Looking beyond the LOFAR central processor

Chris Broekema

ASTRON

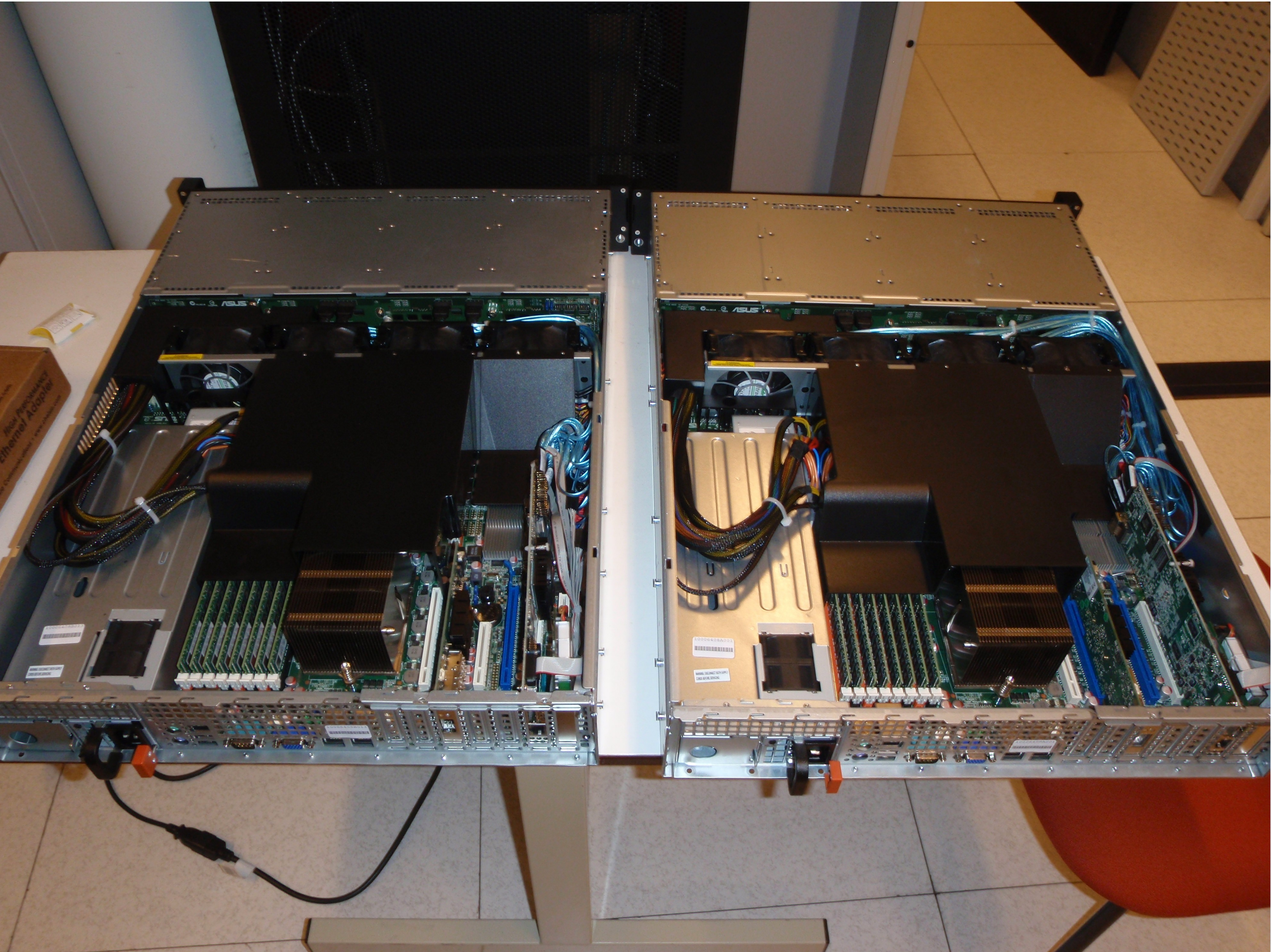
CEP Phase II cluster

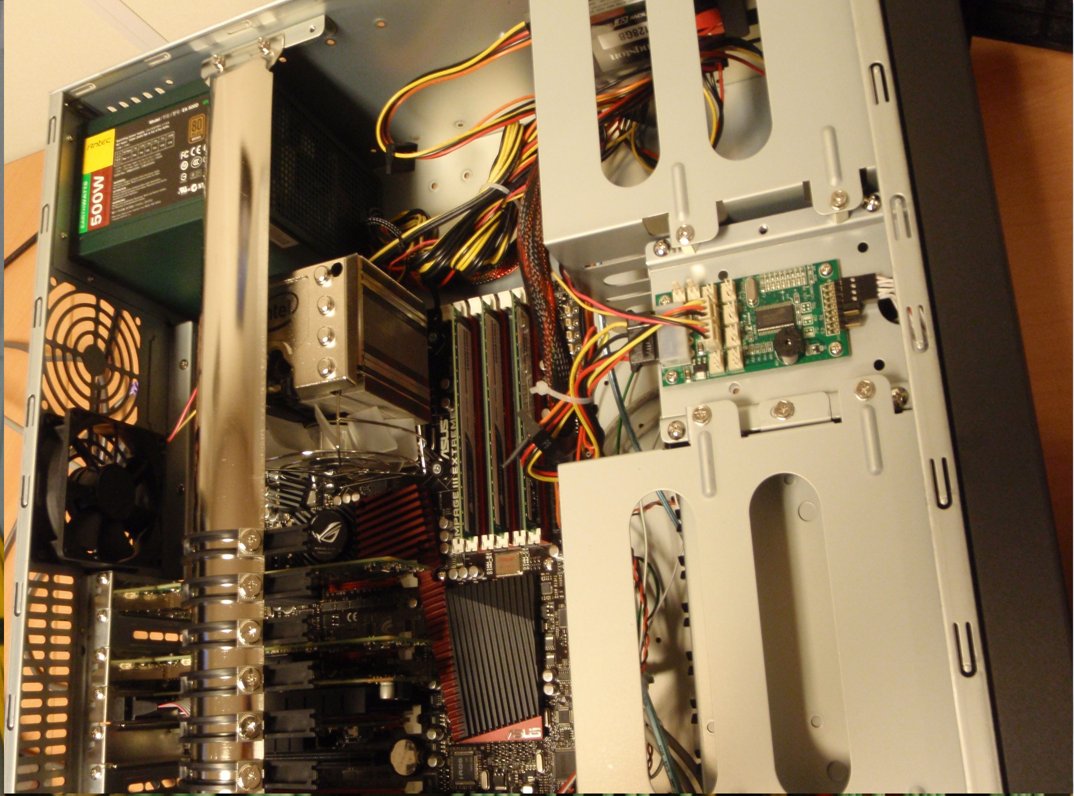
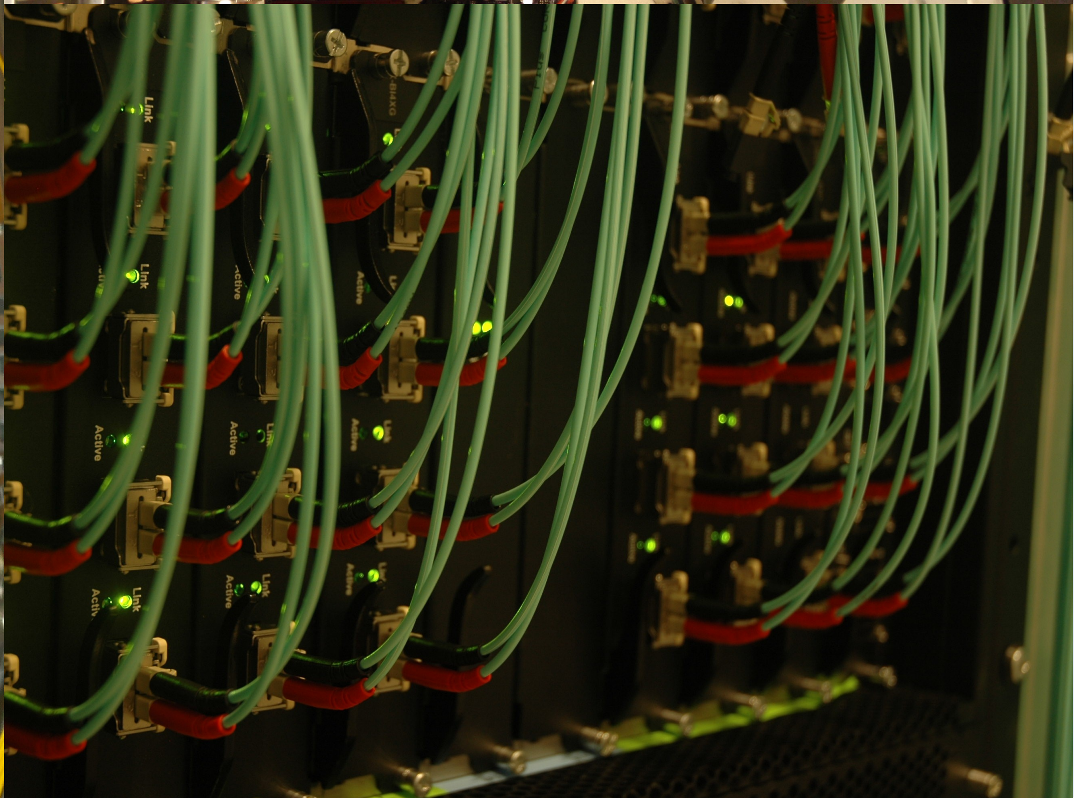
- Final part of LOFAR procurement
- Completed successfully December 2010
- Installation February 2011
- Declared operational April 2011
- 4x capacity previous cluster
  - Combines storage and compute in single node
  - Introduces low latency interconnect (QDR IB)
- Located 'offsite'

CEP Phase II cluster

- 100 Hybrid compute/storage nodes
- 24 AMD Opteron 6172 cores per node
  - @ 2.1 GHz
- 64 GB memory per node (2.67 GB/core)
- 12 2TB Disks per node
  - 20 TB usable disk space per node
- 20.6 Tflops peak performance
- 2 PB storage capacity
- Ubuntu 10.04 LTS

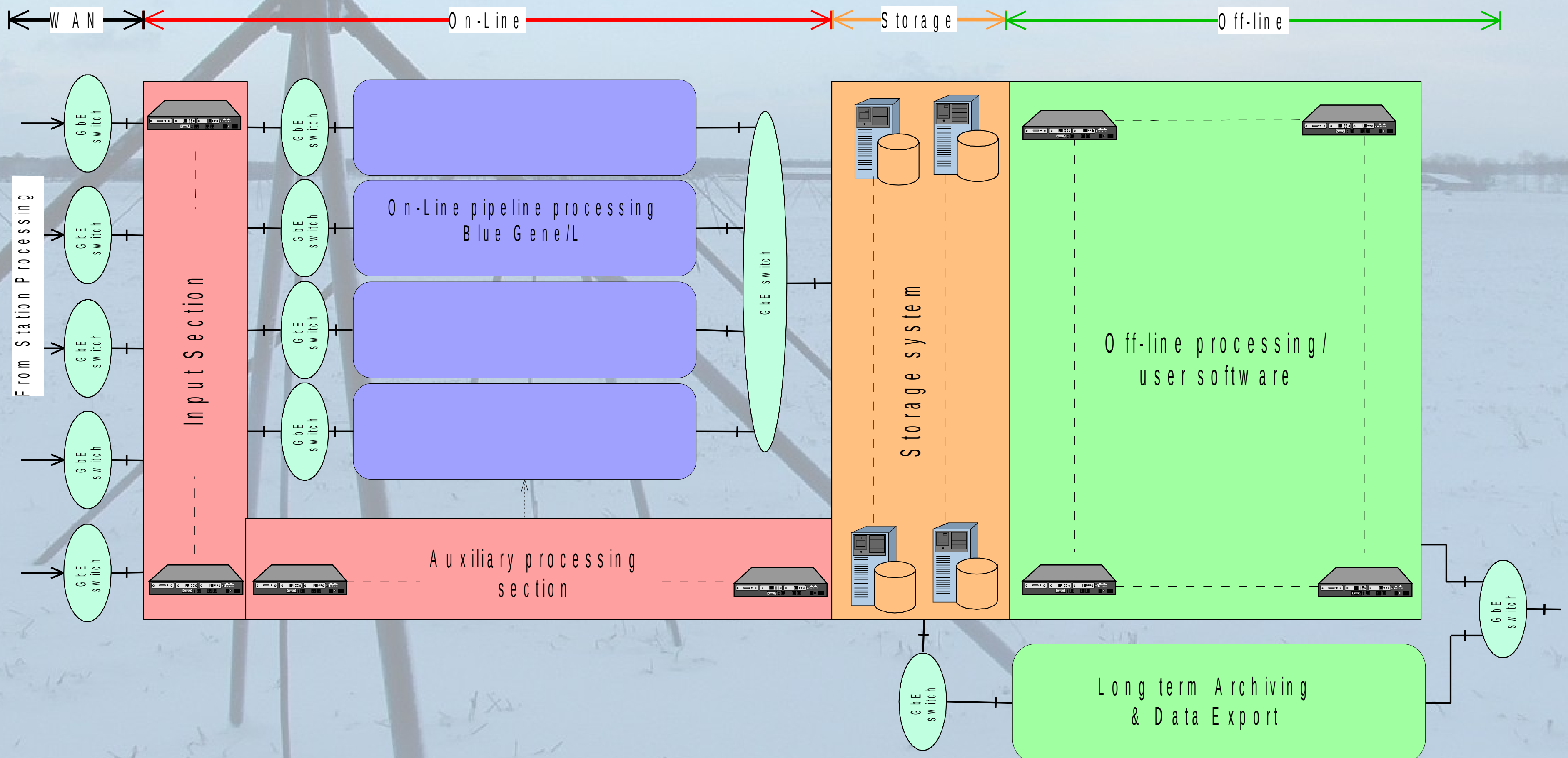



ASUS
ASUS


Router 1
Router 2

500W

Link
Active

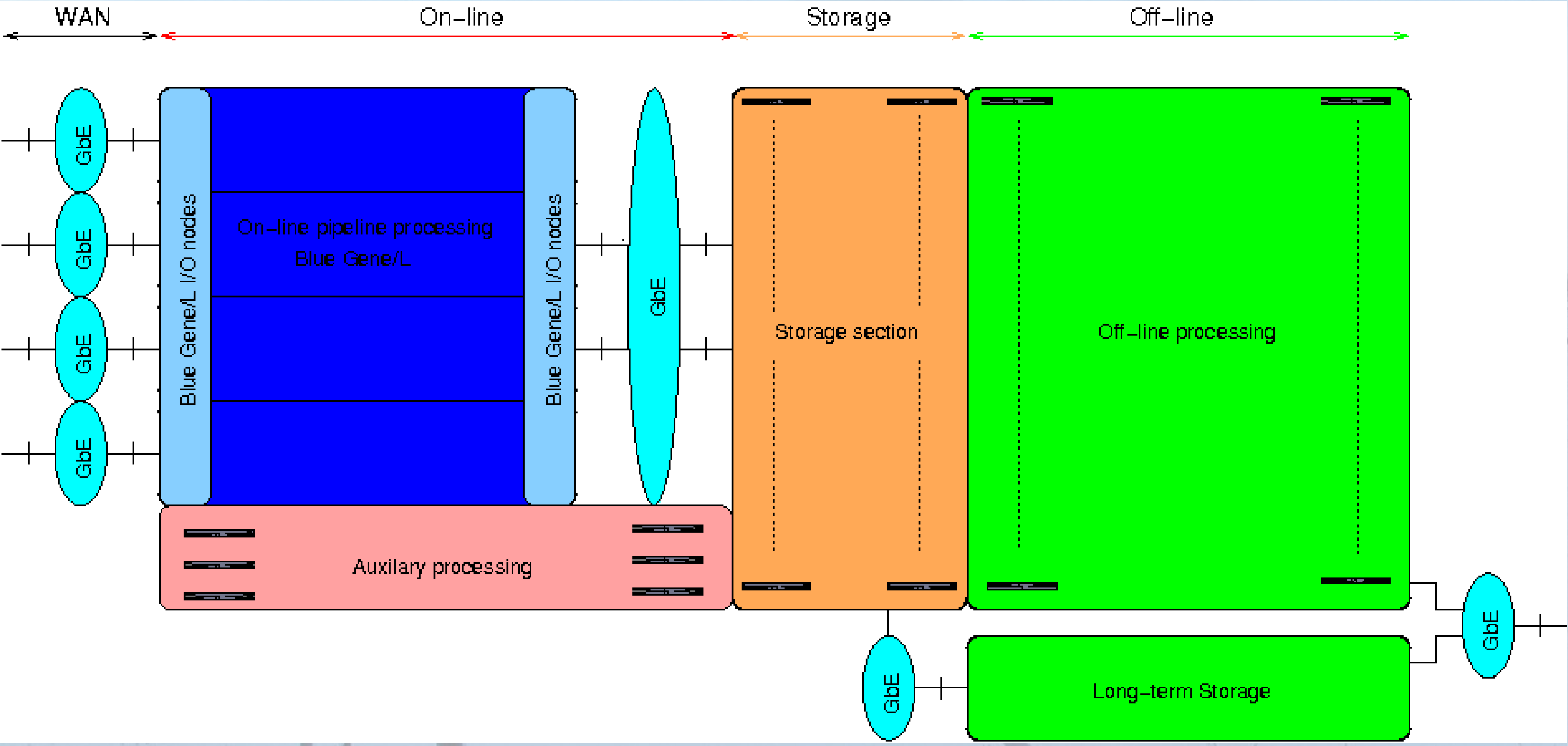
The evolution of the LOFAR Central Processor design

Evolution of CEP design (2005)

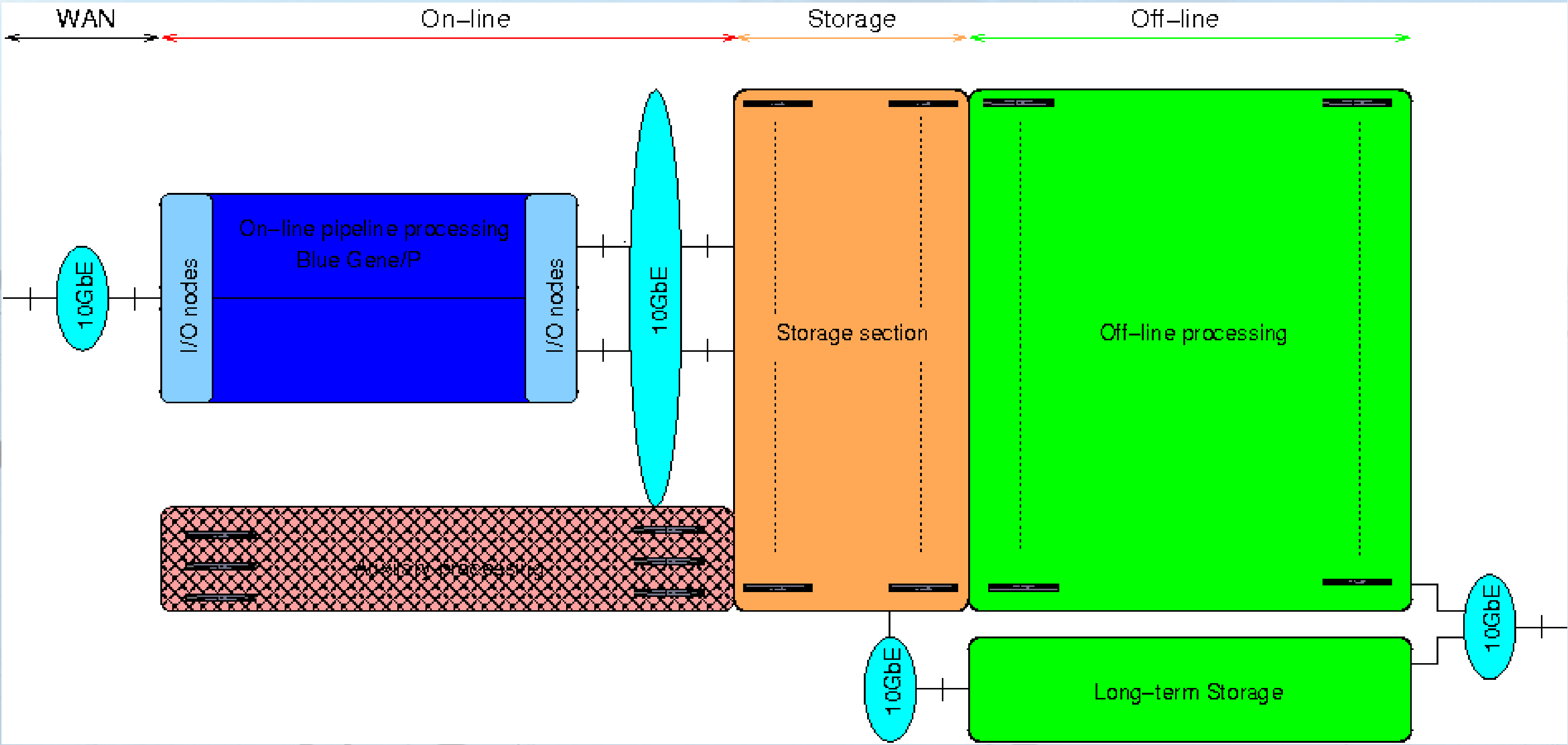


ASTRON LOFAR NWO

Evolution of CEP design (2007)

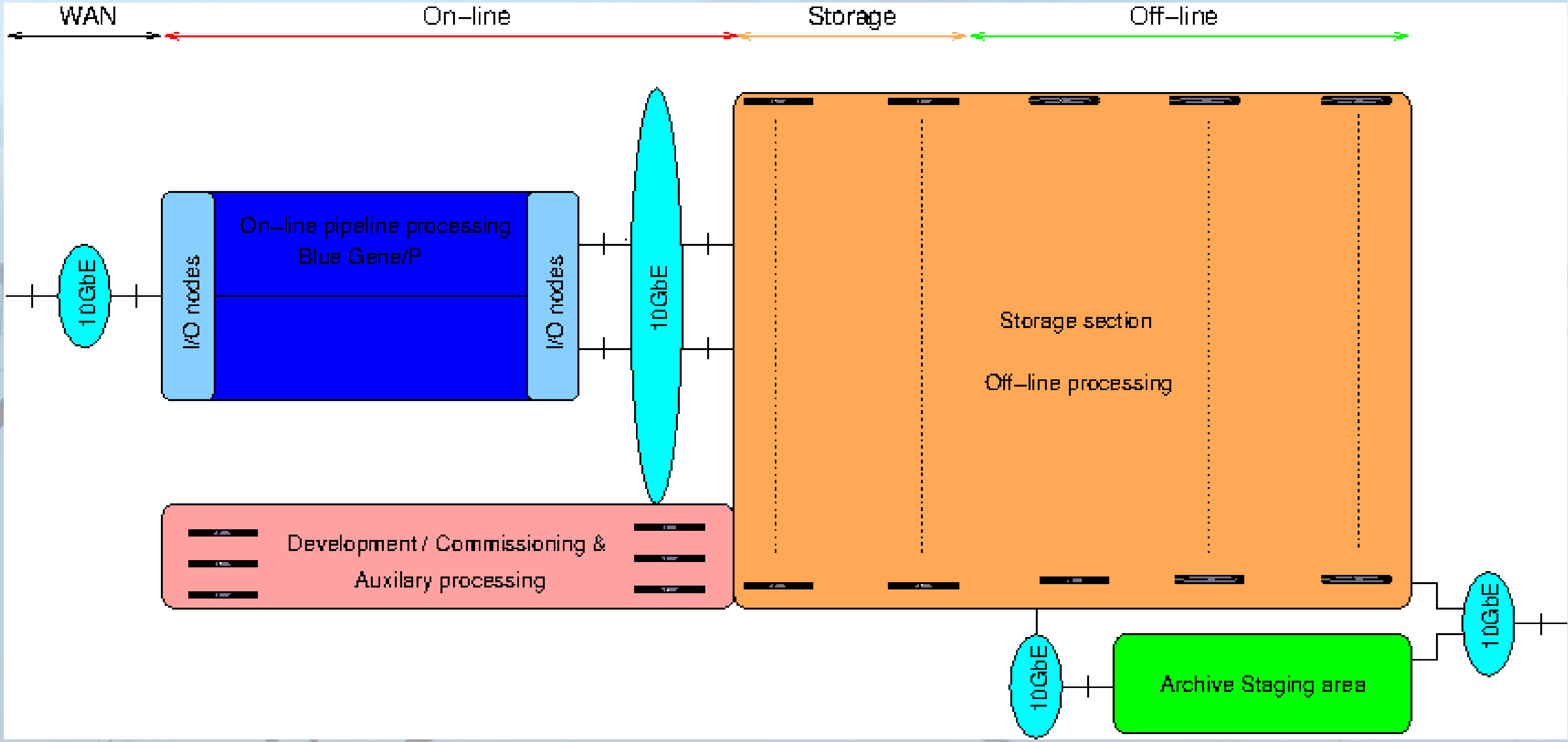


Evolution of CEP design (2008)



ASTRON LOFAR NWO

Evolution of CEP design (2010)



Lessons learned

- 4 iterations, each more integrated
- Increased efficiency by combining tasks
- Move from compute centric to data centric
- Increased awareness of data flow
- We effectively run several tasks on a single node
  - Requires careful tuning of node OS
  - Some application code added to facilitate
  - Linux Capabilities

ASTRON LOFAR NWO

Technology trends

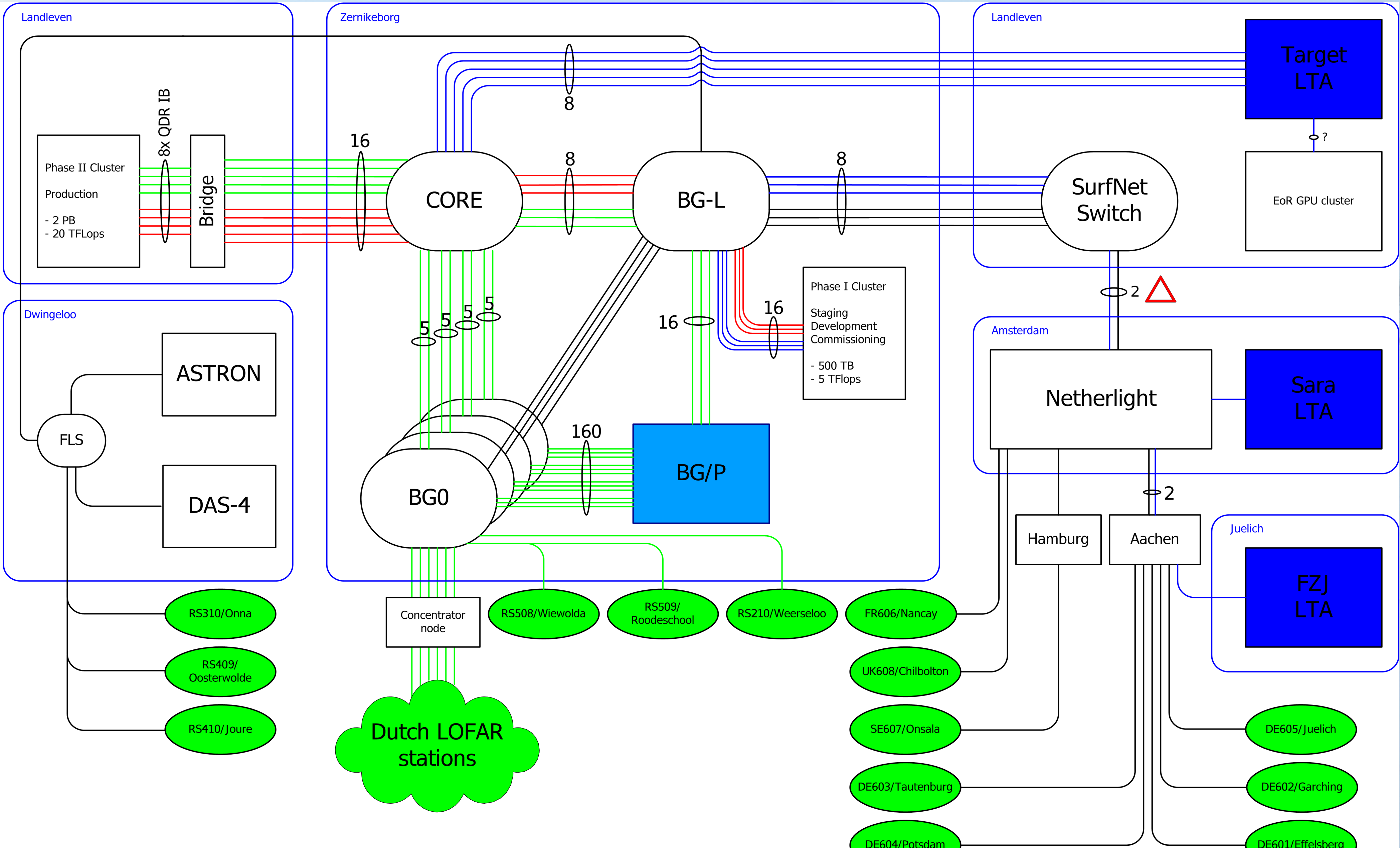
- Tremendous increase in cores per node
  - 2 in 2006, 24 in 2010
- Capabilities of a node improved a lot
  - Need to combine tasks on a node to keep it busy
- Memory size keeping up, speed not so much
  - 4 GB – 64 GB --- 2.67 GB/s – 10.67 GB/s
- Performance gap compute – storage
  - Bring storage to compute resources

ASTRON LOFAR NWO

The world according to Chris

- Shows all high-bandwidth (> GbE) links
- Focusses on central processor
- Gives a highly simplified but clear overview of the LOFAR system
- Shows data flow from station to archive

The world according to Chris



TITLE	VERSION	DESCRIPTION

Looking beyond the LOFAR Central Processor

Amdahl's laws

Gene Amdahl (1965): laws for a balanced system:

I. Parallelism: Max speedup is S/(S+P)

II. One bit of IO/sec per instruction/sec (BW)

III. One byte of memory per instruction/sec (Mem)

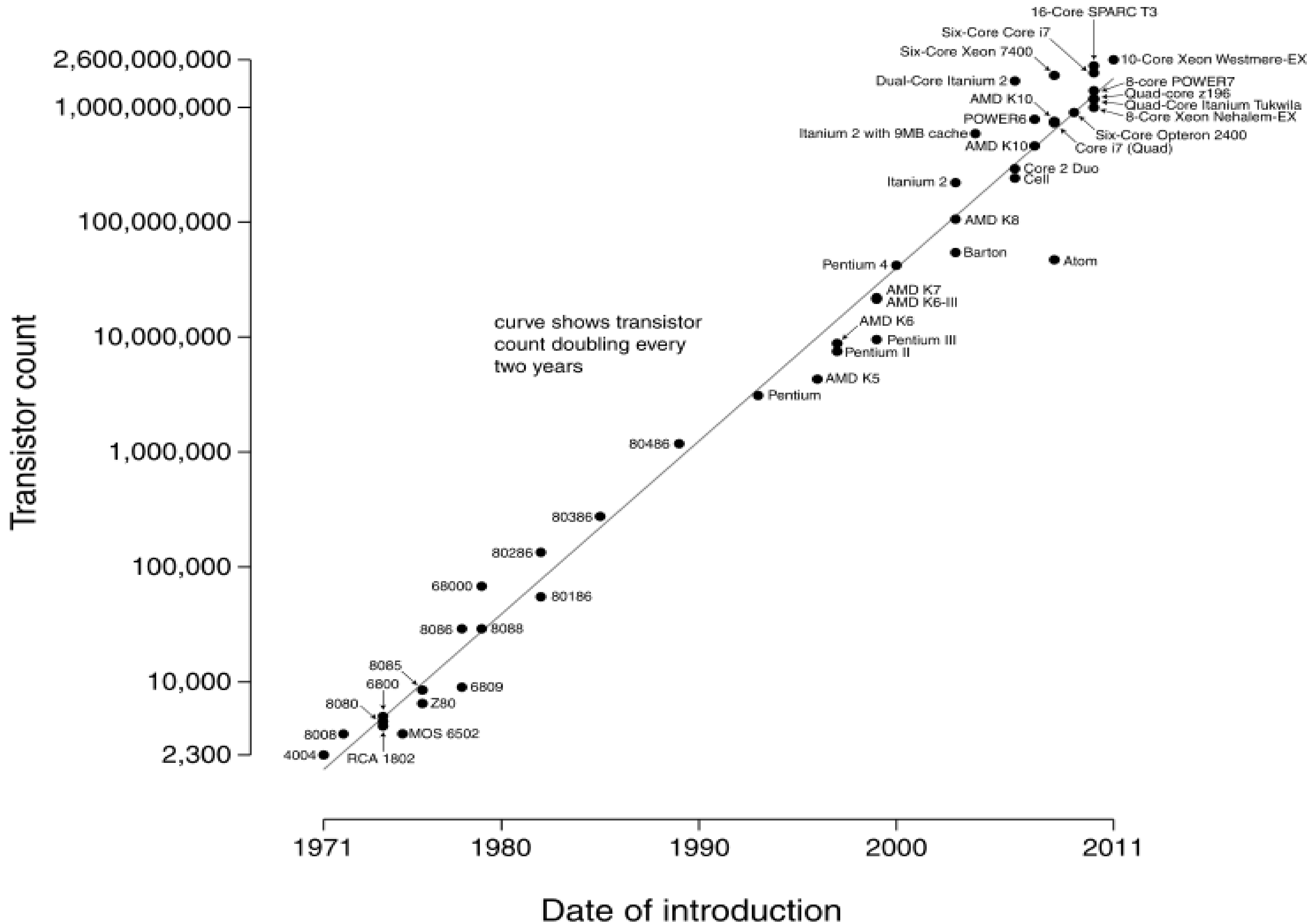
Modern multi-core systems move further away from Amdahl's laws (Bell, Gray & Szalay 2006)

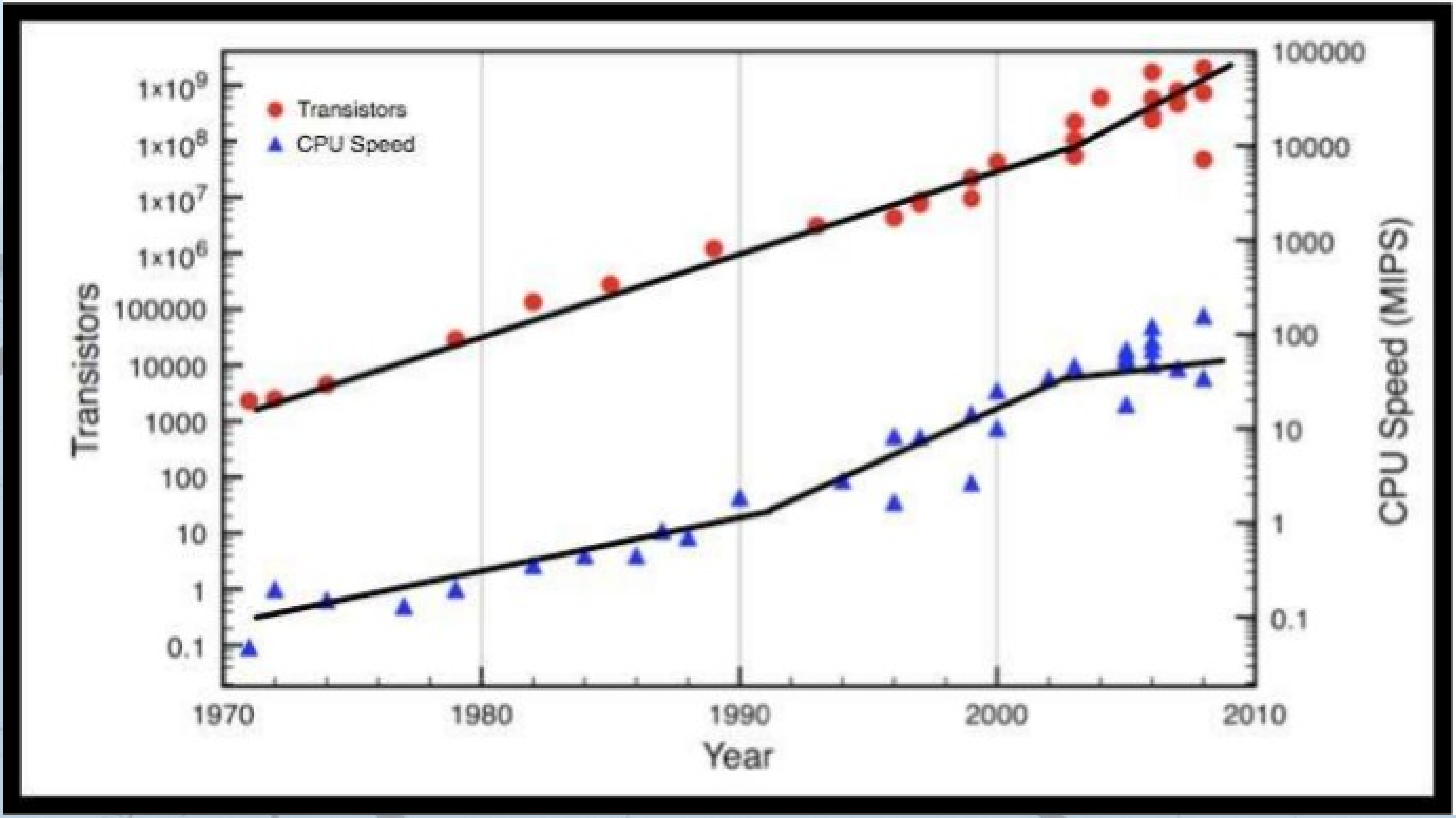
Exascale=GigaHz KiloCore MegaNode

Systems	2009	2018	Difference Today & 2018
System peak	**2 Pflop/s**	**1 Eflop/s**	**O(1000)**
Power	**6 MW**	**~20 MW**	
System memory	0.3 PB	32 - 64 PB [**.03 Bytes/Flop**]	O(100)
Node performance	125 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s [**.002 Bytes/Flop**]	O(100)
Node concurrency	12	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion) [O(10) to O(100) for latency hiding]	O(10,000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO	0.2 TB	60 TB/s (how long to drain the machine)	O(100)
MTTI	days	O(1 day)	- O(10)

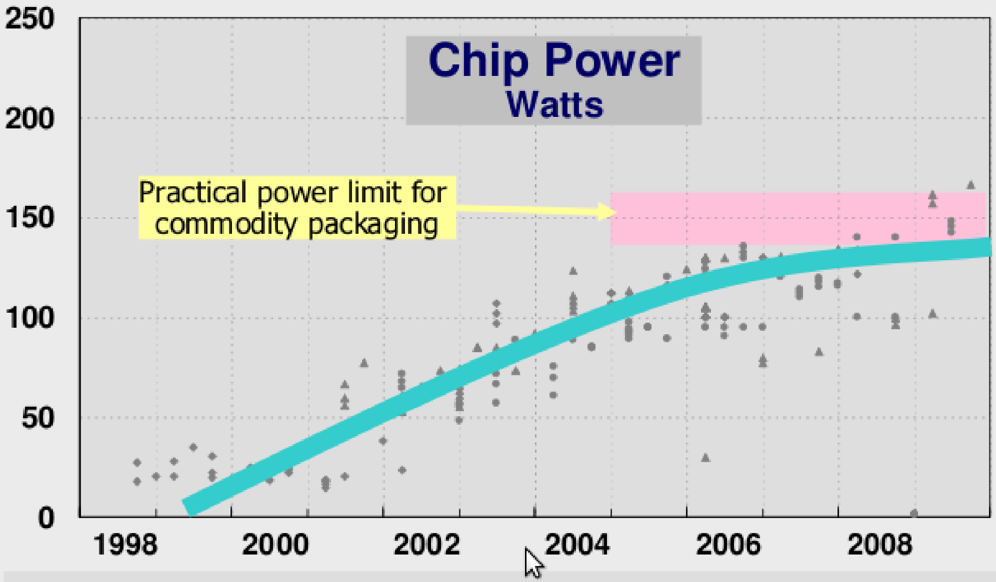
CSIRO

Microprocessor Transistor Counts 1971-2011 & Moore's Law



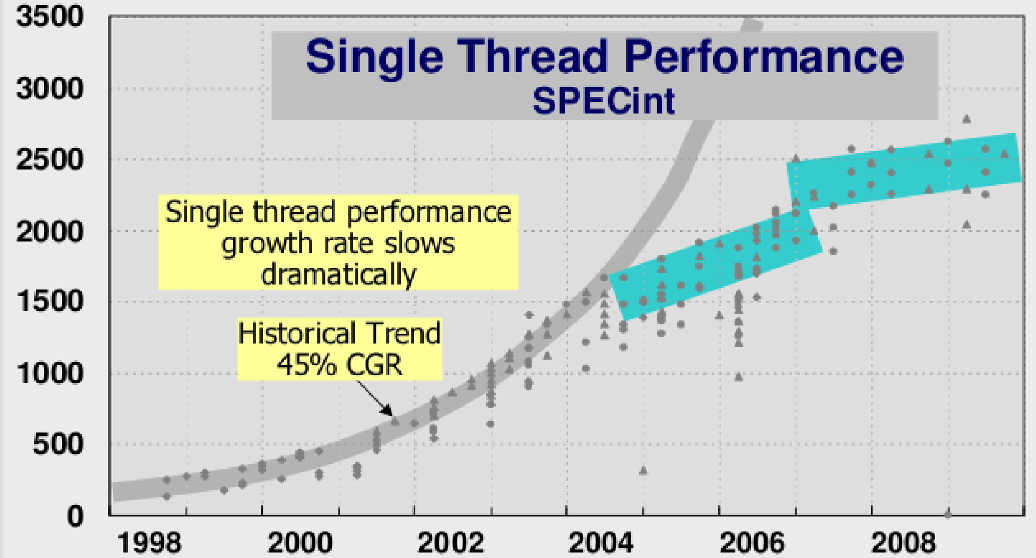


ASTRON LOFAR NWO



Chip Power has leveled off to packaging limit

Single thread performance has almost leveled off too



ASTRON LOFAR NWO

The 2-year cycle for technology shrinkage will continue, but the performance increase may slow beyond 22 nm

IBM GTO08

ASTRON LOFAR NWO

Computing ~2020

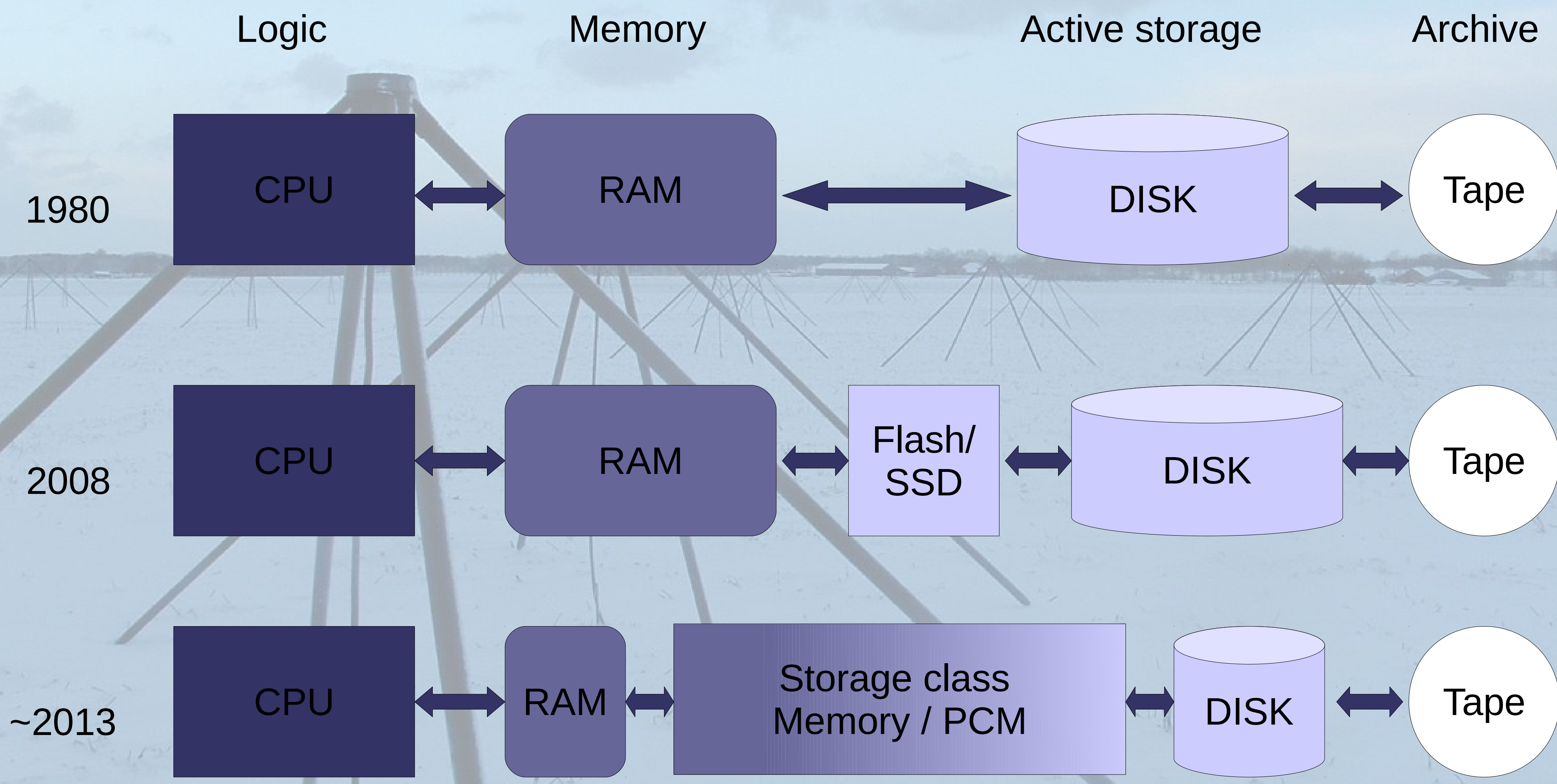
- Shrinkage will continue until at least 2020
- Reduced performance improvement expected
- Continued scaling through shrinkage <u>and</u> new technologies
  - 3D stacking
  - PCM
  - Hybrid
- A series of one-time solutions

ASTRON LOFAR NWO

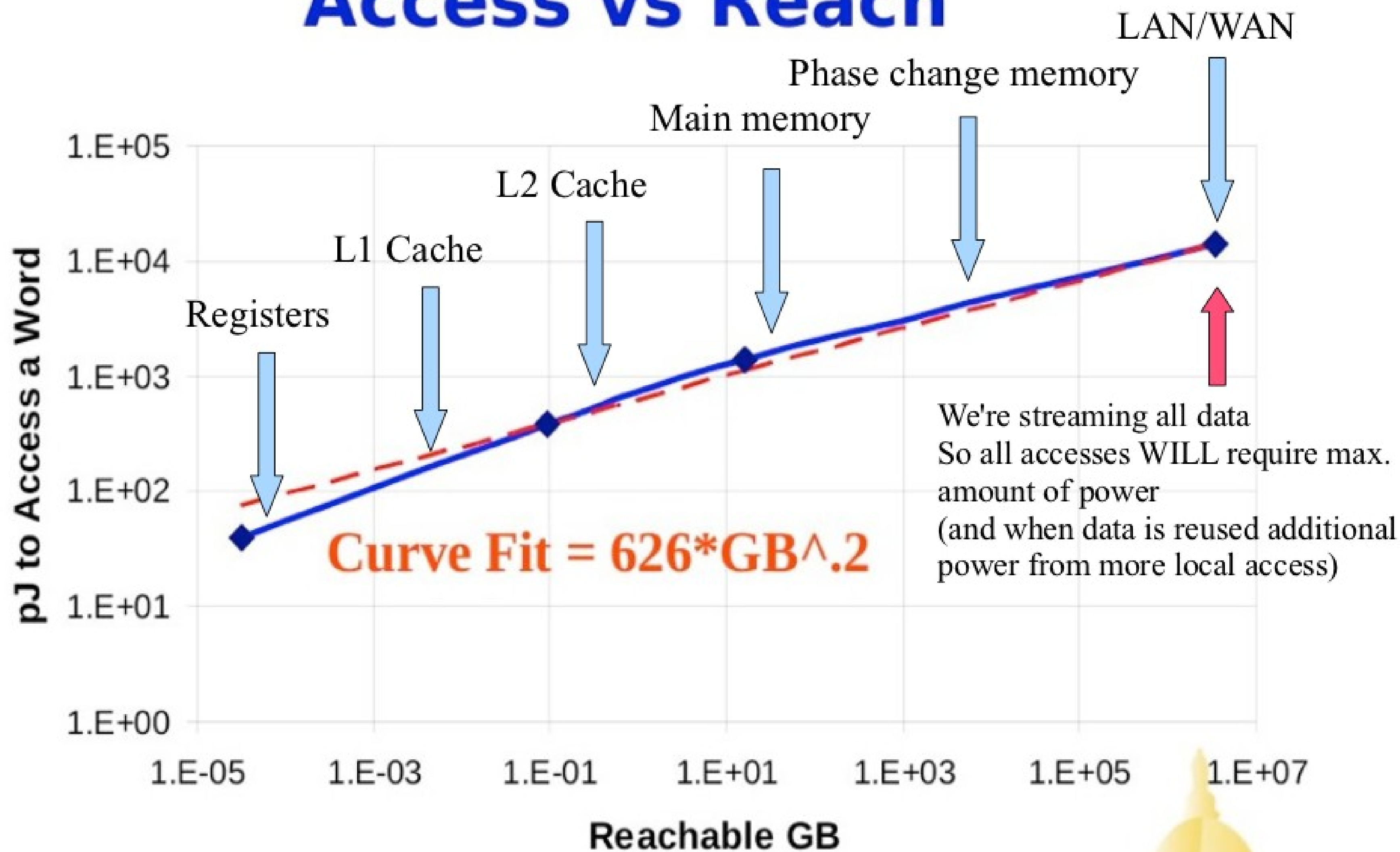
Computing ~2020

- Smaller gate size gives designer lots of real estate
- Possible developments
  - Many heterogeneous superscalar out of order cores
  - Many heterogeneous scalar in order cores
  - Heterogeneous cores (cpu / gpu combination)
  - Heterogeneous system with specialized accelerator cores
- Challenge industry to add accelerators we can use

Storage hierarchies



Access vs Reach



Source: Energy at ExaFlops, Peter M. Kogge, SC09 Exa Panel

Technology trends

- Many core architectures
  - GPUs – Nvidia Tesla and AMD Firestream
  - Intel Knights Ferry and Knights Corner
  - Tilera TileGX
- Low power alternatives
  - ARM based mobile solutions
- Hybrid CPUs
  - Combining large superscalar, out of order cores with smaller accelerator cores

ASTRON LOFAR NWO

Technology trends

- Growing memory gap
  - GPUs lead the way
- Continued trend towards many cores
  - Experience teaches software will lag behind
- Distinction CPU - GPU will disappear
  - (along with CUDA, OpenCL, etc)
- Power is going to dominate system design
  - Network power mostly ignored so far
  - Energy req'ed for I/O is going to dominate CPU

Conclusions

- Compute power will essentially be free ~2020
- Node configuration unclear
  - But quite different to current hardware
  - Likely heterogeneous
- Challenge: how to use this power efficiently
- Algorithm design is required to follow hardware
- Power consumption will guide design
  - System and algorithm development