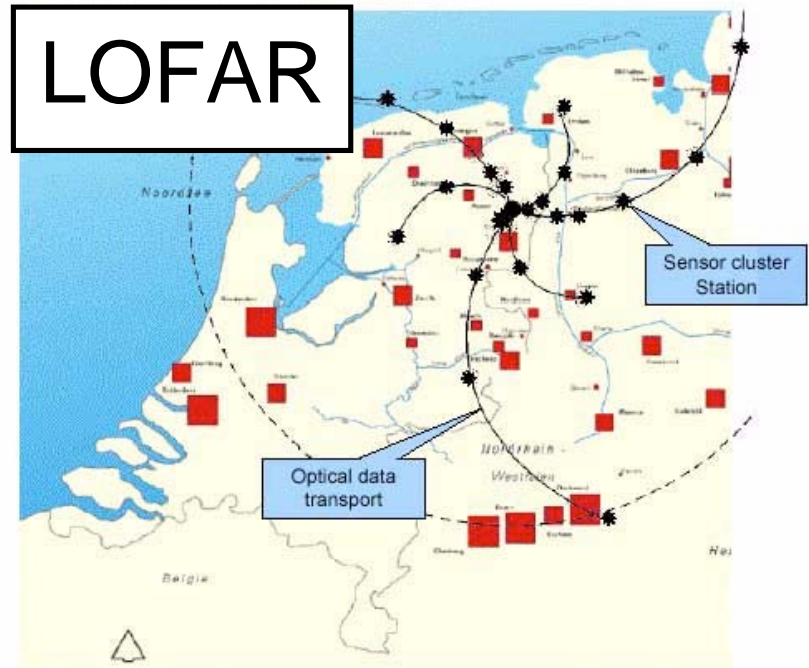


# Use of BlueGene/L in LOFAR

Bruce Elmegreen  
IBM Watson Research Center  
914 945 2448  
bge@us.ibm.com



16 racks of BG/L, 71 Tf  
#1 in the world



# LOFAR is Unique

1 300 km baseline with streaming data processing (VLBI meets WSRT)

- country-wide science project, very public, IT good for infrastructure

2 Enormous data rates (~20 Tbps from antennae, 300 Gbps from stations)

- must process data as it streams and then discard.

3 Low frequency: 10-250 MHz

- HI prior to recombination ( $z=10$ ), pulsars, cosmic ray showers, galactic and extragalactic synchrotron, solar wind, ...

4 Terrestrial and ionospheric noise

- requires mitigation of bad channels, nulling of Terrestrial transmitter directions, active response to ionospheric "seeing" -- must clean signals before correlation

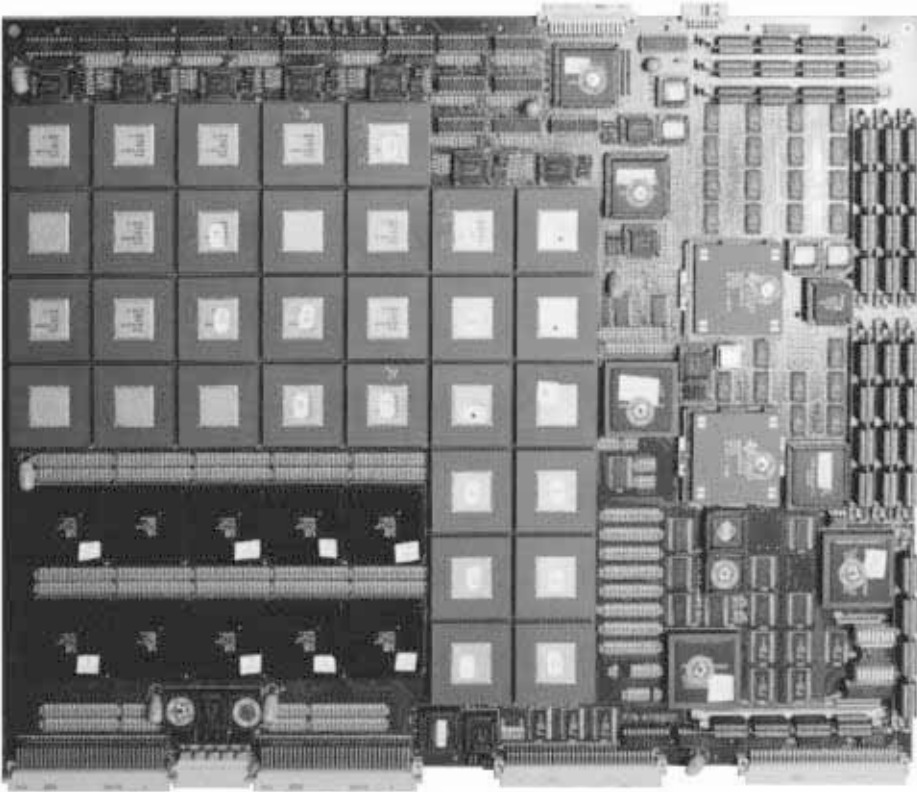
5 Each station is a phased dipole array, not a parabolic dish

- low cost, no moving parts, able to "point" in multiple directions
- stations sum dipole signals, producing 12-16 bit datawords

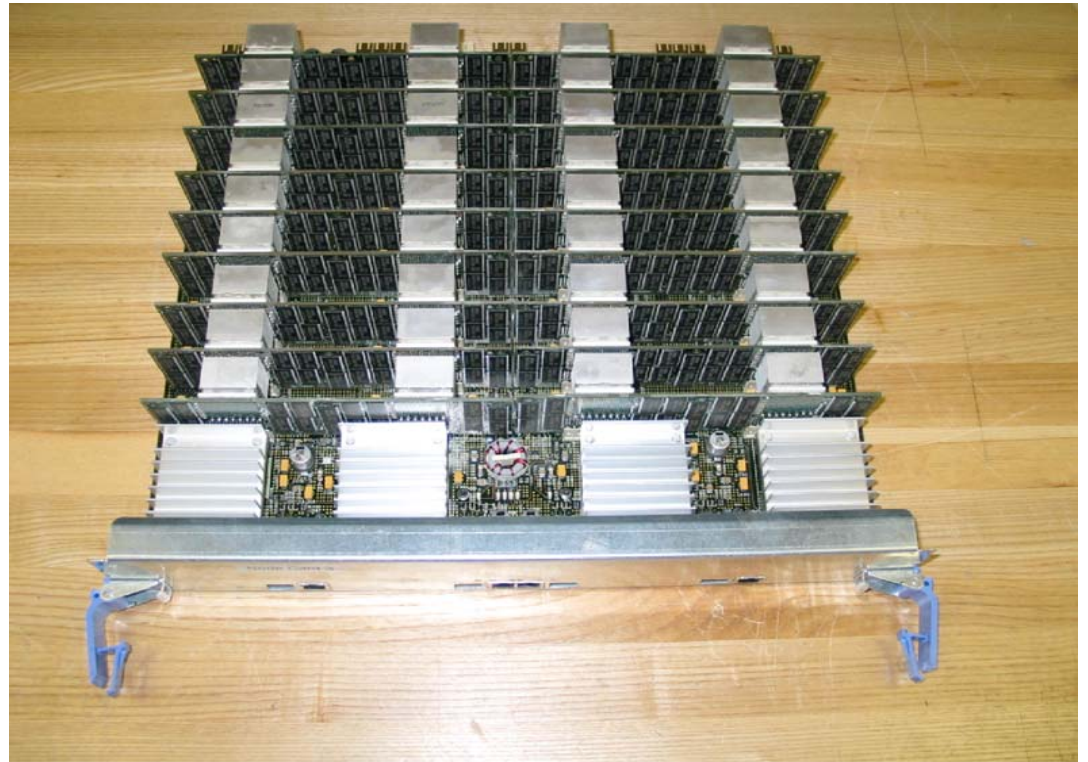
# IT Revolution Has Made This Possible

- 1 High bandpass optical fiber makes 300 km baseline possible
  - National image highlights the role of the central processing facility
    - Important for the next generation of scientists and computer scientists
- 2 High data rates require a central computer with enormous IO, fast node-to-node communications, and fault tolerance
- 3 New frequency range guarantees breakthrough results, minimizing the scientific risks inherent in the innovative design.
- 4 Noise mitigation, nulling, ionospheric corrections etc. require much additional processing and post-construction flexibility
  - 12 or 16-bit datawords increase to 24-32 bits after correlation, and 34-42 bits after 1000-step time sum, requiring "normal" processors
    - harder to make special purpose correlators, but gain from COTS costs
  - Data buffering in memory is essential
- 5 Phased arrays require high-tech chips and processors in the field too.

# The Processor Revolution is Part of This



32-processor Mark IV digital correlator (MIT, Jodrell Bank, ASTRON)



32-node BlueGene/L board with

1. 64x64 bit comp. prod every 2 clock cycles
2. Four Gbps ethernet IOs
3. One chip type (dual core PowerPC)

# BlueGene/L does Fast Digital Correlations

- For sky maps, cross correlate 110 LOFAR Station inputs,
  - $110 \times 110 / 2$  product pairs (including autocorrelation)
    - \* 32,000 channels per millisecond
    - \* 4 polarization cross products
  - giving 0.8 Tera ( $10^{12}$ ) complex products/second
- For Epoch of Recombination using the LOFAR Central Core
  - $64 \times 64 / 2$  pairs \* 32,000 ch/ms \* 4 pol. prod. \* 5 beams  
= 1.3 Tera complex products/second
- One rack of BlueGene/L can run 2048 processors in "Virtual Node Mode" performing  $2048 \times 0.5 \times 700M$  64-bit complex products per second  
= 0.7 Tera complex products/second \* some efficiency (50-80%)
- Allows one half-rack per beam for eight beams with one rack for ionospheric corrections and one rack for transient follow up if needed.

# BlueGene/L does Required IO

- 110 LOFAR Station inputs @ 1-2 Gbps is ~200 Gbps input
- 64 EoR central core inputs @ 5 Gbps is ~320 Gbps input
- Each BlueGene rack has 128 Gbps ethernet IO connections  
@ >70% expected (measured) efficiency into 4 racks, that is 360 Gbps
- IO nodes connected to compute nodes at 2.8 Gbps bi-directional
- Internal communication (all-to-all, broadcasting ionospheric corrections, ...) has 1.4 Gbps both ways in 6 dimensions (=350 Mbps in each dimension)
  - Time for all to all = Number of Bytes \* Number of hops /350 MBps
    - Average number of hops = 1/4 longest torus dimension = 2 in midplane
    - much less than compute times and IO times (< 1 percent)
  - allows "Virtual Node Mode" use of processors for double-speed computing

# BlueGene/L is Reliable

- Few building blocks (PowerPC for IO and compute, linkchip for midplane connectivity)
- Easy to assemble, homogeneous & scalable
  - Modular construction means parts are easily replaced
- System on a chip design
- Simple operating system on compute nodes (no daemons, interrupts, etc.)
- Redundant power supplies, fans, DRAM bits
- ECC or parity/retry on most busses
- Extensive data logging (voltage, temp, recoverable errors) and failure forecasting

# SUMMARY

- Innovative LOFAR design requires high IO, high compute speeds, fault tolerance, float and double words, software flexibility, ease of programming
- BlueGene/L is the most compact design ever for a supercomputer
  - with the lowest power consumption (~12 W/CPU total)
  - in a LINUX environment (mostly\*), optimized for MPI, based on PowerPC chips (COTS advantage),
    - designed for high efficiency FFTs, complex products
    - no LINUX on compute nodes (forks, shell creation, ...)
- All LOFAR specs matched and tested by BlueGene/L hardware
- BlueGene/P xyz extensions offer innovative options to meet SKA needs based on the same fundamental designs
  - requirements for SKA likely to exceed 10 Pflops