LOFAR on BlueGene/L

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

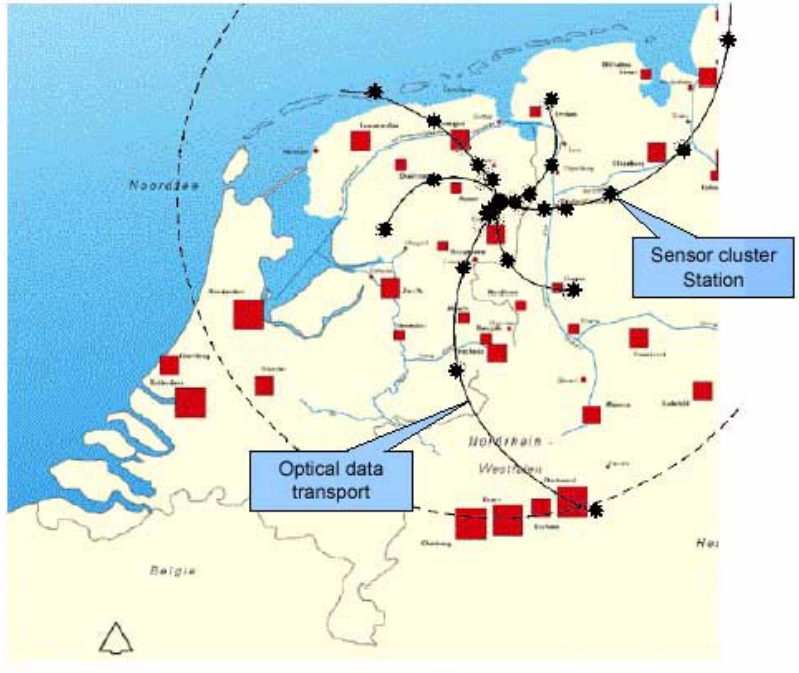
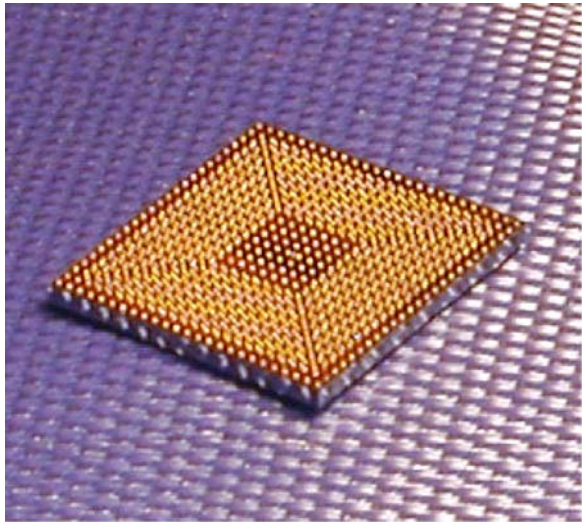
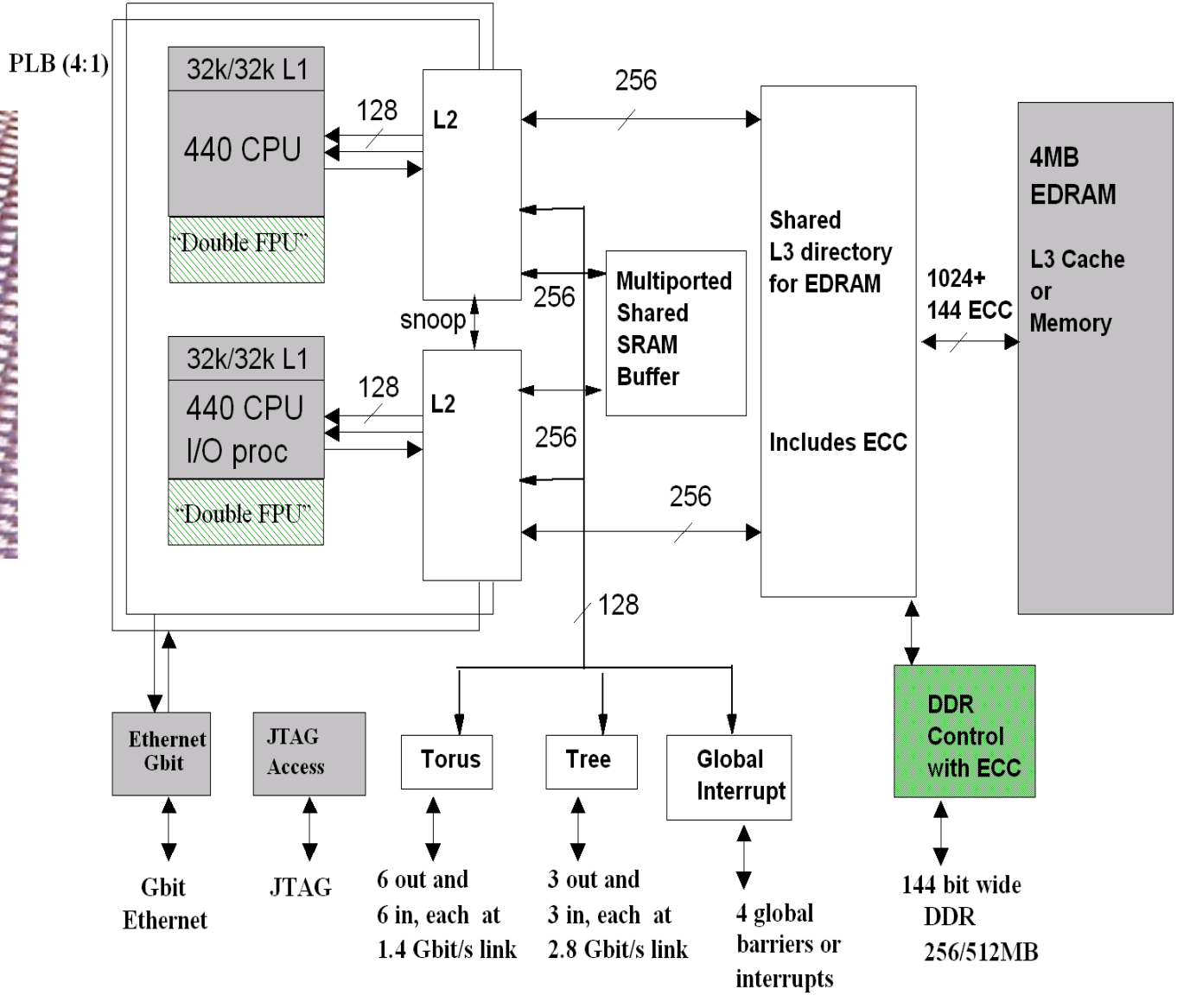


Figure 2.2 Schematic example configuration of the research infrastructure. The final configuration will be chosen in 2003, based on scientific, economic and environmental considerations.

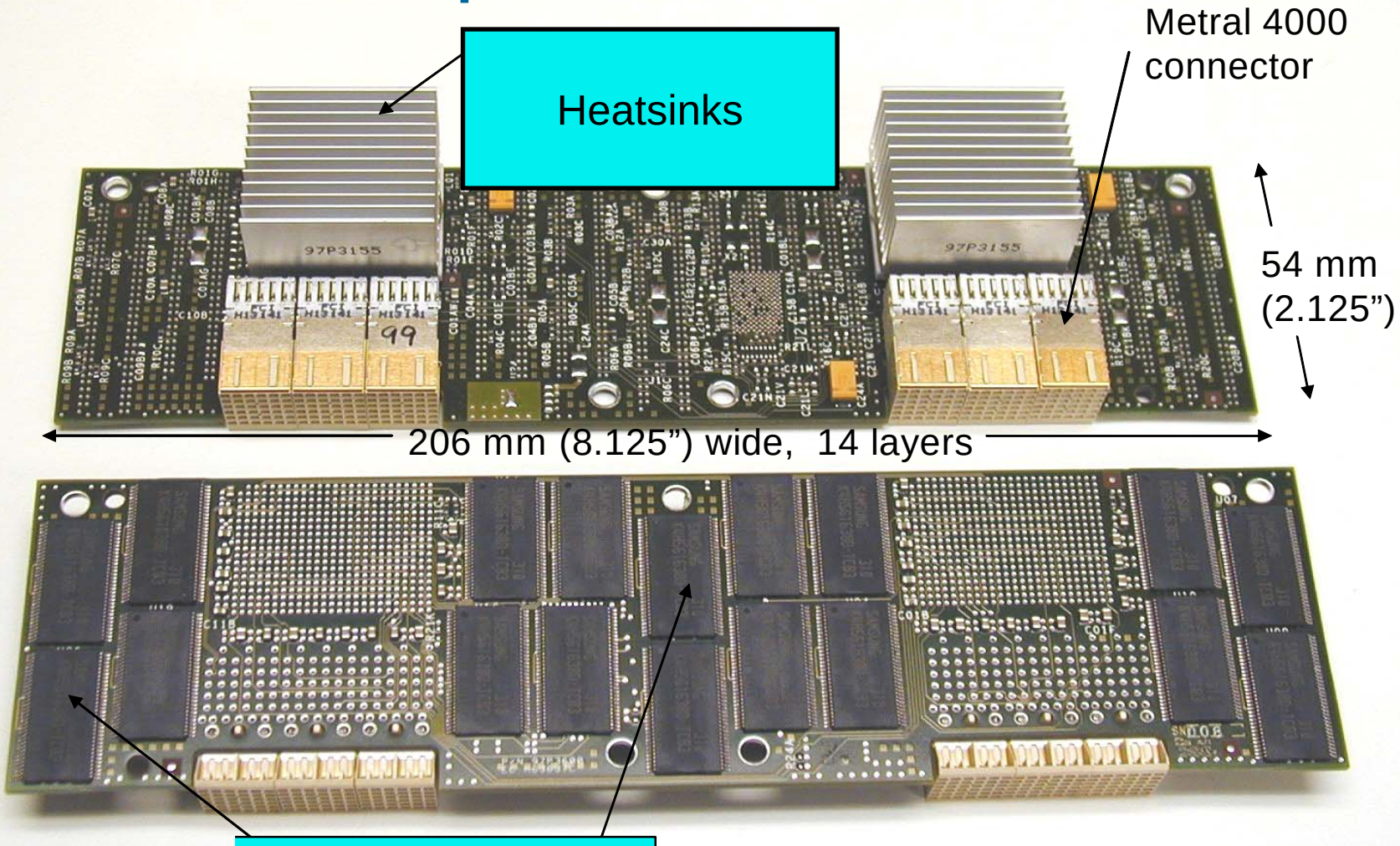
BlueGene/L chip



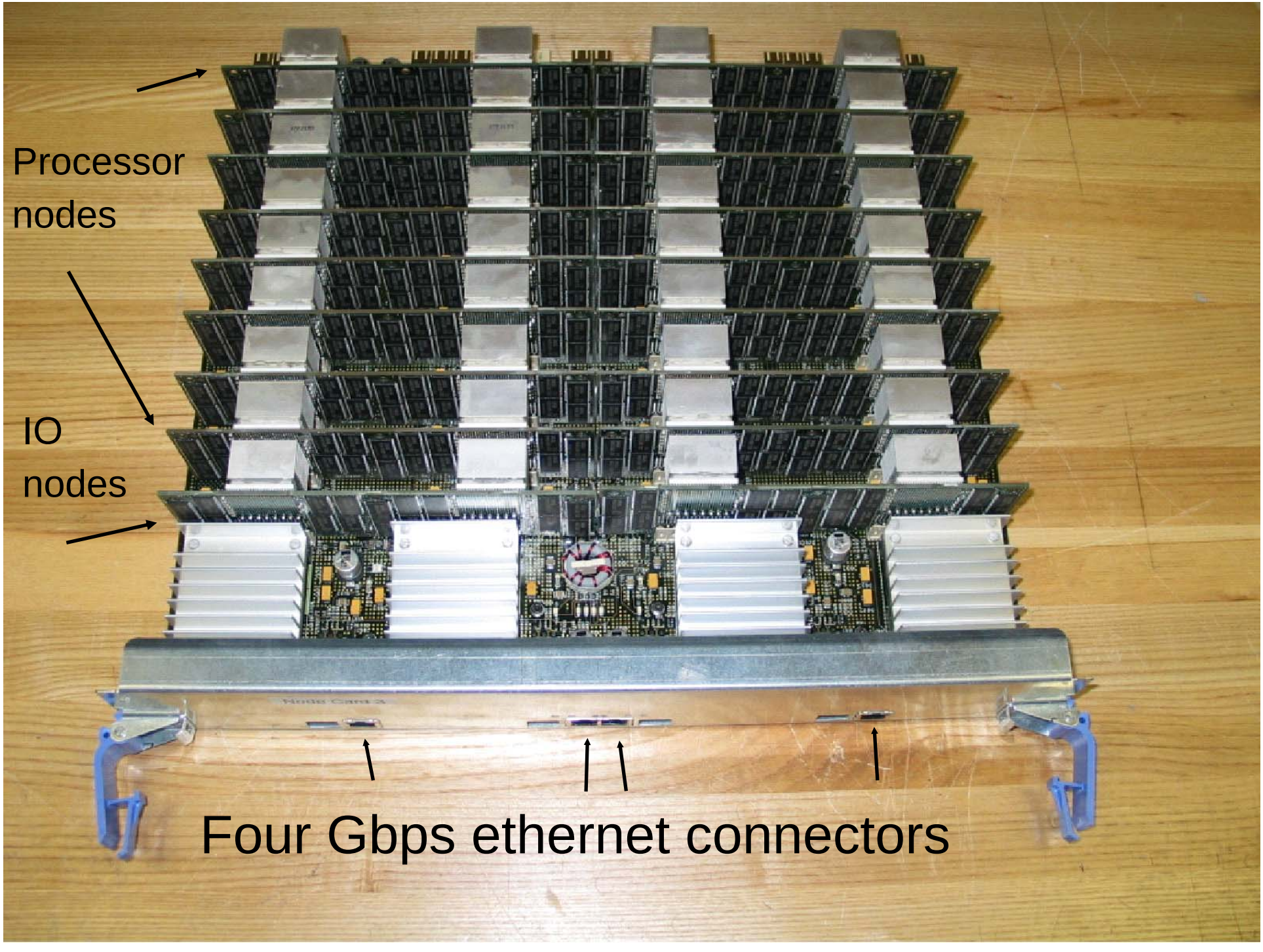
IBM CU-11, 0.13 µm
11 x 11 mm die size
25 x 32 mm CBGA
474 pins, 328 signal
1.5/2.5 Volt



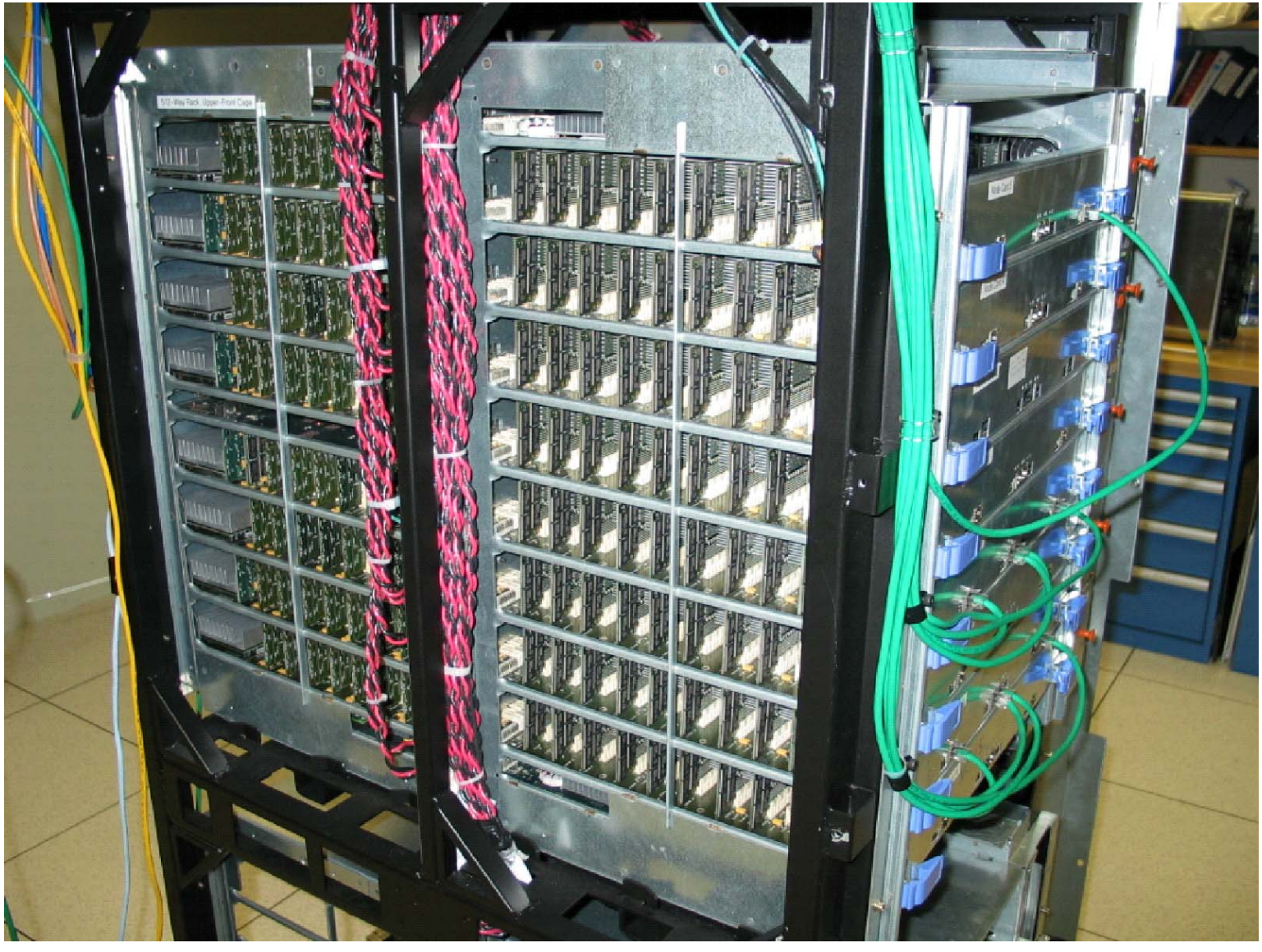
Dual Node Compute Card



0.5 GB RAM/node


Processor nodes
IO nodes
Four Gbps ethernet connectors

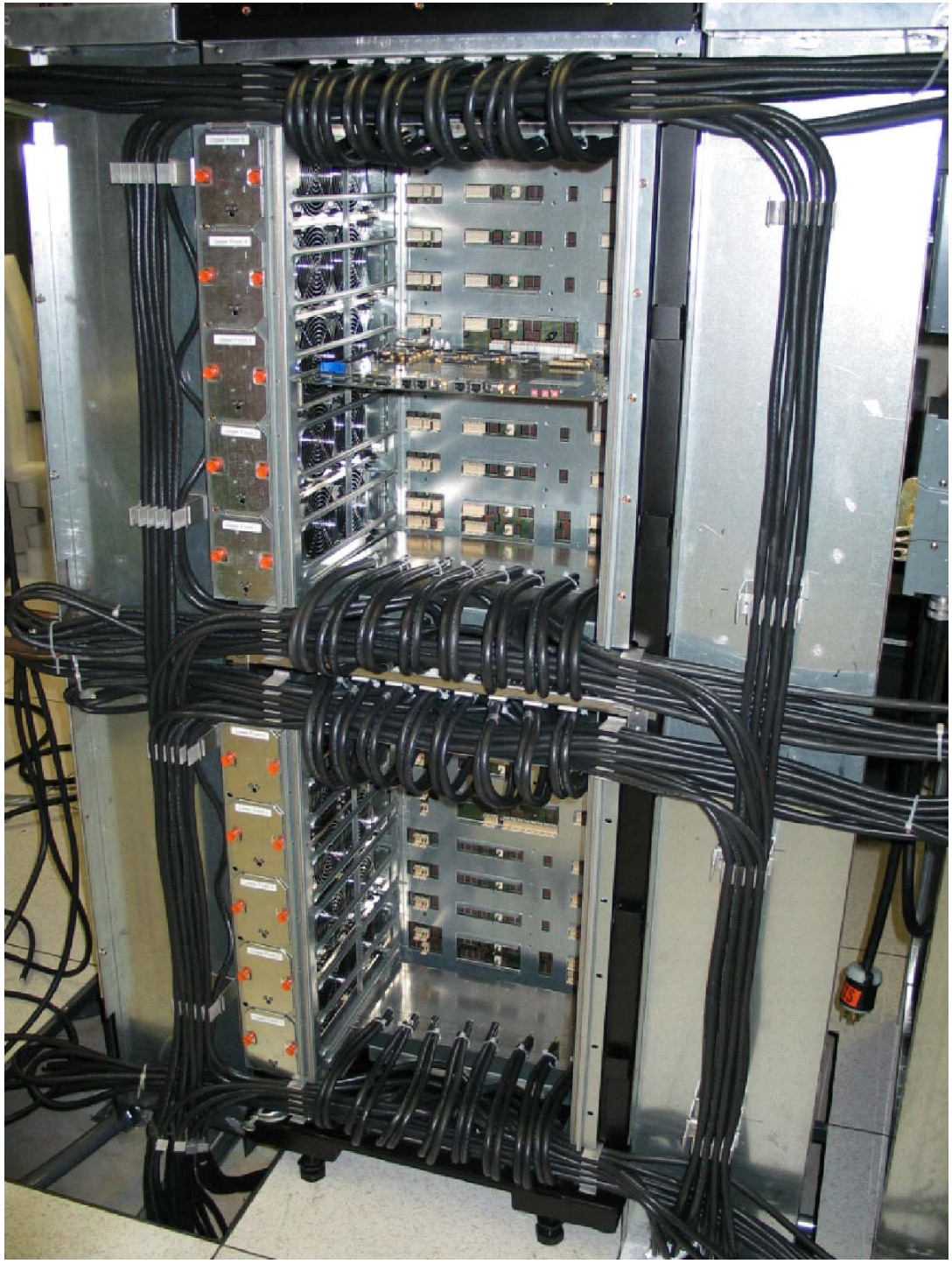
512 Way BG/L Prototype



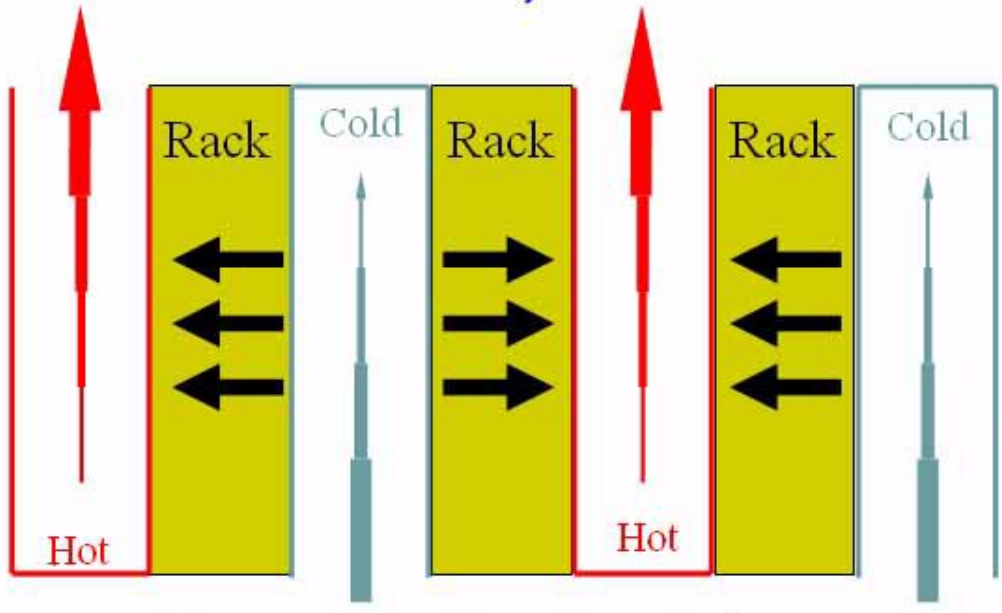
midplane = half rack

Cables for torus in 3 dimensions

6 racks have ~1 km of cables each 1/2" thick



BG/L L<->R airflow, direct from raised floor



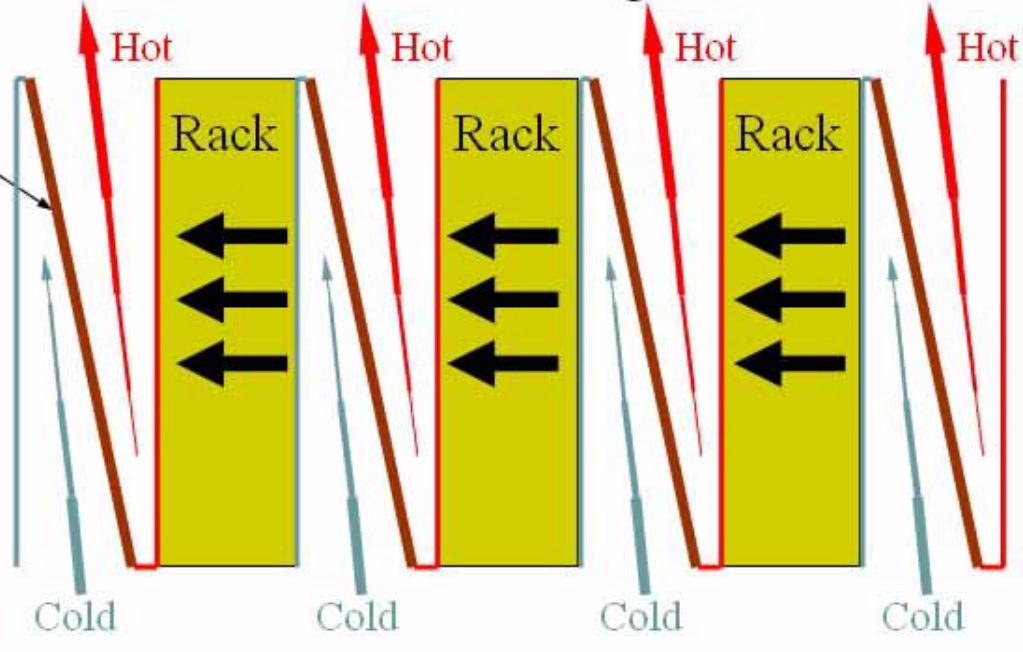
etc.

Flow rate in cold duct is largest at bottom; flow rate in hot duct is largest at top.

This scheme has same duct area, top to bottom, regardless of flow rate.

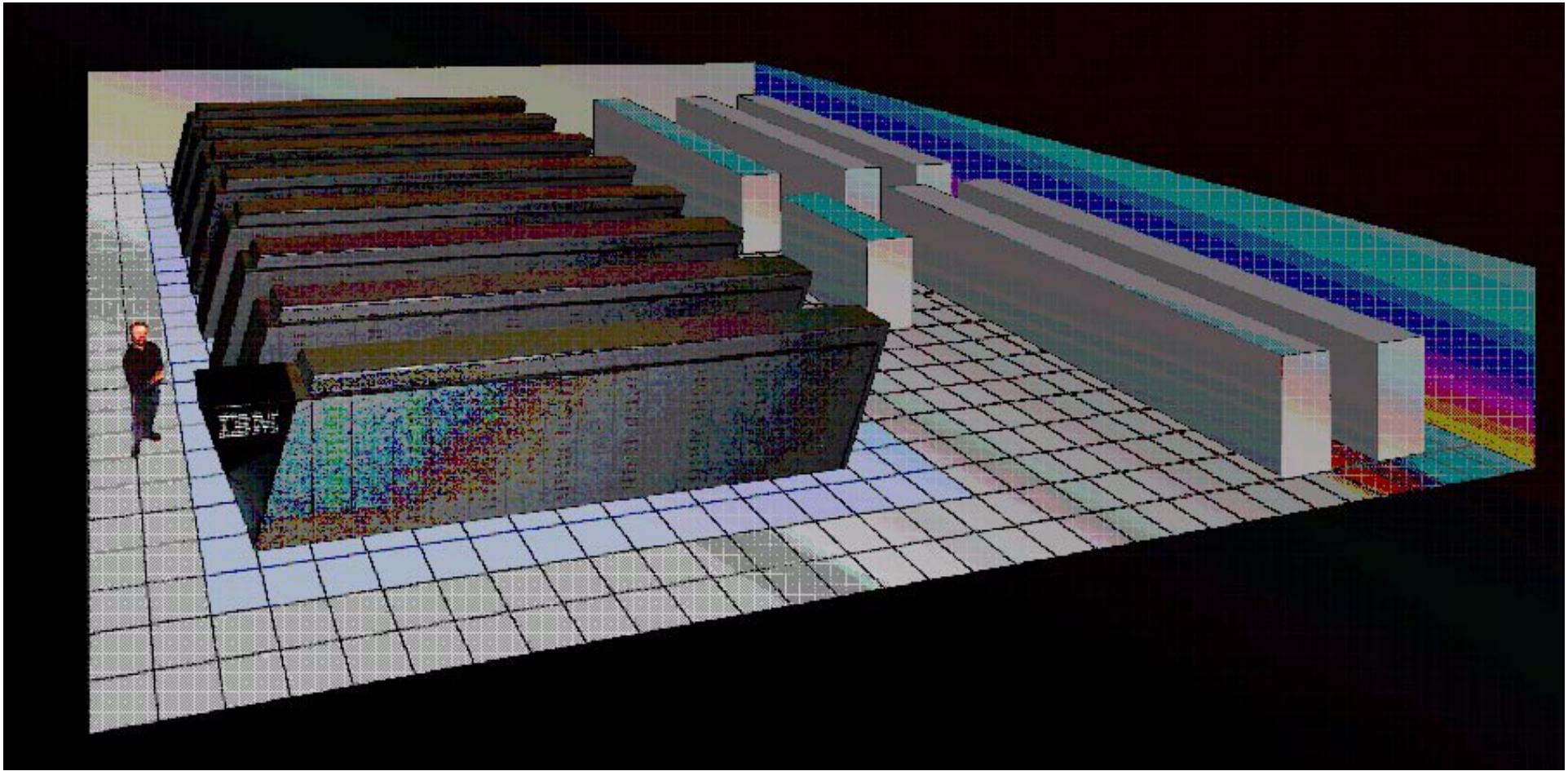
Alternative Ducting Scheme

Thermal-Insulating Baffle



etc.

Alternative Ducting: Ducts are larger where flow is greater (Tj ~10C lower)

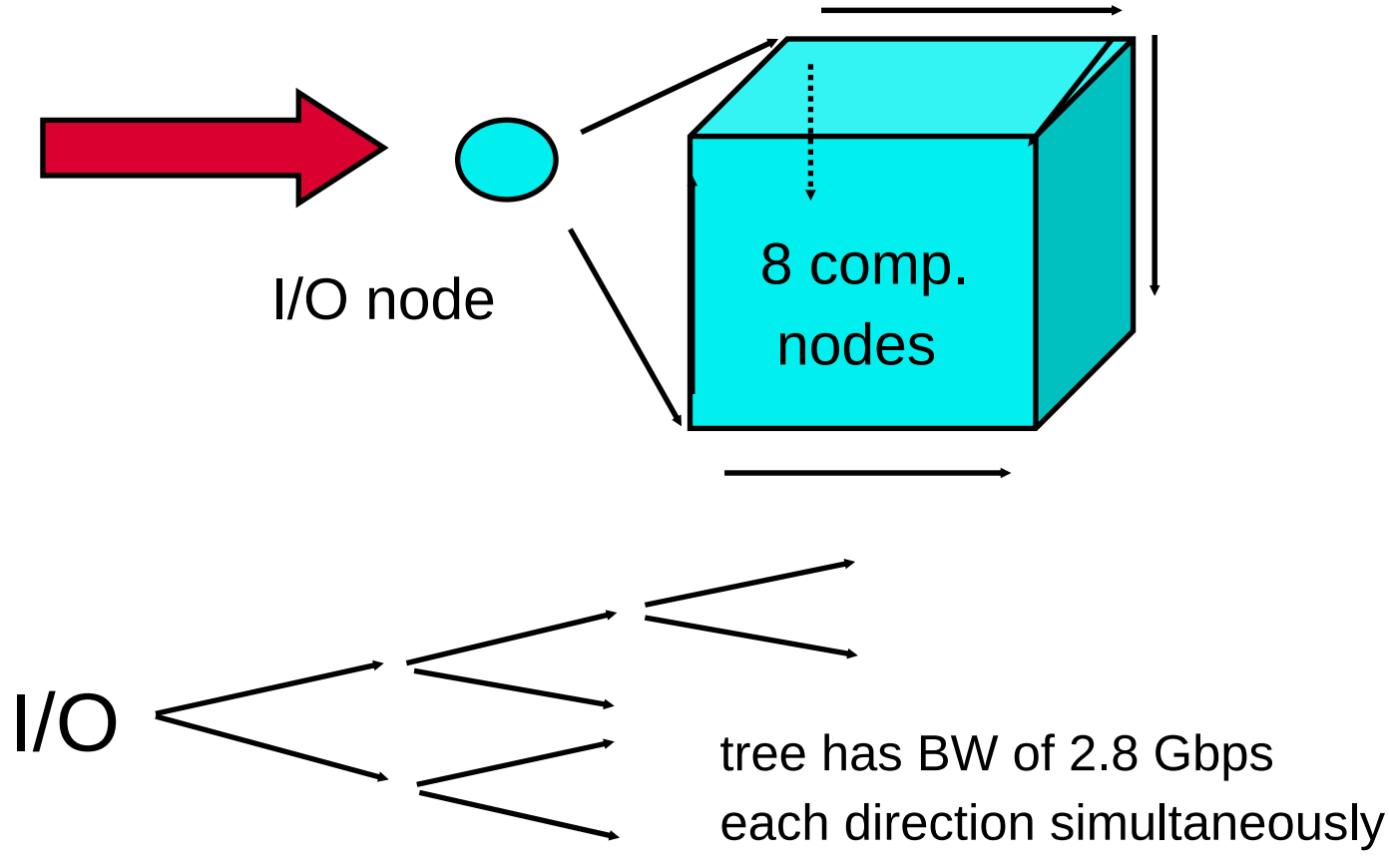


64 racks at LLNL

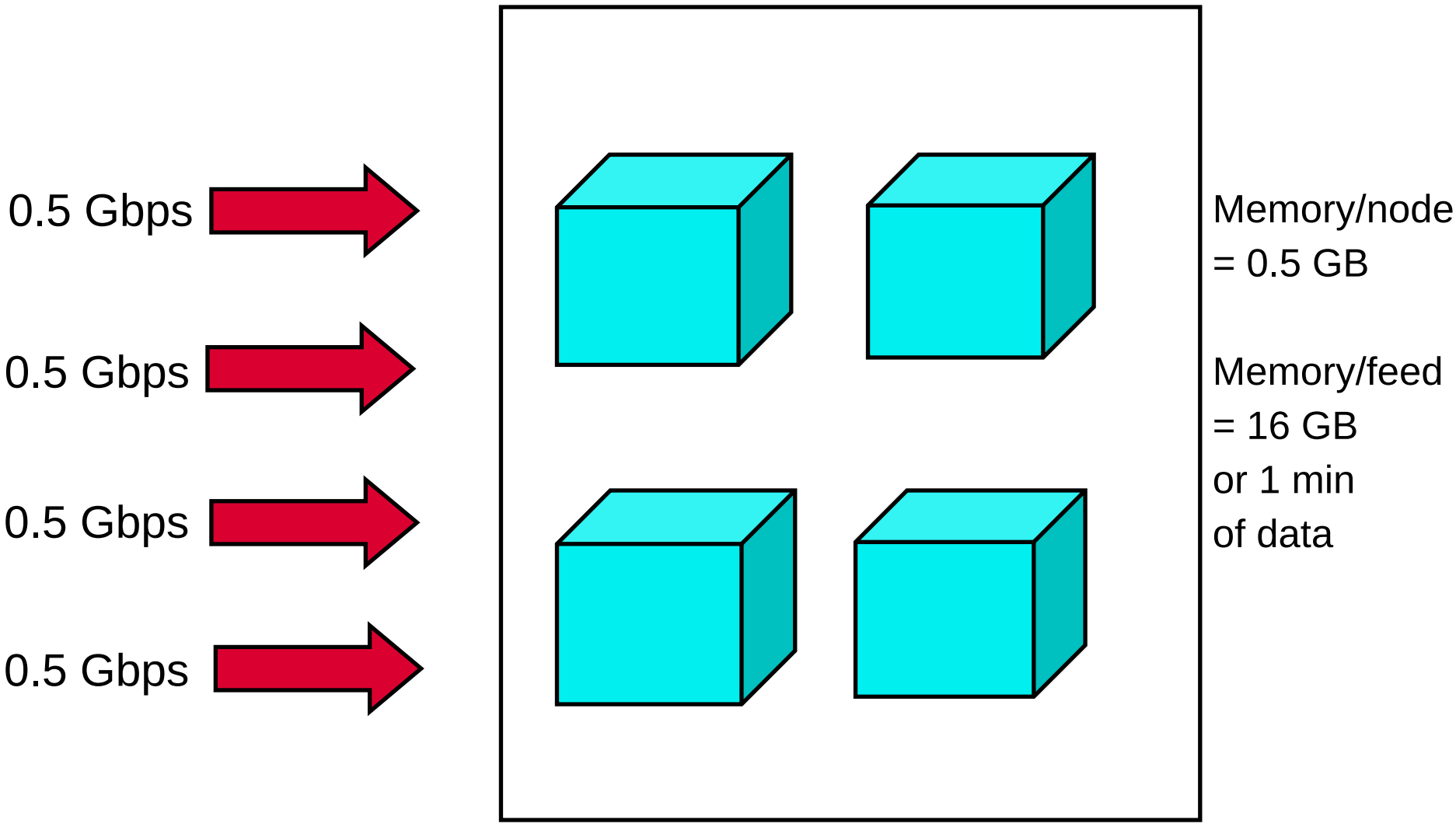
ASTRON will get 6 racks

Data Flow in one Rack

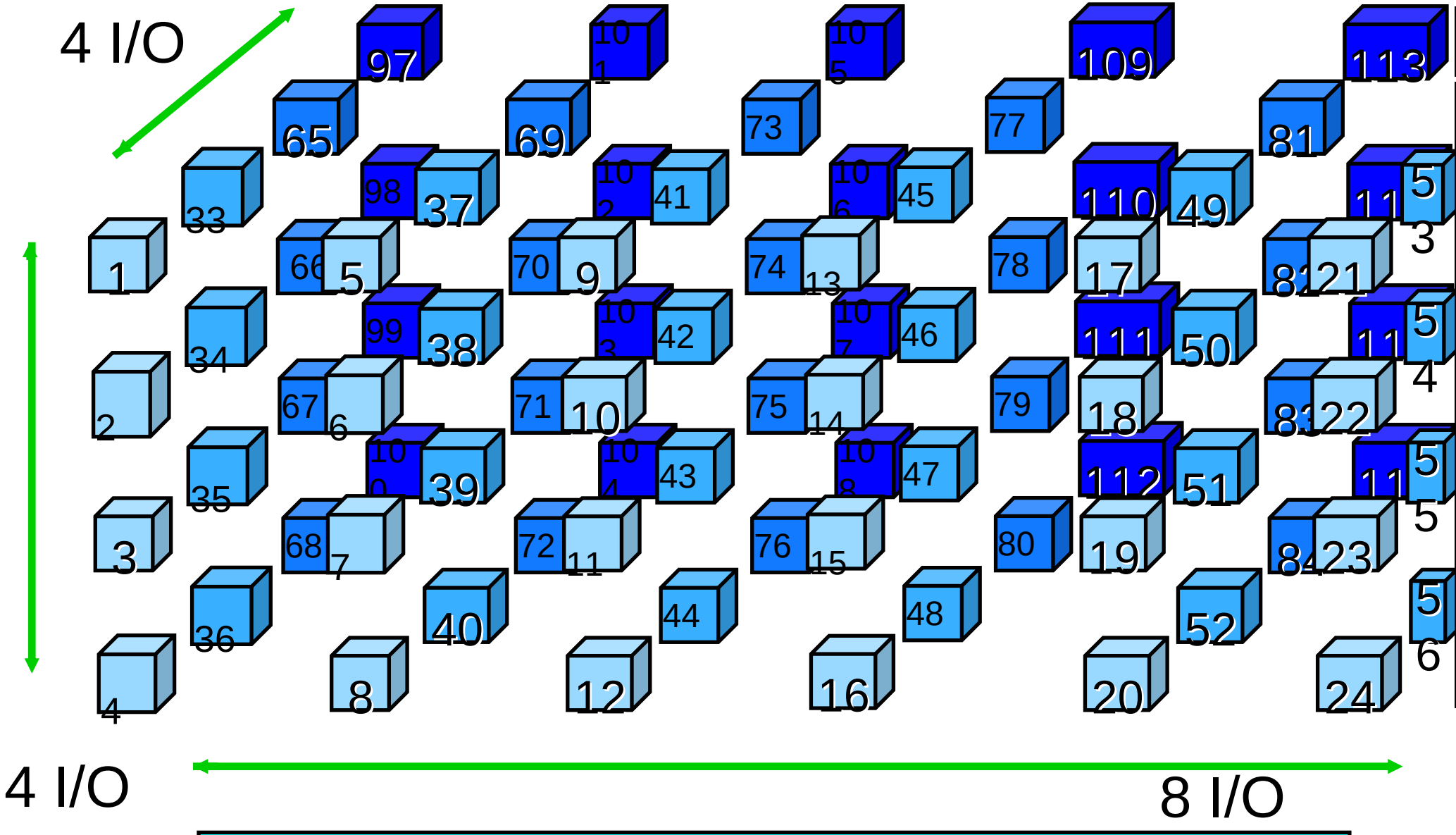
- 128 1-Gbps ethernet I/O nodes, two directions simultaneously
- each IO feed goes to an IO node, which is connected to 8 compute nodes by a hierarchical tree that pipes data at 2.8 Gbps bi-directional



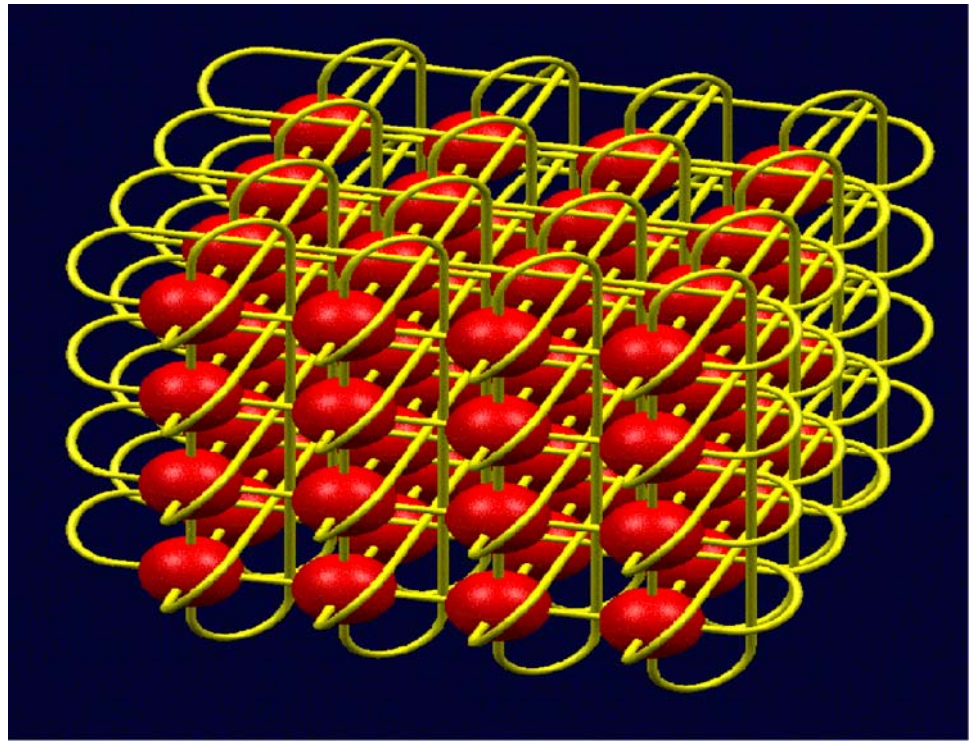
- Each antenna feed at ~2 Gbps can be divided over 4 IO nodes
- each compute node reads from a socket of antenna data at ~0.5 Gbps



4 I/O nodes along with their 32 compute nodes = 1 node card

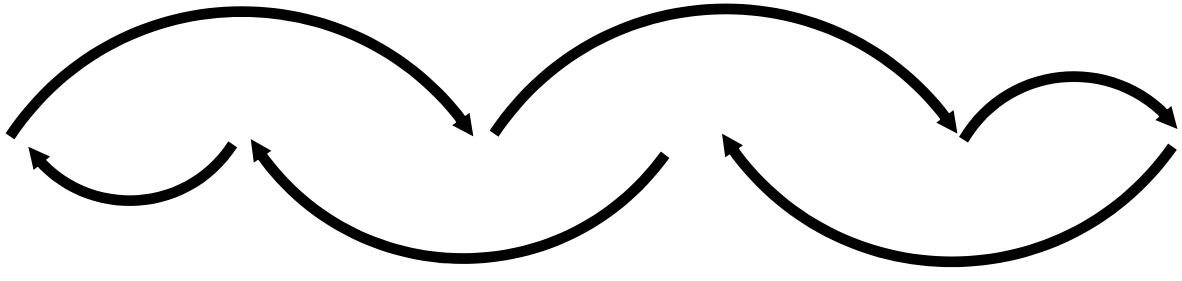

4 I/O
4 I/O
8 I/O
128 I/O cubes; 1024 processor nodes; 8x8x16 node torus
6 racks = 16x16x24 node torus

A torus has independent connections in 3 dimensions.



The torus bandwidth is 1.4 Gbps each way in all 3 dimensions simultaneously

Wiring actually jumps over the physical neighbors to prevent large timing mismatches between distant edges.



Two other networks

- barrier network to all nodes: allows programs to stay in synch
- control network to all nodes: boot, monitor, partition
  - partitions set up in software using linkchips
  - smallest partition is a midplane (512 nodes, half rack)
  - may partition system as midplanes, racks, multiple racks
  - each partition runs a different job

I/O SUM for 6 racks

- 768 IOs @ 1Gbps streaming can read from sockets at ~384 Gbps

Processing Power

- Each node has two processors, and
- each processor has 2 FPUs ("double hummer") and can do a complex*8 product in 2 clock cycles if data is streamed with 16-bit alignment in L1 cache (32K L1 storage ~ 2000 double complex numbers)

- clock speed is 700 MHz, so c.p. rate is 1/2 of this, or 350 M c.p./sec/processor

- this is maximum complex product rate per node if 100% pipelined and second processor inside each node is used only for message passing

  (use of second processor for complex products would increase this rate)

- 6144 nodes can do c.p. at ~1 T c.p/s for 1 proc/node and 50% effic.
  - note: 95% efficiency measured on complex product for pipelined data

Timing for Message Passing on Torus

- How busy is the second processor for MPI?
- Each node receives data at 2 Gbps/32 nodes = 64 Mbps
- all-to-all command rearranges data along torus at 1.4 Gbps each direction
- average number of node-to-node hops in longest dimension is 6
  - (=24 nodes in longest direction divided by 2 for bi-directional and 2 for average)
- fraction of time doing all-to-all is

  64 Mbps * 6 hops / 2.8 Gbps each dimension = 0.14

  - for 50% efficiency, the extra processor is doing MPI for 0.28 of the time, leaving ~3/4 of the time for complex multiplies on 2nd processor
    - note 50% efficiency measured for MPI alltoall using compiler
    - 95% efficiency measured for all to all in low level compiler language

Sample c.p. Rate: Virtual Core Beamforming

- 3200 antenna inputs, 32000 ch/ms, in 2 polarizations:
- Total complex product rate
  - f = (3200 weighted c.p.)*(32,000 ch/ms)*(2 pol)*(2 prod/pol)=400M c.p./ms
- divided among 6144 nodes = 66,000 complex products/ms/node
- compared to maximum c.p. rate of 350,000 c.p. per millisecond per node
- BG/L can handle c.p. rate, but the IO into BlueGene is not high enough for all 3200 antennae, so probably should do this VC Beamforming outside

Complex product rate for Station Beam Correlations

- After 64 VC and 45 RS beams are formed,
- Central processor has to do 109^2/2 station products * 32,000 channels per ms * 2 polarizations * 2 polarization products per polarization
- divide by 6144 nodes = 123 M c.p./s/node
- compared to max single-processor rate of 350 Mc.p./s/node

Sample Data Flow for station correlations

- 64 VC+45 RS inputs @ 2 Gps each = 440 IOs distributed over 6 racks
- Each IO directly linked to 8 compute nodes by hierarchical tree
- Each station's 2 Gbps of data is initially distributed among 32 nodes
  - 30 second data buffer takes half the RAM per node (0.5 GB RAM/node)
- MPI_alltoall redistributes the data so each node has some of the channels in both polarizations from all telescopes
  - 32000 channels in 2 polarizations = 11 channel-pols/node
  - 1000 ms of data expanded to 16B complex for 11 channels and 110 stations = 19 MB out of the remaining (non-buffer) 250 MB RAM per node
  - each channel fits in L3 cache: 1000 ms, 2 pol, 16By, 110 stations = 3.5 MB out of 4 MB/node cache.
  - L1 cache holds 32kB = 2000 double-complex numbers --> allows streaming
- Each node does all cross correlations for its own channels

Pulsar Tied Array Beamformer

- For 110 input streams making 128 beams in 2 polarizations,
- need a total rate of complex products to be
- (110 stations)*(128 beams)*(2 pol.)*(32000 channels/ms)=0.9 Tc.p./s
- Divided among 6144 nodes gives a rate of 147 Mc.p./sec/node, compared to peak rate of 350 M c.p./sec/node using 1 processor per node

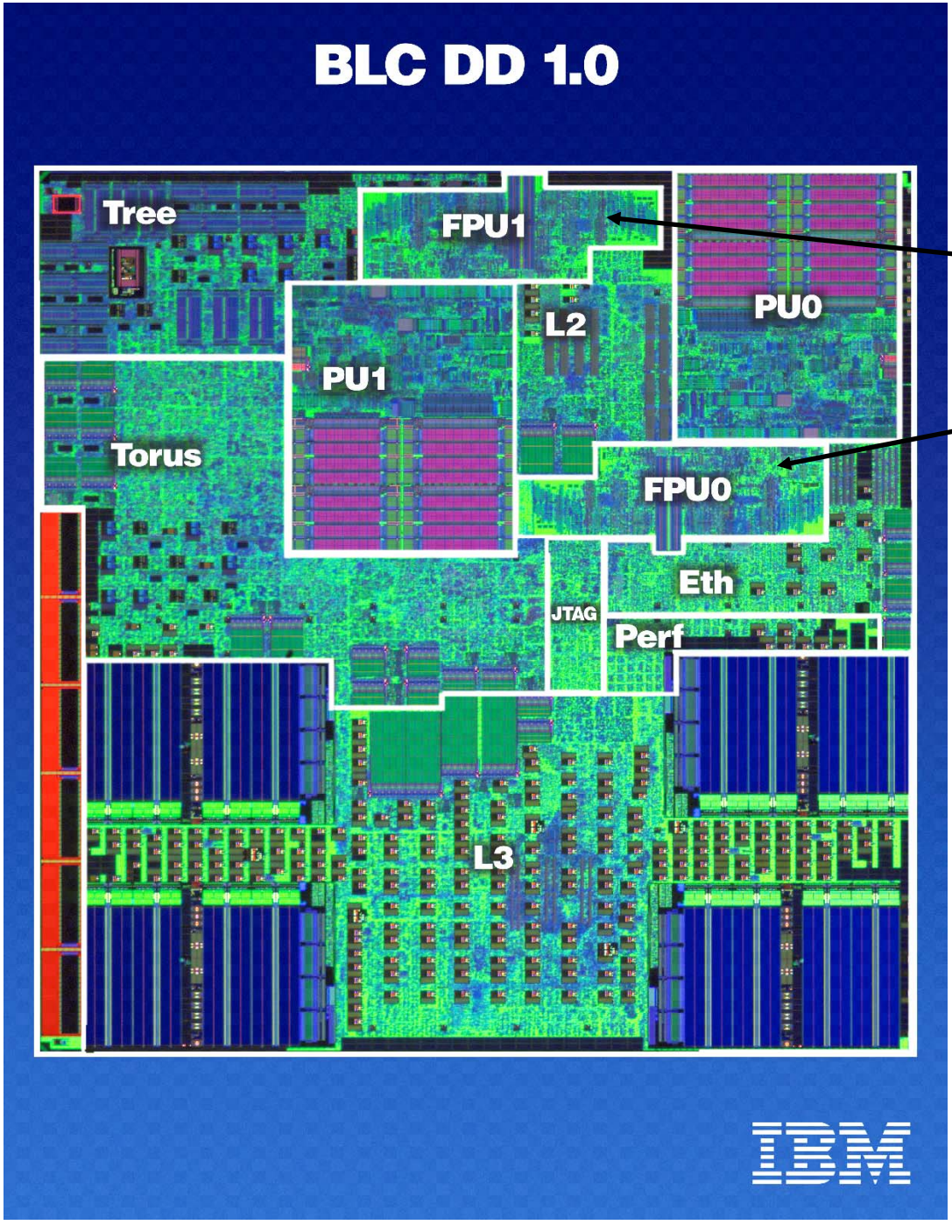
Epoch of Reionization

- For 64 V.C. input streams making 25 beams with 3200 channels in 2 polarizations,
- Need a total rate of complex products to be
- (64^2/2 pairs)*(25 beams)*(4 pol. pairs)*(3200 channels/ms)
- Divided among 6144 nodes gives a rate of 105 Mc.p./sec/node, compared to peak rate of 350 M c.p./sec/node using 1 processor per node

Summary of LOFAR on BG/L

- BlueGene/L can handle LOFAR station data rates and c.p. rates
- configuration can change with software commands
- handles data with high bit counts
- also usable as general purpose computer (~20 Tf sustained)
  - good for Dutch Infrastructure, an attraction for Industry partners, science ...

- BlueGene/L Innovations:
  - system on a chip design (2 processors, all networks, memory)
  - 4 independent networks: tree, torus, barrier, control
  - variable torus sizes (controlled by software using link chips)
  - moderate clock speed (700 MHz)
    - good for RAM reads, good for low power consumption (25 kW/rack)
  - LINUX kernel on IO nodes



How fast can BG be?

Construction is modular.
Can replace one or both
FPU with something else

128 bit loads

memory is already on
chip (L1,L2,L3)

32-4 bit units
seems possible

Options for Faster Computations with smaller dataword sizes (IRQA, SKA)

thanks to Ruud Haring (Mgr. Cellular Systems Chip Development) and George Chiu (Mgr., Advanced Server Hardware Systems) -- BG/L team

- Future BlueGene-type machines not known, but suggest special chips:
  - replace FPUs with auxiliary processing units (APUs) on same chip
    - APU = digital signal processor, gate array, etc.
  - keep PowerPC cores as controllers to preserve familiar software environment with development tools, debuggers, etc.
  - keep BlueGene format with same communication & packaging on chip
- Use current BlueGene chip technology to contain cost and leverage fabrication experience
  - could start design now based on existing BG/L chip
- When contents and instruction set for APU are known, IBM Rochester Engineering &Technology Service can do it all, from chips to racks
- if no commercial interest for IBM, would require 100% outside funding.