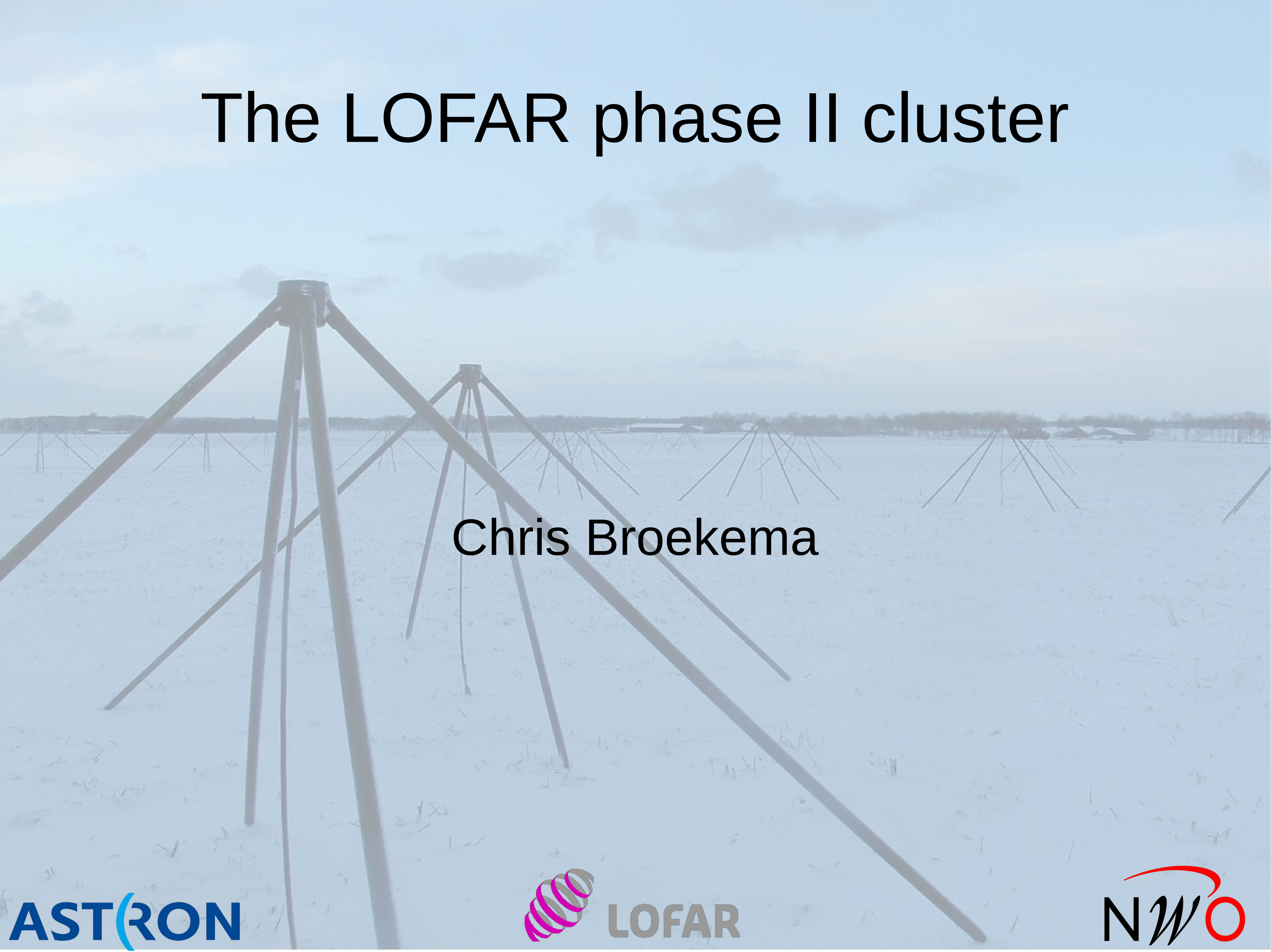
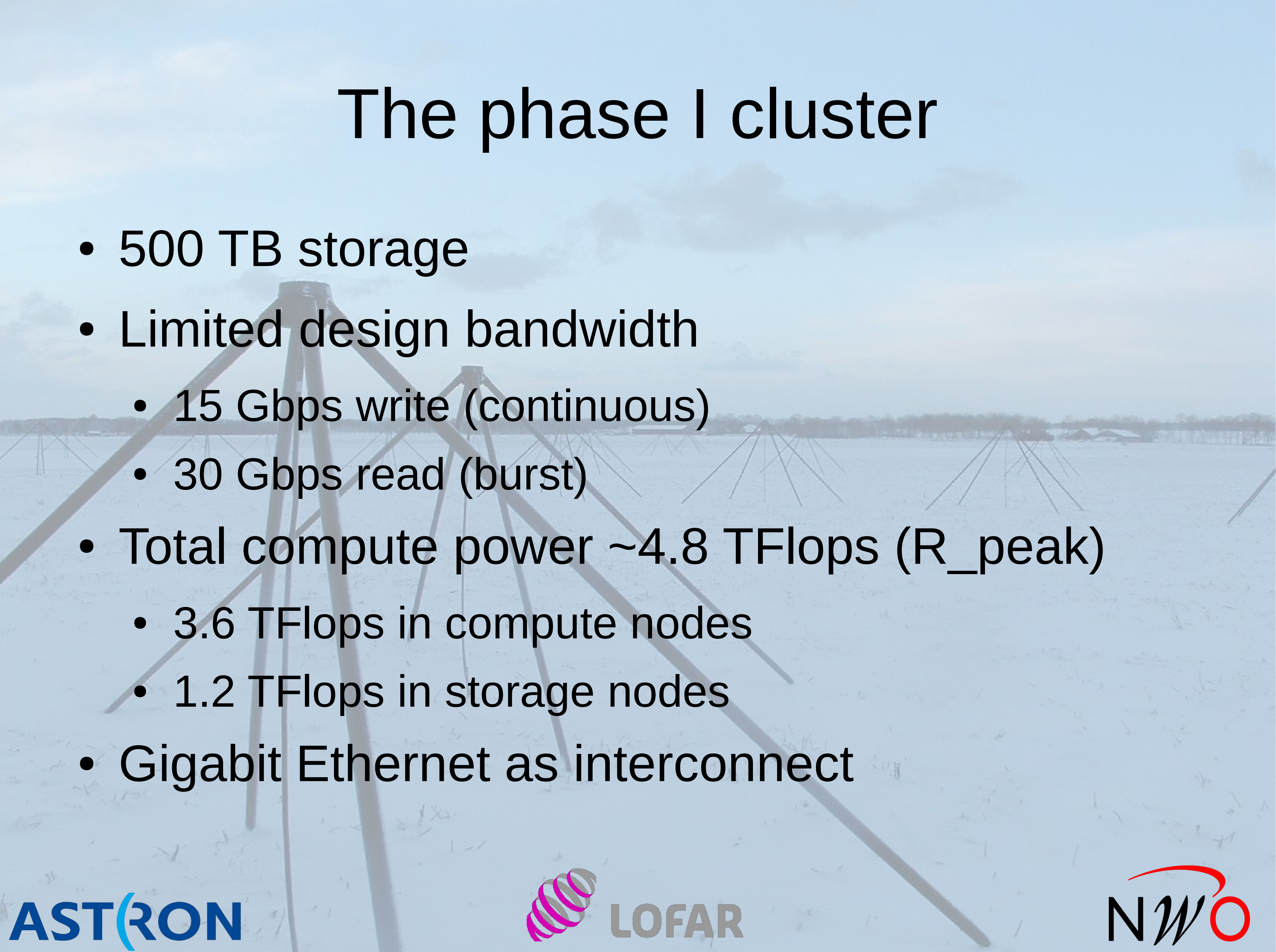
The LOFAR phase II cluster



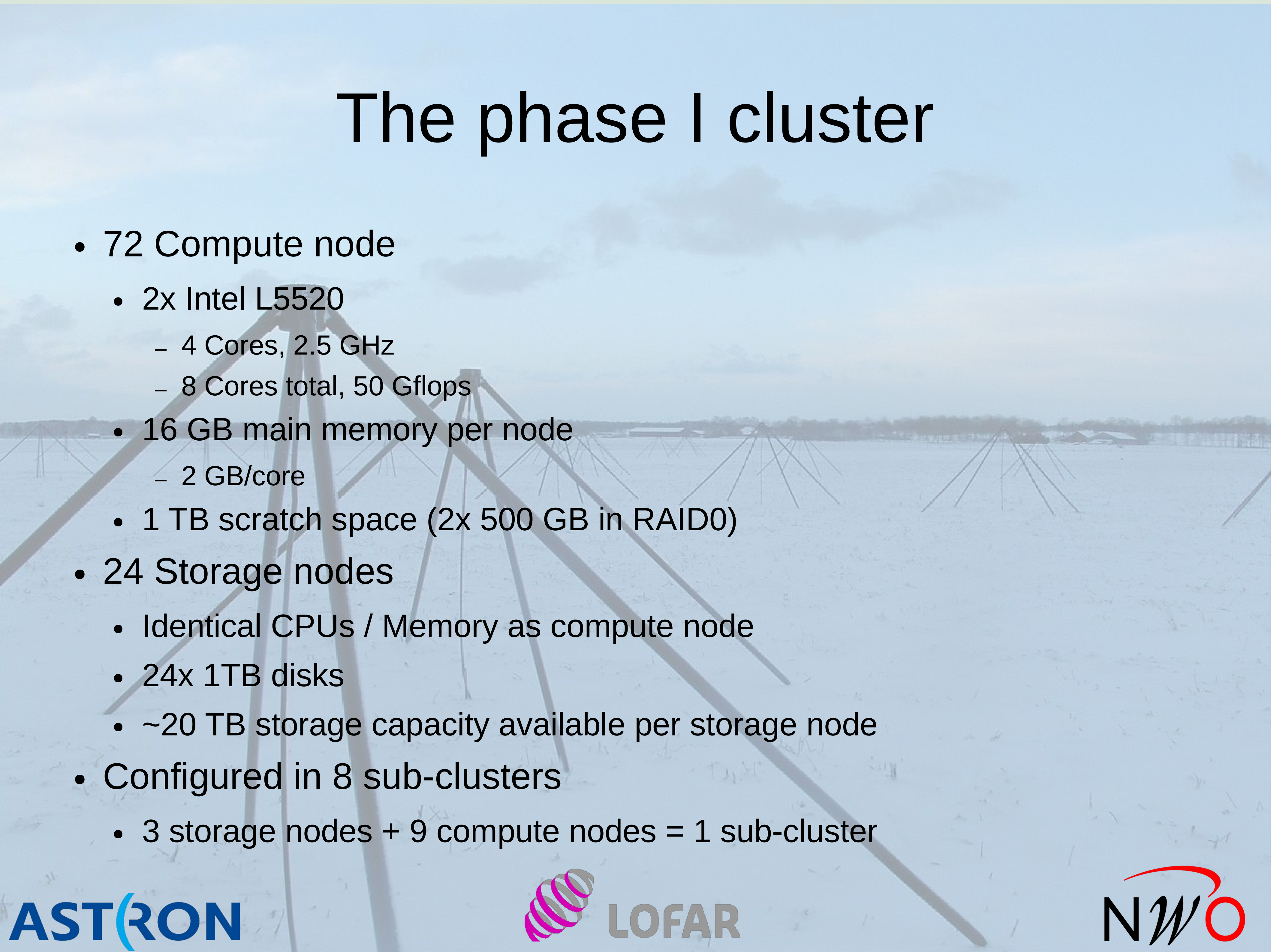
Chris Broekema

The phase I cluster



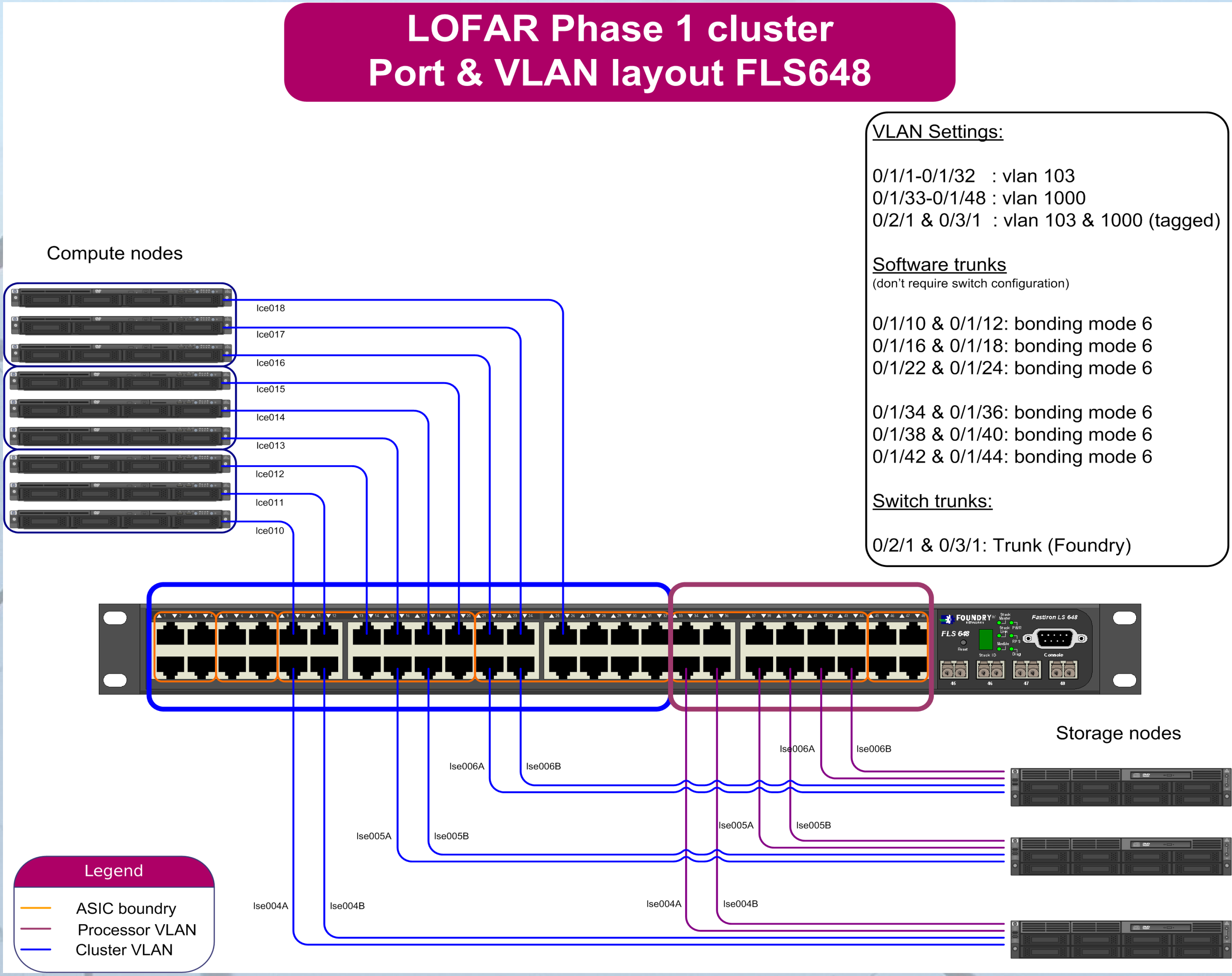
- 500 TB storage
- Limited design bandwidth
  - 15 Gbps write (continuous)
  - 30 Gbps read (burst)
- Total compute power ~4.8 TFlops (R_peak)
  - 3.6 TFlops in compute nodes
  - 1.2 TFlops in storage nodes
- Gigabit Ethernet as interconnect

The phase I cluster



- 72 Compute node
  - 2x Intel L5520
    - 4 Cores, 2.5 GHz
    - 8 Cores total, 50 Gflops
  - 16 GB main memory per node
    - 2 GB/core
  - 1 TB scratch space (2x 500 GB in RAID0)
- 24 Storage nodes
  - Identical CPUs / Memory as compute node
  - 24x 1TB disks
  - ~20 TB storage capacity available per storage node
- Configured in 8 sub-clusters
  - 3 storage nodes + 9 compute nodes = 1 sub-cluster

The phase I cluster



LOFAR Phase 1 cluster
Port & VLAN layout FLS648

VLAN Settings:

0/1/1-0/1/32 : vlan 103
0/1/33-0/1/48 : vlan 1000
0/2/1 & 0/3/1 : vlan 103 & 1000 (tagged)

Software trunks
(don't require switch configuration)

0/1/10 & 0/1/12: bonding mode 6
0/1/16 & 0/1/18: bonding mode 6
0/1/22 & 0/1/24: bonding mode 6

0/1/34 & 0/1/36: bonding mode 6
0/1/38 & 0/1/40: bonding mode 6
0/1/42 & 0/1/44: bonding mode 6

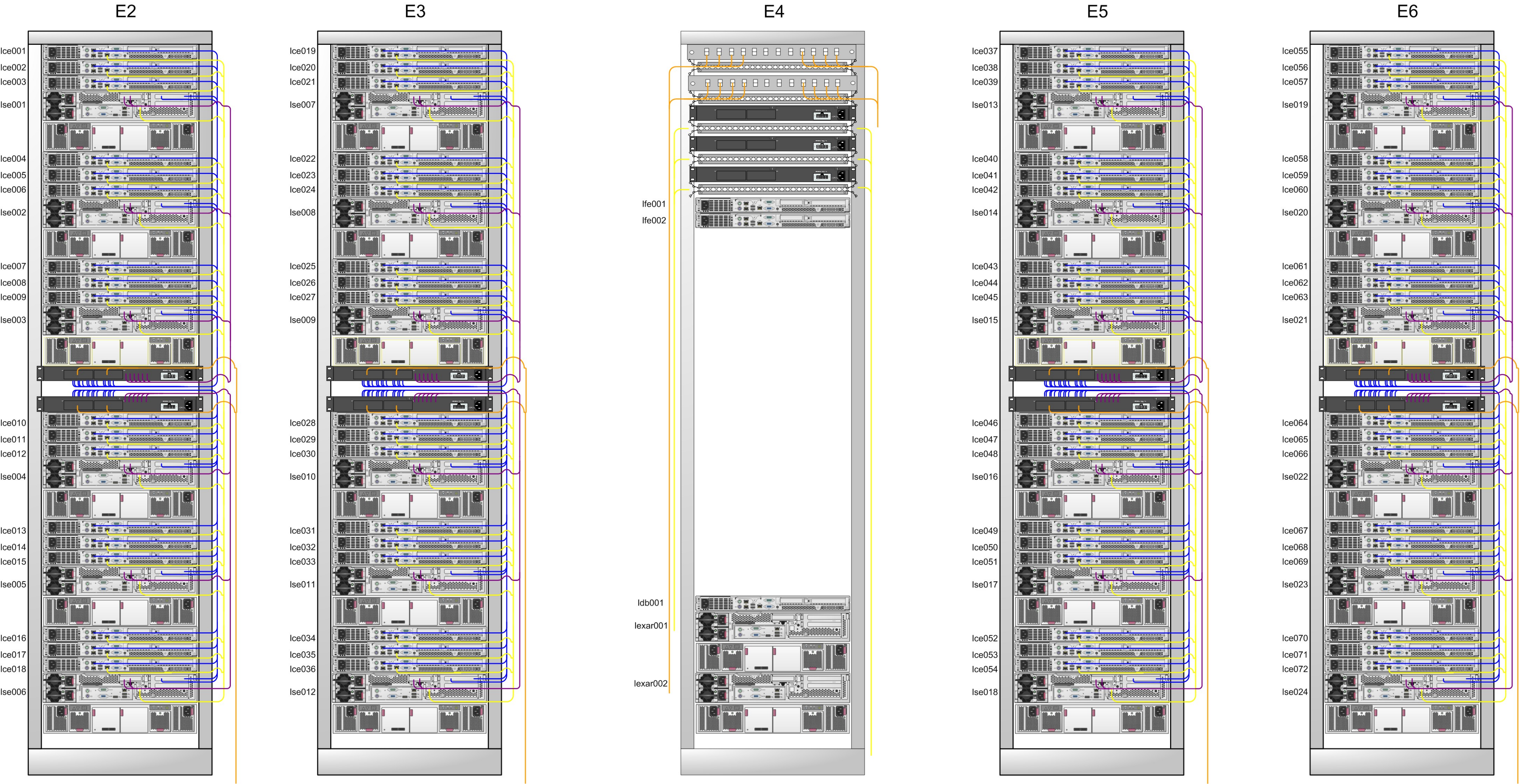
Switch trunks:

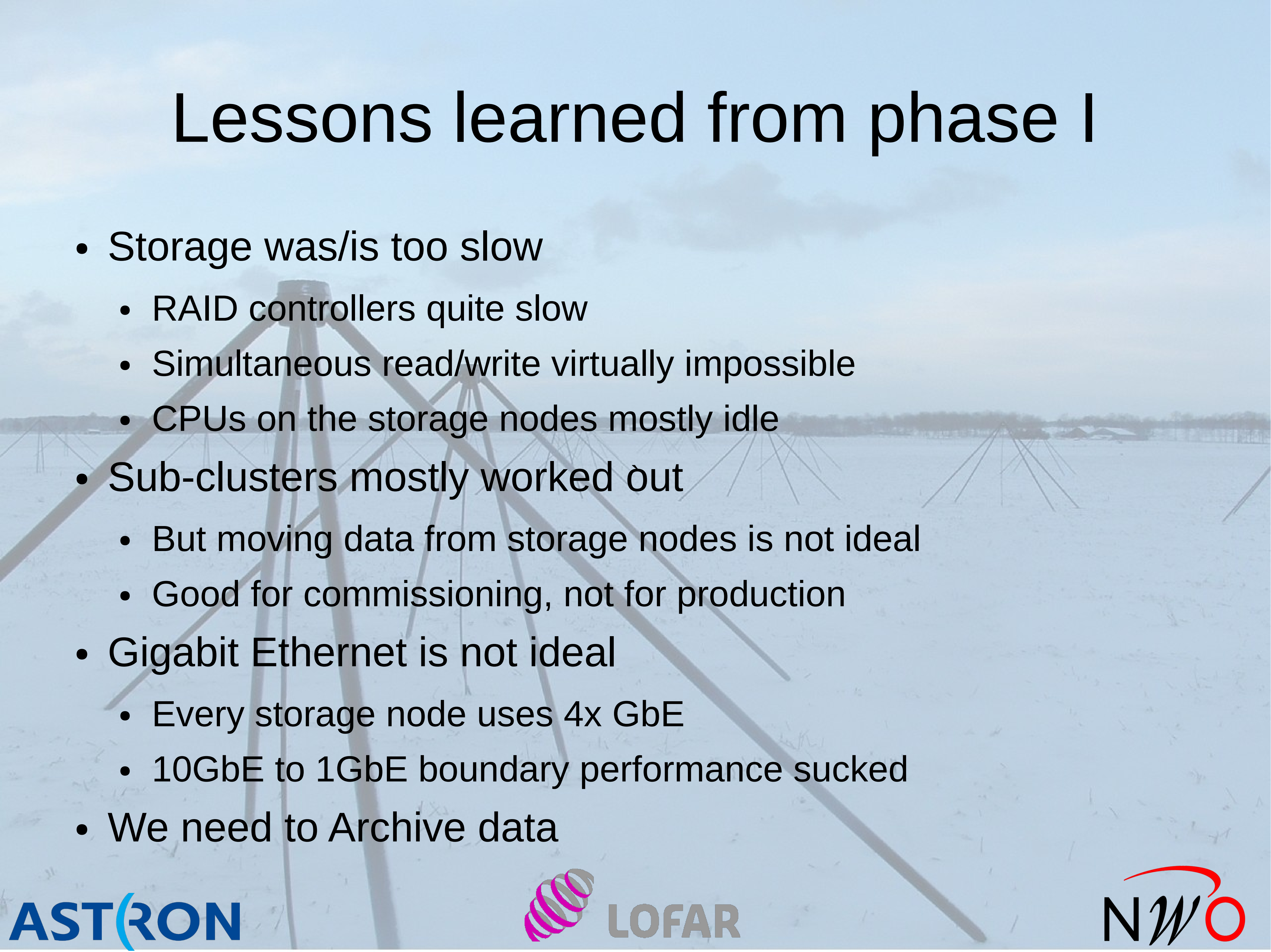
0/2/1 & 0/3/1: Trunk (Foundry)

LOFAR Phase 1 cluster
Rear view

Legend

- 10GbE Uplink
- 1 GbE Data line (RSP & TBB)
- 1 GbE offline data link
- Management network

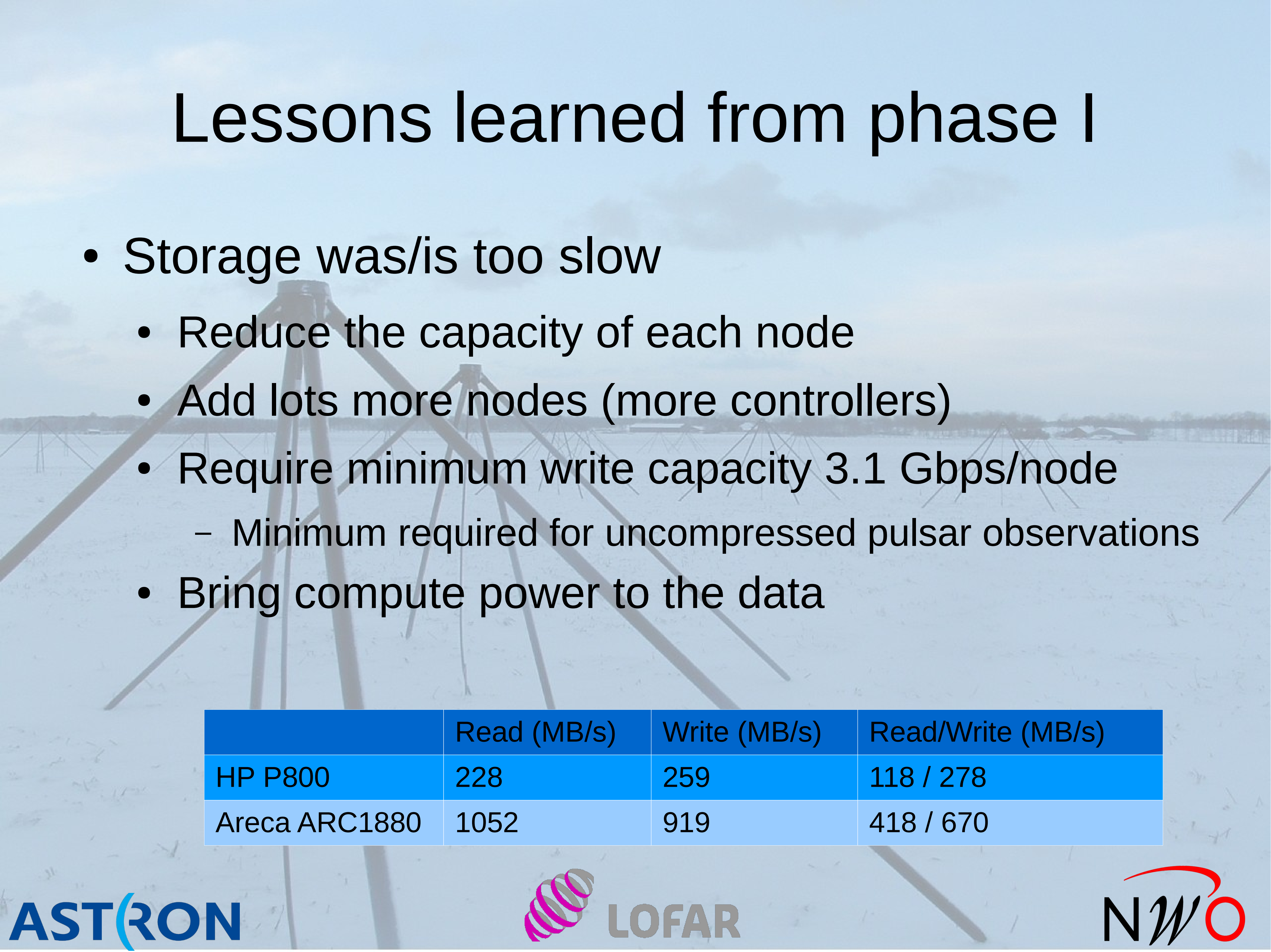




Lessons learned from phase I

- Storage was/is too slow
  - RAID controllers quite slow
  - Simultaneous read/write virtually impossible
  - CPUs on the storage nodes mostly idle
- Sub-clusters mostly worked out
  - But moving data from storage nodes is not ideal
  - Good for commissioning, not for production
- Gigabit Ethernet is not ideal
  - Every storage node uses 4x GbE
  - 10GbE to 1GbE boundary performance sucked
- We need to Archive data

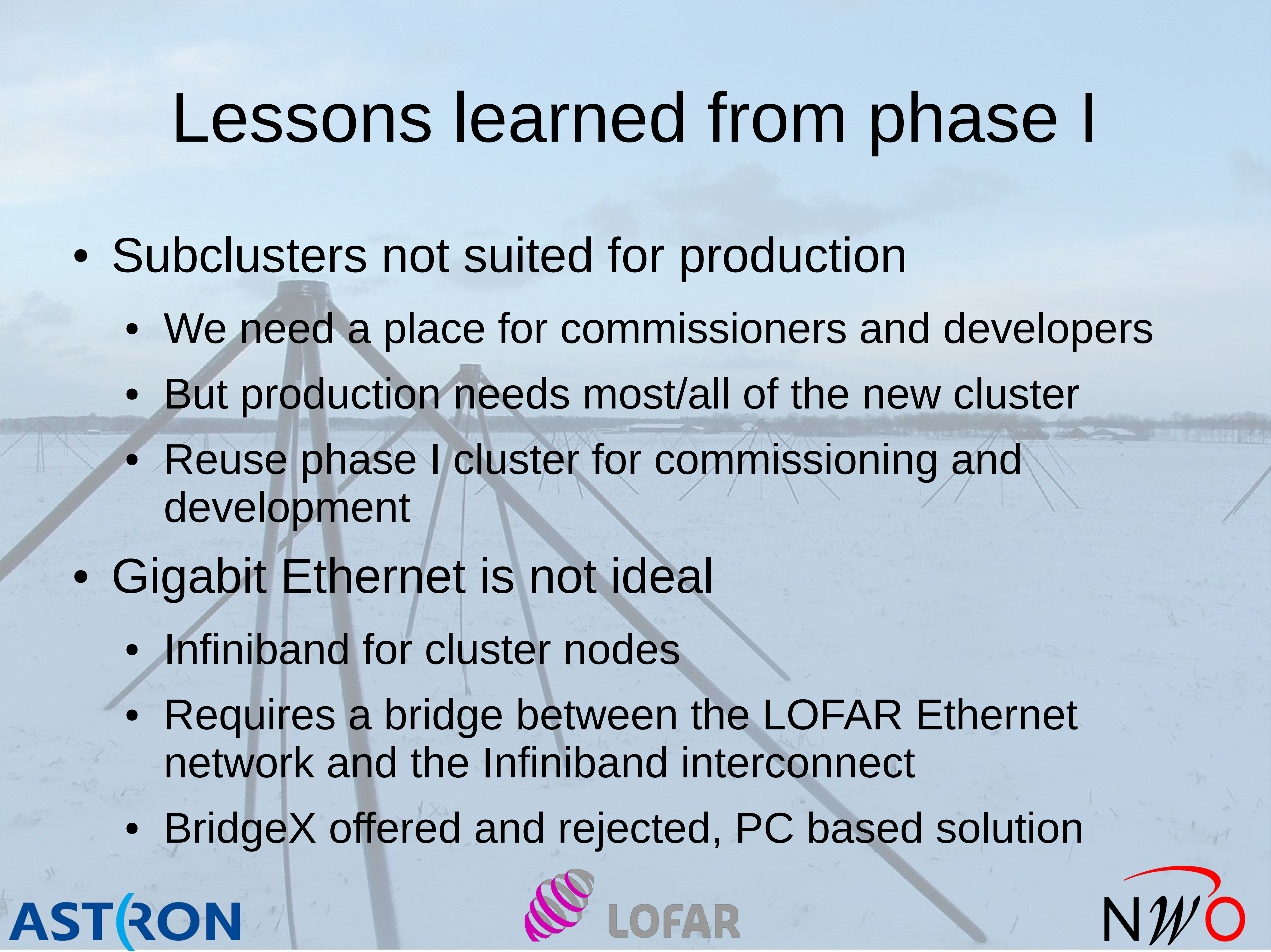
ASTRON LOFAR NWO



Lessons learned from phase I

- Storage was/is too slow
  - Reduce the capacity of each node
  - Add lots more nodes (more controllers)
  - Require minimum write capacity 3.1 Gbps/node
    - Minimum required for uncompressed pulsar observations
  - Bring compute power to the data

	Read (MB/s)	Write (MB/s)	Read/Write (MB/s)
HP P800	228	259	118 / 278
Areca ARC1880	1052	919	418 / 670



Lessons learned from phase I

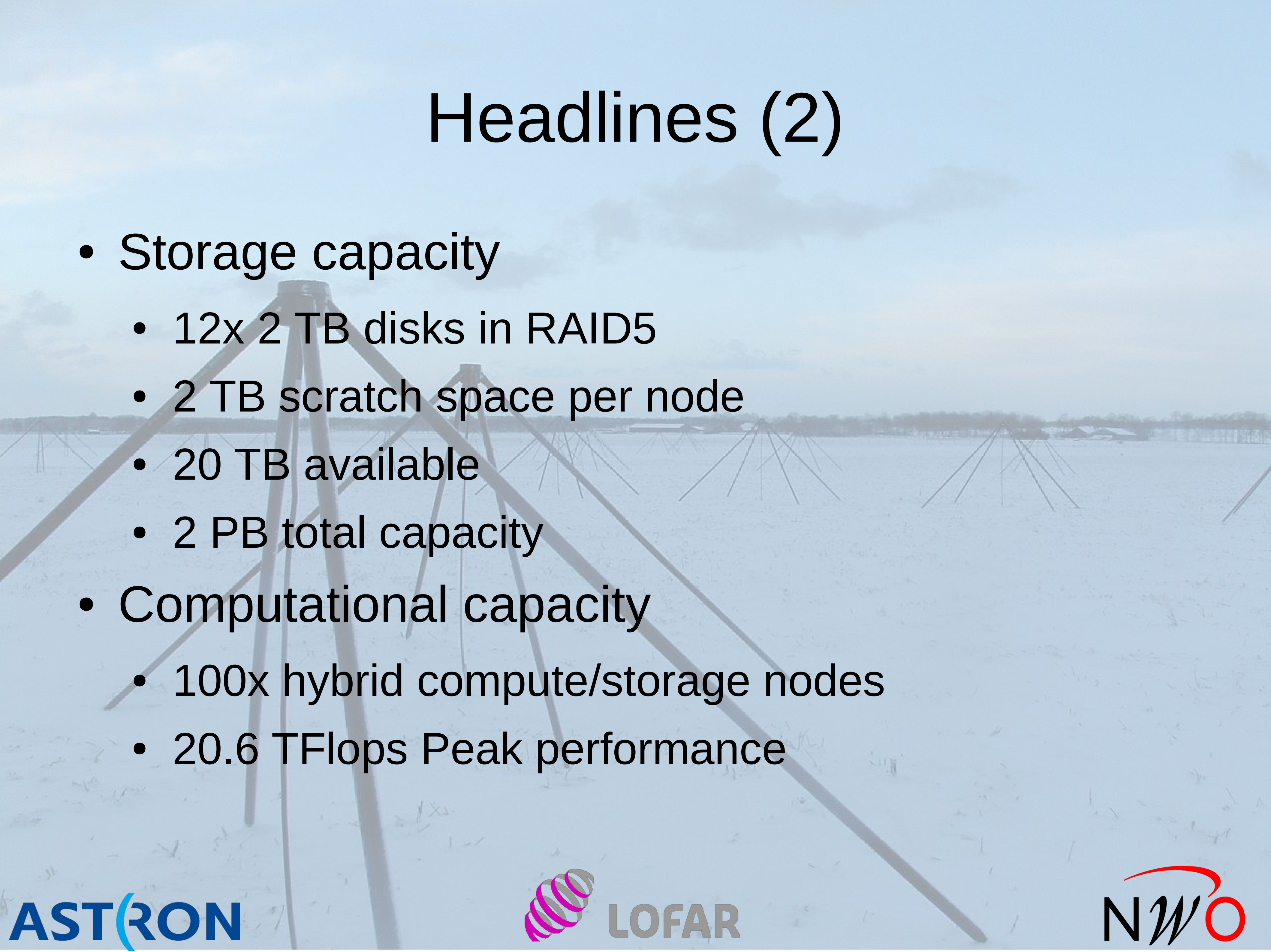
- Subclusters not suited for production
  - We need a place for commissioners and developers
  - But production needs most/all of the new cluster
  - Reuse phase I cluster for commissioning and development
- Gigabit Ethernet is not ideal
  - Infiniband for cluster nodes
  - Requires a bridge between the LOFAR Ethernet network and the Infiniband interconnect
  - BridgeX offered and rejected, PC based solution

Lessons learned from phase I

- We need to archive data
  - Was not fully considered when building phase I
  - Reuse storage nodes phase I as staging cluster
  - Target project in Groningen
    - Provide high bandwidth archive facility
    - For end products and raw data
    - Possible to stream directly to Target from phase II cluster
    - Hopefully will provide significant additional compute power too
  - LTA's in Amsterdam and Juelich
    - Currently less bandwidth available (but we're ready for more)
    - Archive only end products here

Headlines (1)

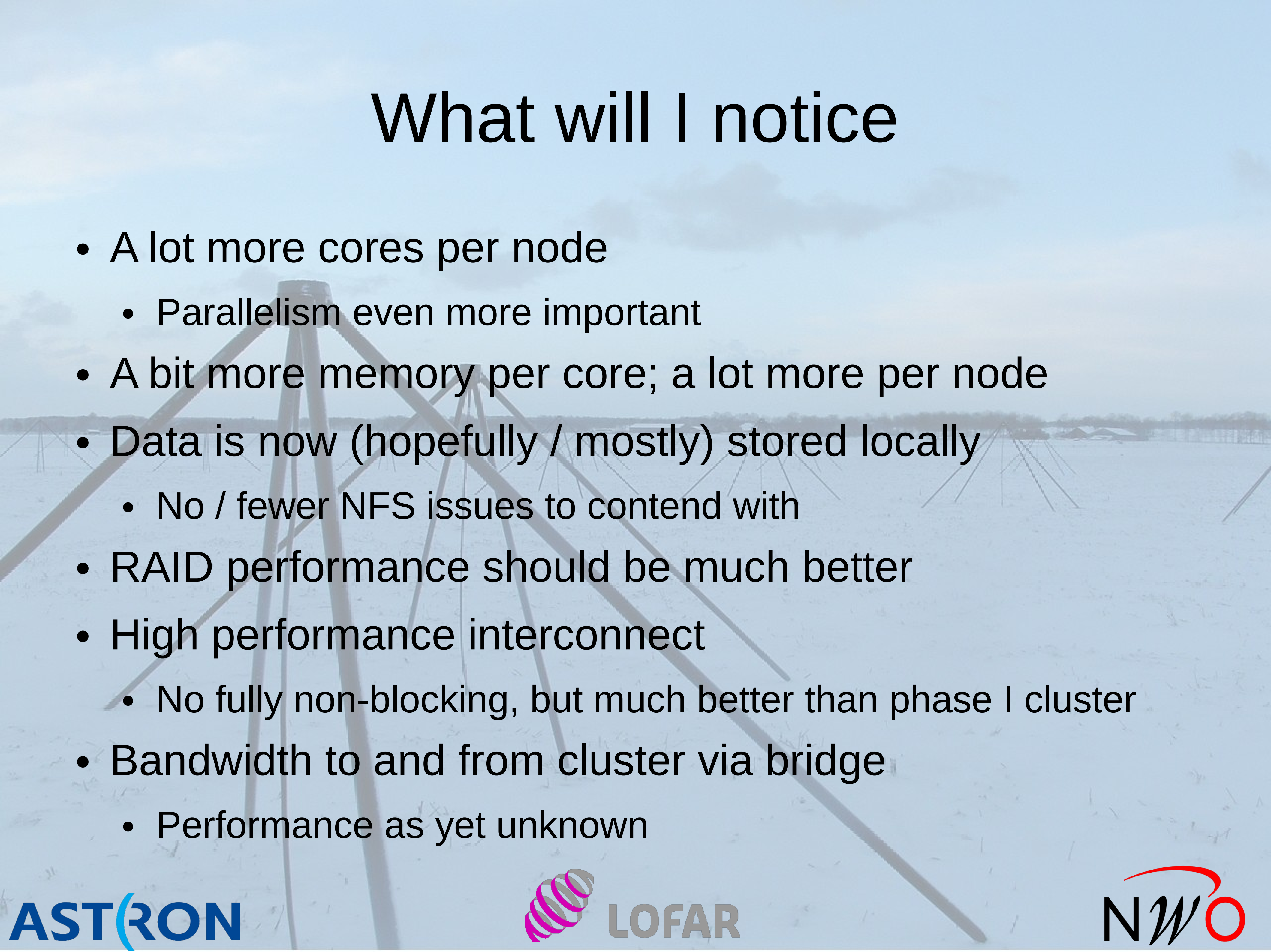
- 100 hybrid storage / compute nodes
  - 2x AMD Opteron 6172 CPUs per node
    - 12 Cores, 2.1 GHz
    - 24 Cores total, 201,6 GFlops
  - 64 GB main memory per node
    - 2.67 GB/core
  - Areca ARC-1880ixl RAID controller
  - 12x 2TB 7200 rpm disks
    - ~20 TB storage capacity available per node
  - QDR Infiniband interconnect
    - 32 Gbp theoretical maximum throughput (~20 Gbps realistic)



Headlines (2)

- Storage capacity
  - 12x 2 TB disks in RAID5
  - 2 TB scratch space per node
  - 20 TB available
  - 2 PB total capacity
- Computational capacity
  - 100x hybrid compute/storage nodes
  - 20.6 TFlops Peak performance

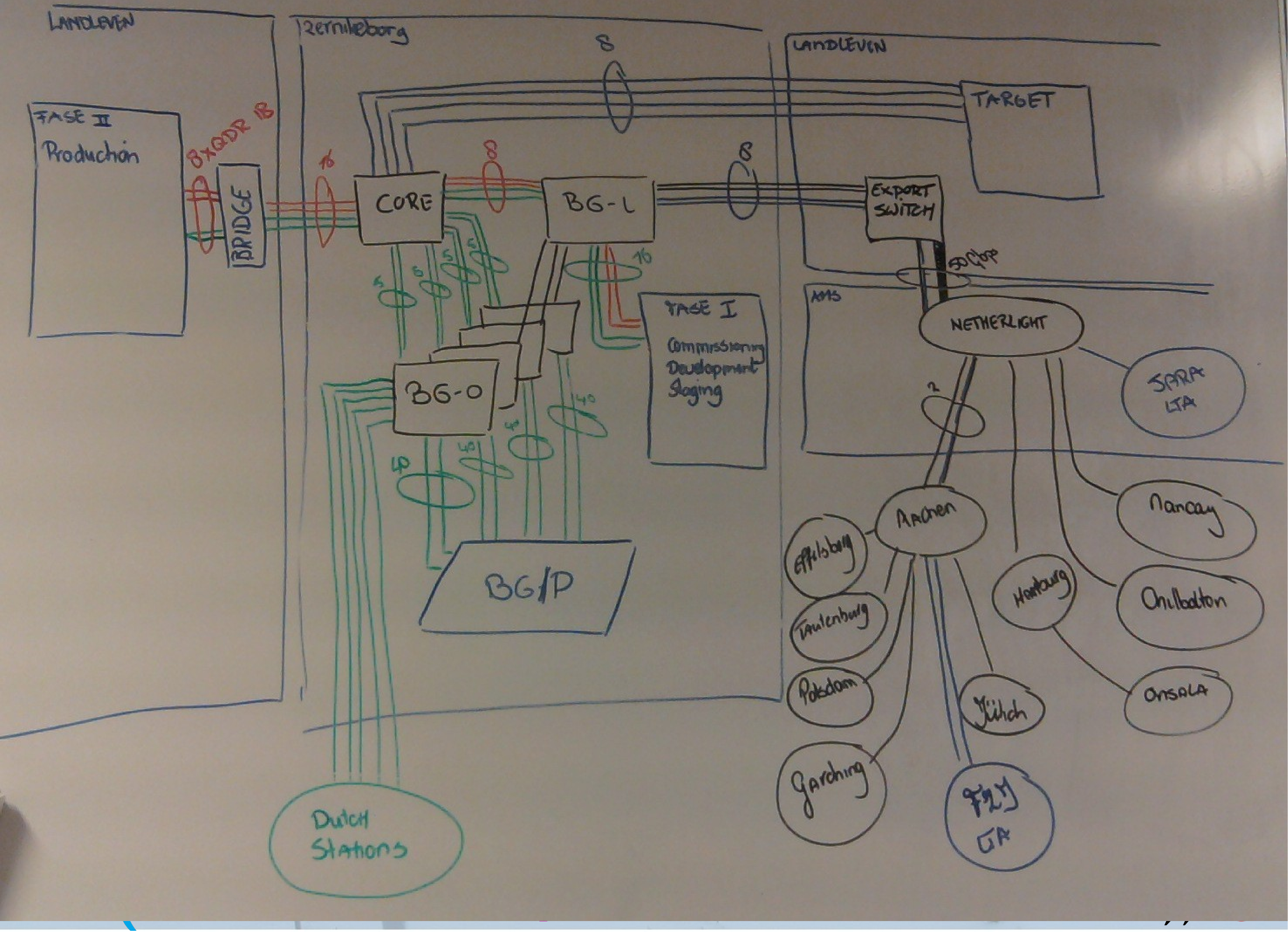
What will I notice



- A lot more cores per node
  - Parallelism even more important
- A bit more memory per core; a lot more per node
- Data is now (hopefully / mostly) stored locally
  - No / fewer NFS issues to contend with
- RAID performance should be much better
- High performance interconnect
  - No fully non-blocking, but much better than phase I cluster
- Bandwidth to and from cluster via bridge
  - Performance as yet unknown

Phase II cluster – Bandwidths (Gbps)

To From	Stations	BG/P	Cluster	Staging	Target	LTA's
Stations	X	200	80-160	80	-	-
BG/P	-	X	80-160	80	-	-
Cluster	-	80-160	X	80	80	-
Staging	-	-	80	X	80	10-80
Target	-	-	-	-	X	-
LTA's	-	-	-	-	-	X


LANDLEVEN
Zernikeborg
LANDLEVEN
FASE II
Production
8xQDR IB
BRIDGE
16
CORE
8
BG-L
8
8
TARGET
EXPORT SWITCH
50Gbp
AMS
NETHERLIGHT
SARA LTA
16
FASE I
Commissioning
Development
Staging
5
6
5
6
BG-0
40
40
40
40
40
BG/P
Dutch Stations
2
Aachen
Effelsberg
Tautenburg
Potsdam
Garching
Jülich
FZJ LTA
Hamburg
Nancay
Chilbolton
Onsala