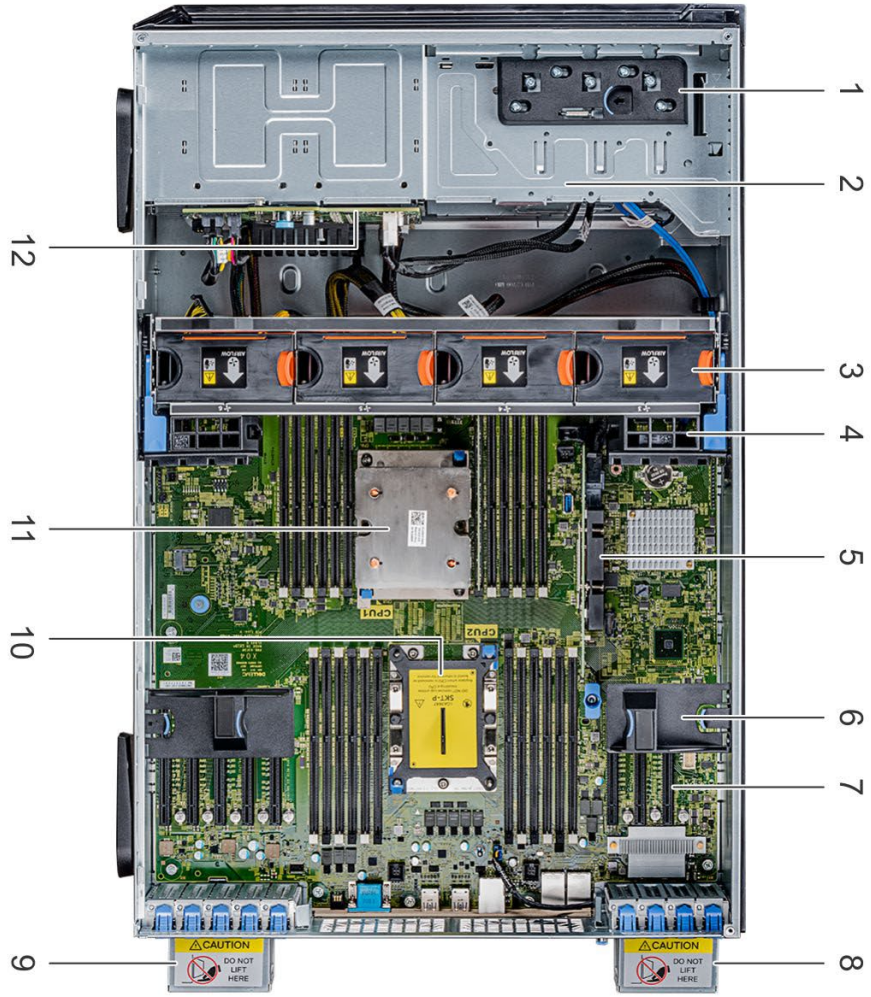
COBALT 1



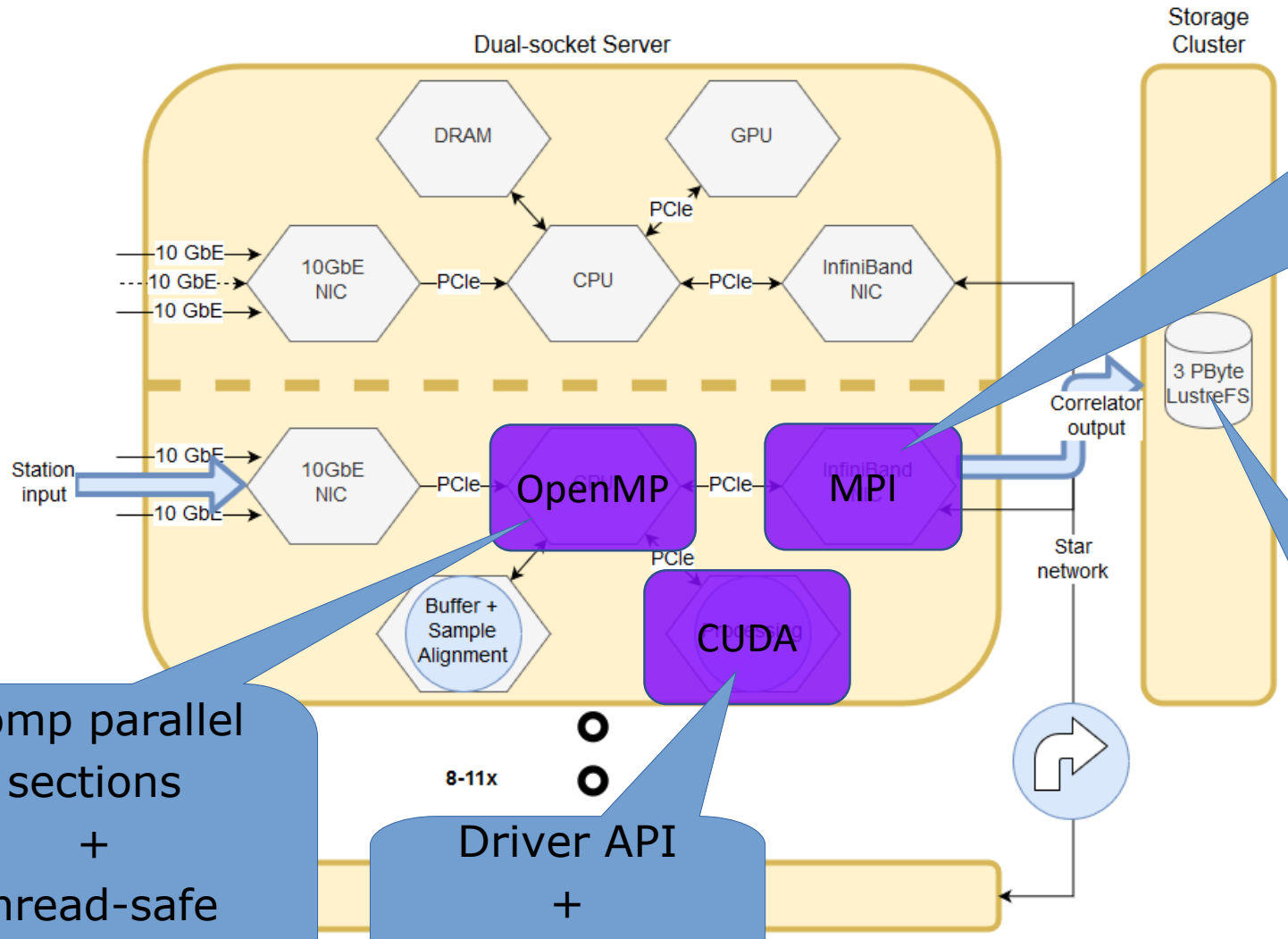
COBALT 2

COBALT software architecture

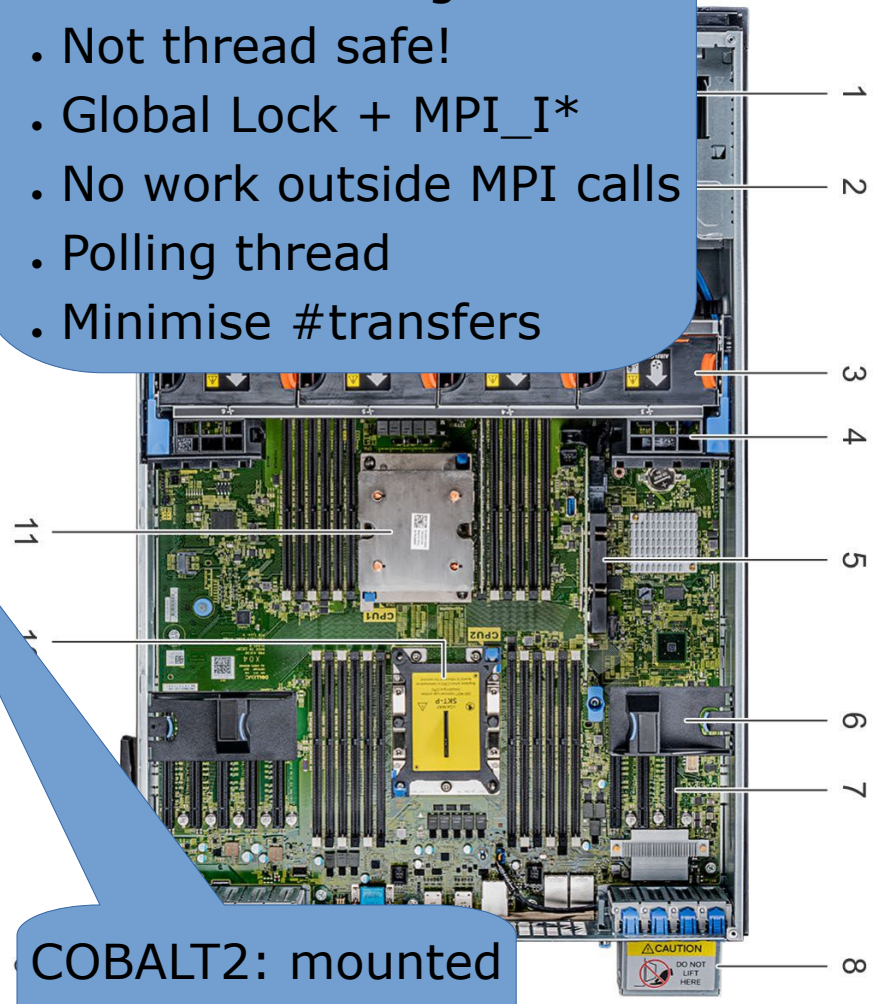




COBALT parallellisation



- Demands tuning
- Not thread safe!
- Global Lock + MPI_I*
- No work outside MPI calls
- Polling thread
- Minimise #transfers

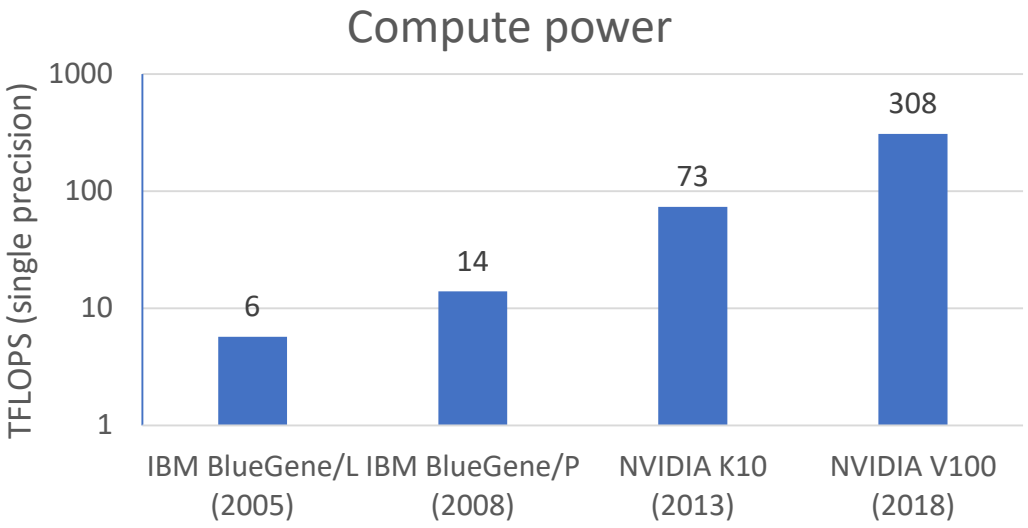


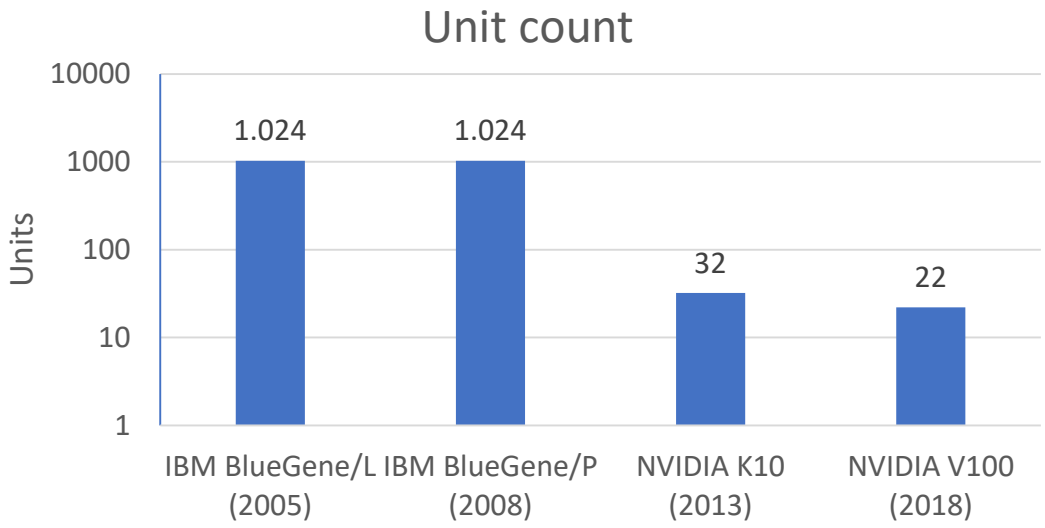
#omp parallel sections + thread-safe queues

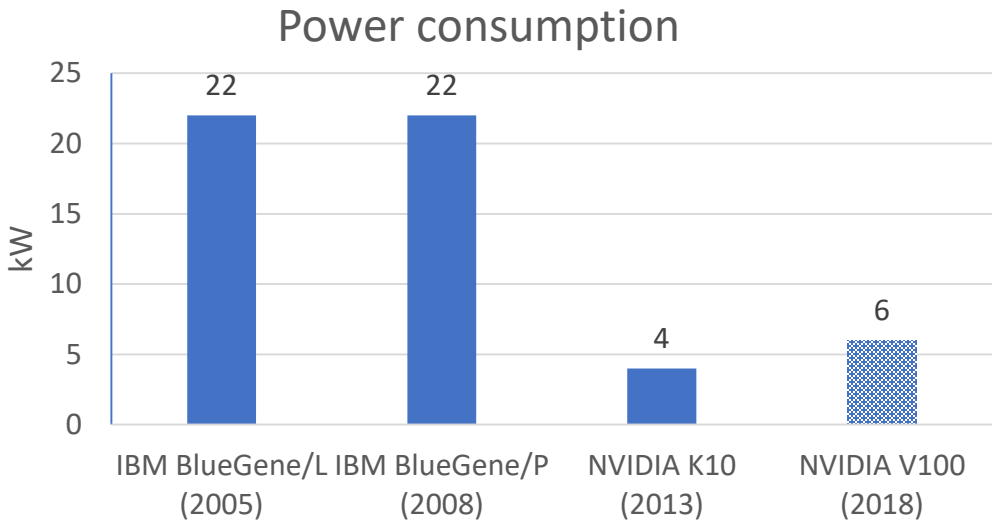
Driver API + JIT compilation

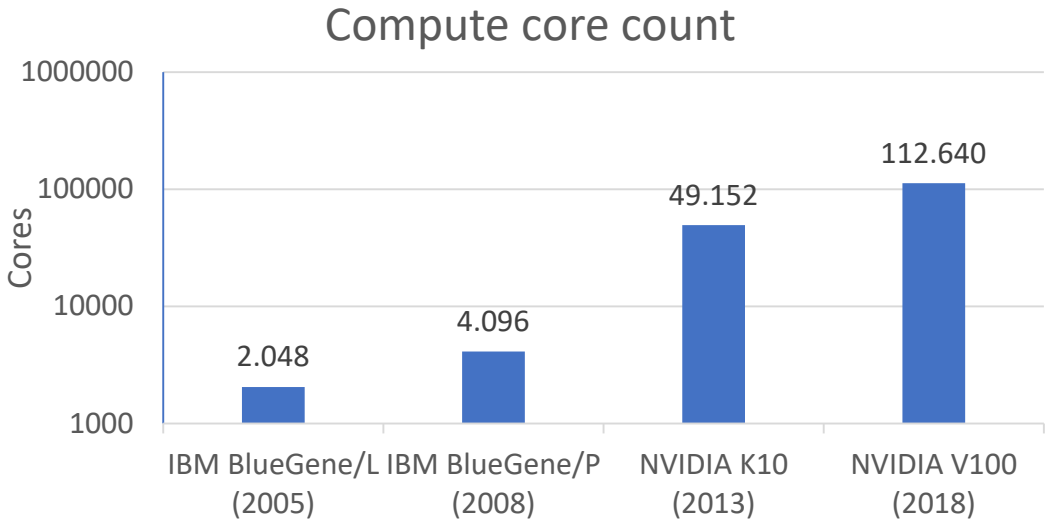
COBALT2: mounted locally

Compute power

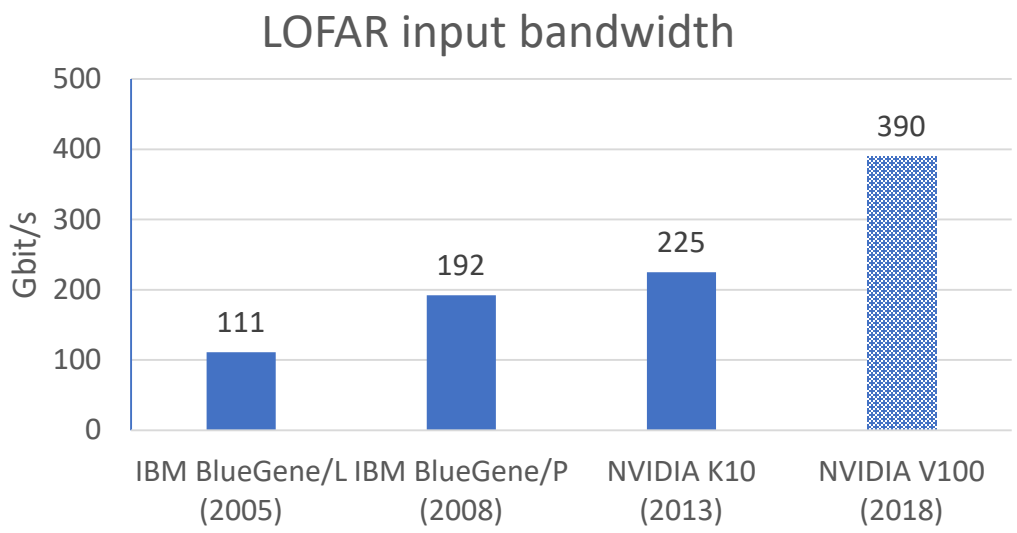


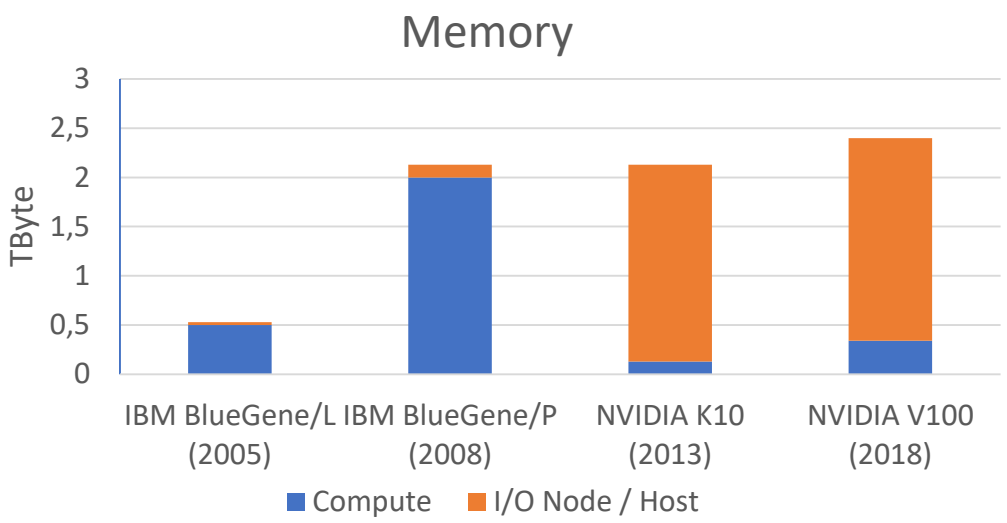


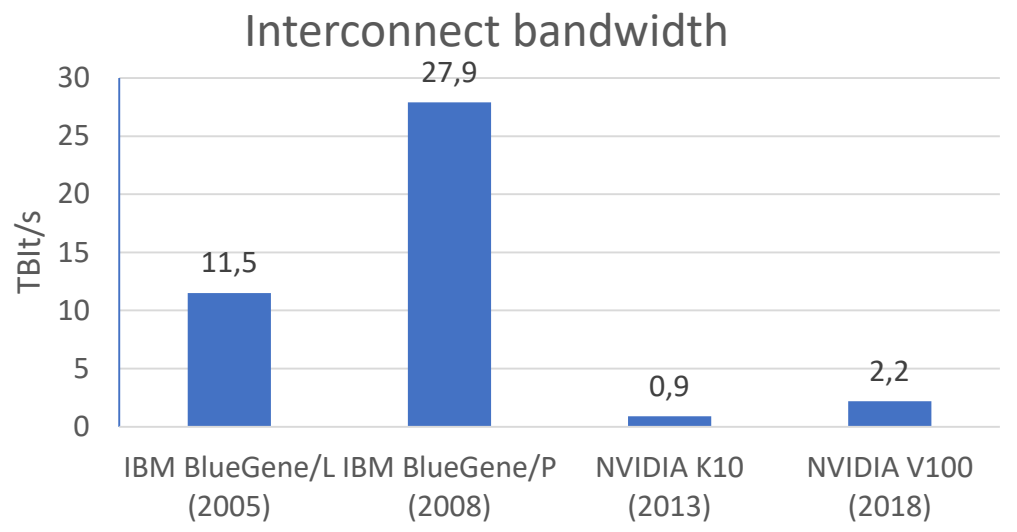


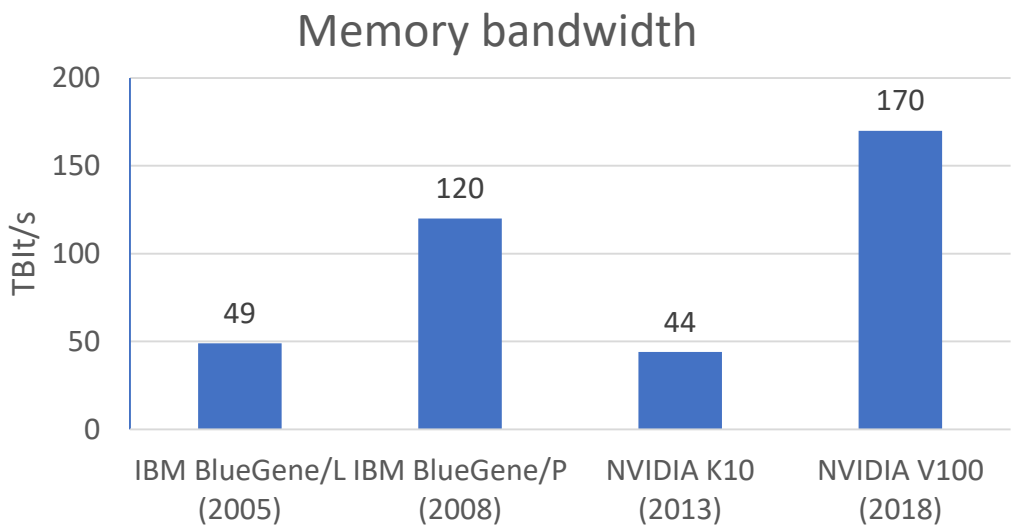


Compute bandwidth









COBALT2: “LOFAR Mega Mode”

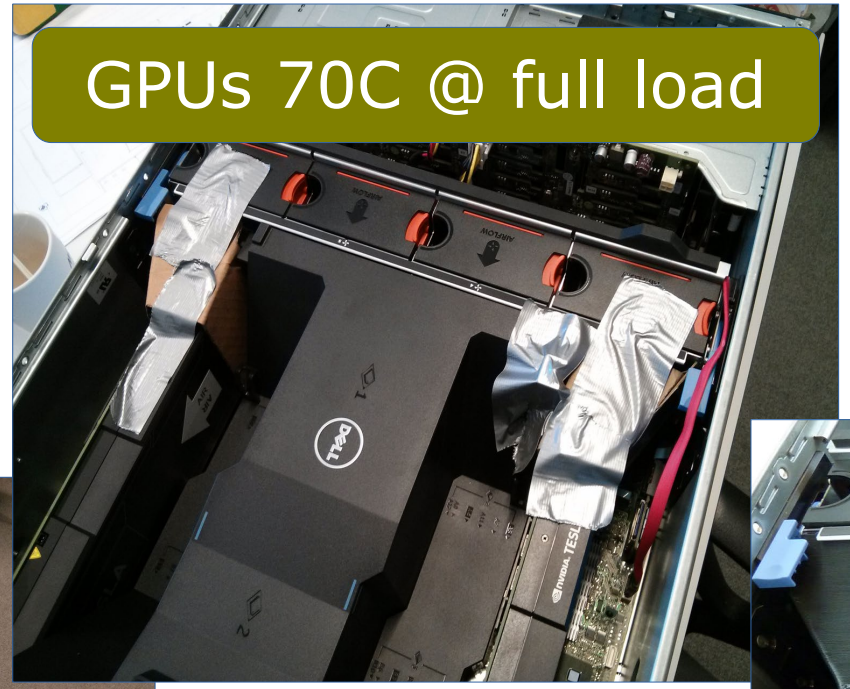
Science (simultaneous)	Processing Type	#P100s	#V100s (est)
All-sky imaging survey	Correlation	2.8	1.9
Targets of opportunity	Correlation	0.3	0.2
Pulsar & Fast transients survey	Tied-array beam forming	4.0	2.6
Recombination lines	Tied-array beam forming	1.6	1.1
Pulsar timing & RM monitoring	Tied-array beam forming	4.9	3.3
Ionospheric scintillation	Per-station time series	2.7	1.8
Total		*16.3*	*10.9 (of 22)*

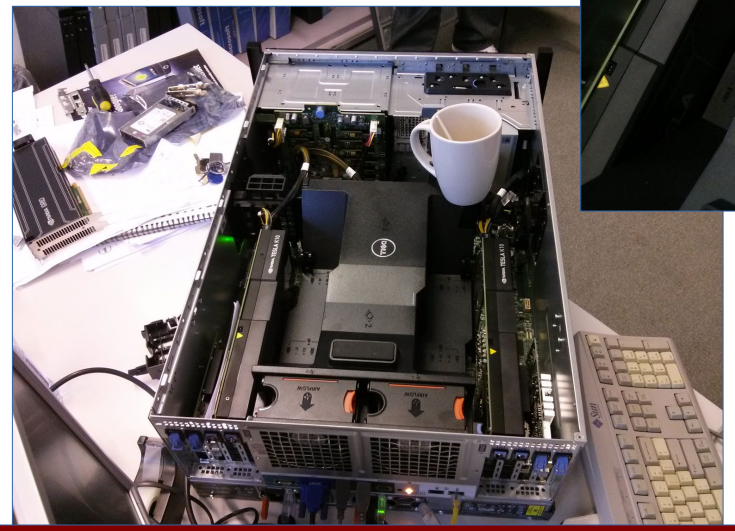
- Highly different settings/submode
- Compute power to spare! Could do f.e.:
  - More tied-array beams
  - Better signal quality (on-line RFI filtering, Doppler tracking, dedispersion)
  - Real-time (uncalibrated) imaging

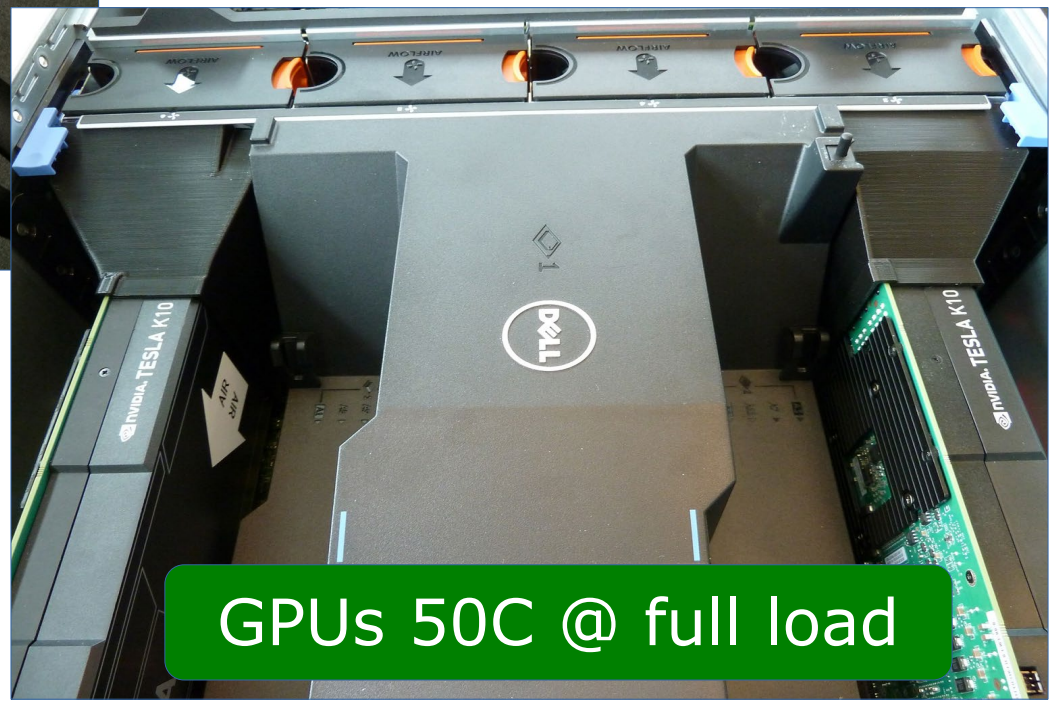
COBALT challenges

- Drop-in replacement
- Hardware/software co-design
  - Real-time processing on non-real-time systems
- EU Tendering (COBALT2)
  - Squeezes vendors
  - Cannot demand end-to-end performance
  - No consumer-grade GPUs, no 32GB V100s…
- Port all code PPC → OpenCL → CUDA
  - Wide range in problem dimensionality
  - JIT kernel compilation
- Driver incompatibility
  - Shared memory vs pinned memory clashes

COBALT1 Temperature Control







Fitness for purpose?

Property	BlueGene	COBALT (GPU servers)
Real-time processing	**√** Little OS interference	**√** Overdimension CPU/DRAM
Stability	**√**	**√**
Platform	**√** Integrated	**X** Frankenstein
Workload distribution	**X** 10-20 compute units/unit of work	**√** 15-20 units of work/compute unit
	X Avoid clashes on tree/torus	**√** Star network
	X Custom low-level tree protocol	**√** Tune common software
Staging memory	**X** Little	**√** Much

Programmability

Property	BlueGene	COBALT (GPU servers)
Compute kernels	**X** Assembly (to use special FP instructions)	**√** Common GPU techniques
New outputs	**X** Manual routing…	**√** Local disk write (Lustre)
Debugging	**X** Track work across 4096 cores	**√** Few units, easy work distribution
	X Remote debugging	**√** Common GPU tools
Crash recovery	**X** 45-minutes (not uncommon)	**√/X** 15-minutes (rare)
Expertise	**X** Rare (PPC assembly)	**√** More common

Conclusions

- Yes, GPUs are powerful and hot
  - We know that
- Whole COTS package makes it shine:
  - Work vs compute unit ratio (easy work distribution),
  - Network topology, global fs (easy input/output distribution),
  - Powerful middleware (many problems are solved for you)
- Allows focus on flexibility
- Allows more science


Thank you
ASTRON
Netherlands Institute for Radio Astronomy
2011 © AEROPHOTO EELDE