Software Correlators for Dish and Sparse Aperture Arrays of the SKA Phase I

Jongsoo Kim
Korea Astronomy and Space Science Institute

Collaborators: Paul Alexander (Univ. of Cambridge)
Andrew Faulkner (Univ. of Cambridge)

Correlators for Radio Interferometry

- ASIC (Application-Specific Integrated Circuit)
- FPGA (Field-Programmable Gate Arrays)
- Software (high level-languages, e.g., C/C++)
  - Rapid development
  - Expandability
  - …

Current Status of SC

- LBA (Australian Long Baseline Array)
  - 8 antennas (Parkes, … 22-64m, 1.4-22GHz)
  - DiFX software correlator (2006; Deller et al. 2007, 2011)
- VLBA (Very Long Baseline Array)
  - 10 antennas (25m, 330MHz - 86GHz)
  - DiFX
- MPIfR (the Max Planck Institute for Radio-astronomy)
  - Mark4 → DiFX

Current Status of SC (cont.)

- GMRT (Giant Metrewave Radio Telescope)
  - 30 antennas (45m, 50MHz-1.5GHz), 32MHz
  - ASIC → software correlator (Roy et al. 2010)
- LOFAR (Low Frequency Array)
  - LBA (Low Band Antennae) 10-90MHz
  - HBA (High Band Antennae) 110 – 250MHz
  - IBM BlueGene/P: software correlation

CoDR for SKA Phase I, Memo 125

- Key Sciences: H I and Pulsars
- Sparse Aperture Array

  70-450 MHz, A/T_{sys} = 2000m^2/K, L_{max}=100Km
- Dish Array

  0.45-3 GHz, A/T_{sys}=1000m^2/K, 250 15m dishes single-pixel feeds, L_{max}=100Km
- Construction: 2016-19
- Budget: 350M Euros

SKA Phase I: Preliminary System, Memo 130

- Dish array
  - 250 15m dishes,
  - Bandwidth: 0.55GHz (0.45-1.0GHz), and 1.0GHz (1.0-2.0 GHz)
  - Dual polarizations
  - Bits per sample: 4 bits

SKA Phase I: Preliminary System, Memo 130

- Sparse Aperture array
  - 50 stations
  - Bandwidth: 380 MHz (70-450 MHz)
  - Number of beam: 480 (160)
  - Dual polarizations
  - Bits per sample: 4 bits

Correlation Theorem, FX-correlator

$$R_i(f) = \int_{-\infty}^{+\infty} r_i(t) \mathrm{e}^{2\pi i f t} dt$$

F-step (FT):
~$\log_2(N_c)$ operations per sample

$$\int_{-\infty}^{+\infty} r_j(\tau + t) r_j(\tau) d\tau \Leftrightarrow R_j(f) R_j^*(f)$$

X-step (CMAC):

~ N operations per sample

FLOPS of the X-step in FX correlator

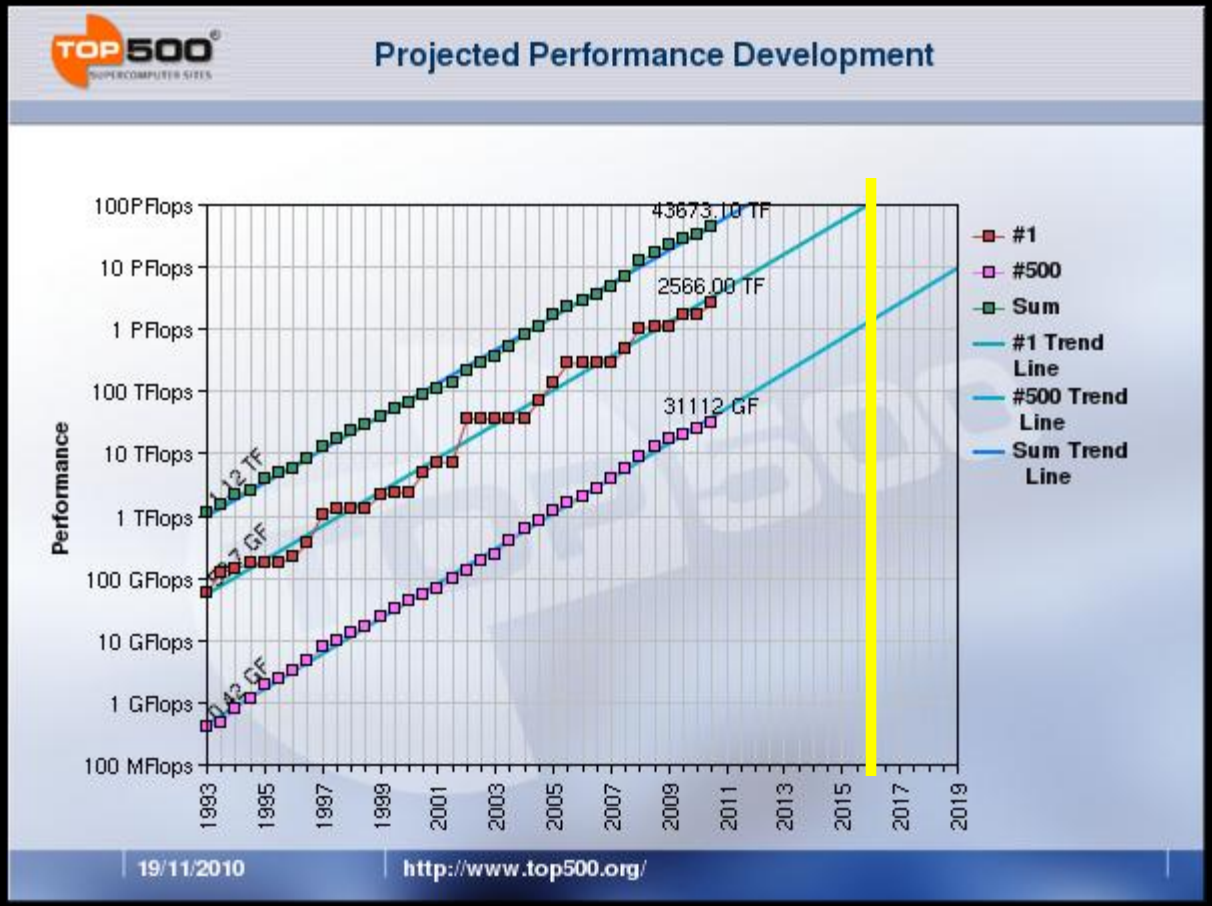
$$4 \times 8 \frac{N(N+1)}{2} N_b \left(\frac{B}{\mathrm{Hz}} \right) [\mathrm{FLOPS}] \approx 16 N^2 N_b \left(\frac{B}{\mathrm{GHz}} \right) [\mathrm{GFLOPS}]$$

- 4 is from $R_i R_j^*, R_i L_j^*, L_i R_j^*, L_i L_j^*$
- 8 is from 4 multiplications and 4 additions:

$$+R_i R_j^* = +(a_i + ib_i)(a_j - ib_j) = +(a_i a_j + b_i b_j) + i(b_i a_j - a_i b_j)$$

- $N(N+1)/2$ is the number of auto- and cross-correlations with antenna (station) N
- Dish array (N=250, B = 1 GHz, N_b=1)
  → 16×250^2 GFLOPS = 1 PFLOPS
- Sparse AA (N=50, B=380MHz, N_b=160)
  → $16 \times 50^2 \times 160 \times 0.38$ GFLOPS = 2.43 PFLOPS

top500


TOP500
Projected Performance Development
100 PFlops
10 PFlops
1 PFlops
100 TFlops
10 TFlops
1 TFlops
100 GFlops
10 GFlops
1 GFlops
100 MFlops
Performance
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
43673.10 TF
2566.00 TF
31112 GF
#1
#500
Sum
#1 Trend Line
#500 Trend Line
Sum Trend Line
19/11/2010
http://www.top500.org/

Design goals

- Connect antennas and computer nodes with simple network topology
- Use future technology development of HPC clusters
- Simplify programming

Data Rate per Dish

- Pure data:

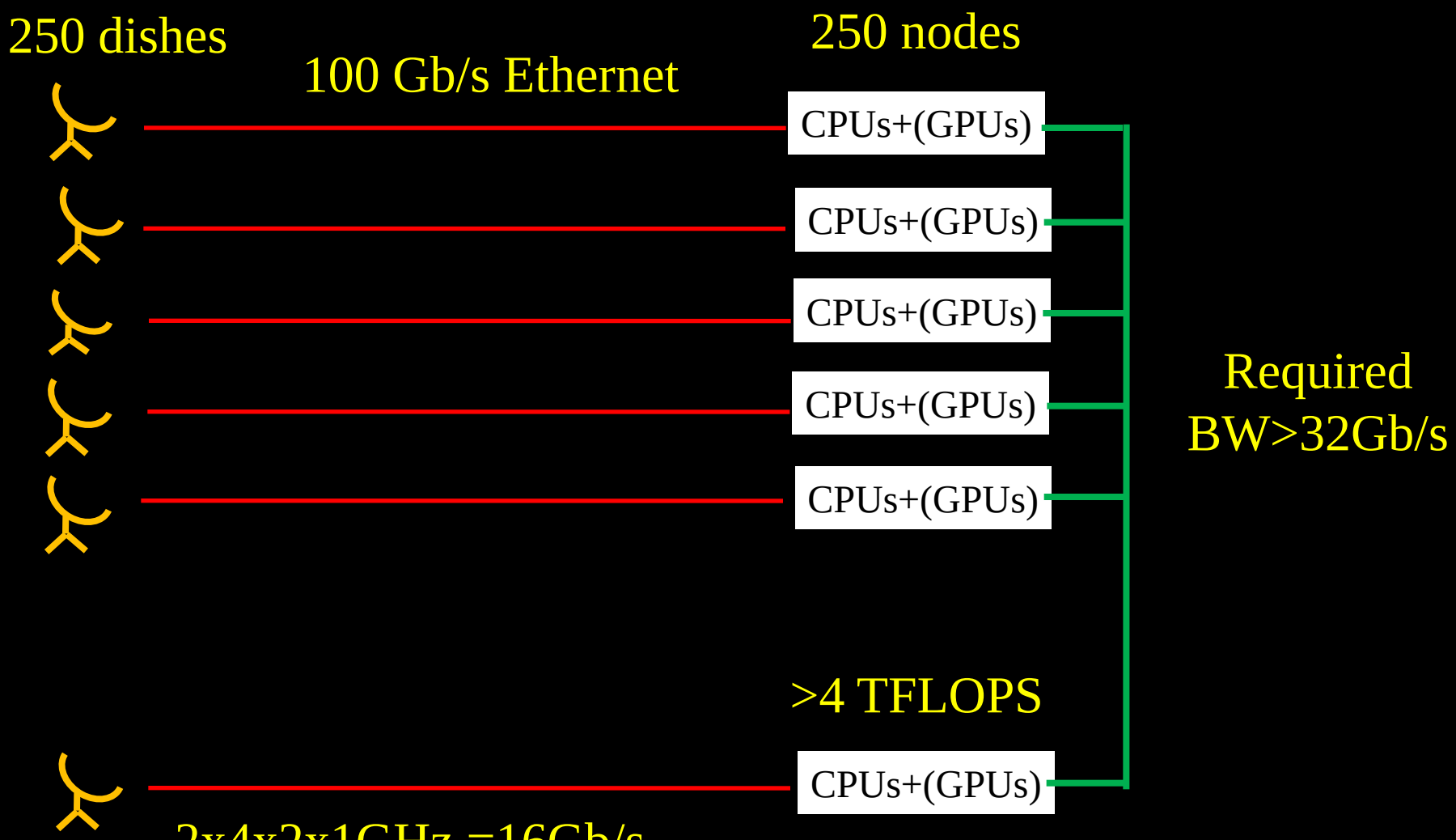
  2 (pol) x 4 (bit/sample) x 2 (Nyquist) x 1GHz (BW) =16Gb/s
- Encoding overhead

  20% (8b/10b; PCIe 2.0) → 1.5% (128b/130b; PCIe 3.0)
- UDP (User Data protocol) overhead

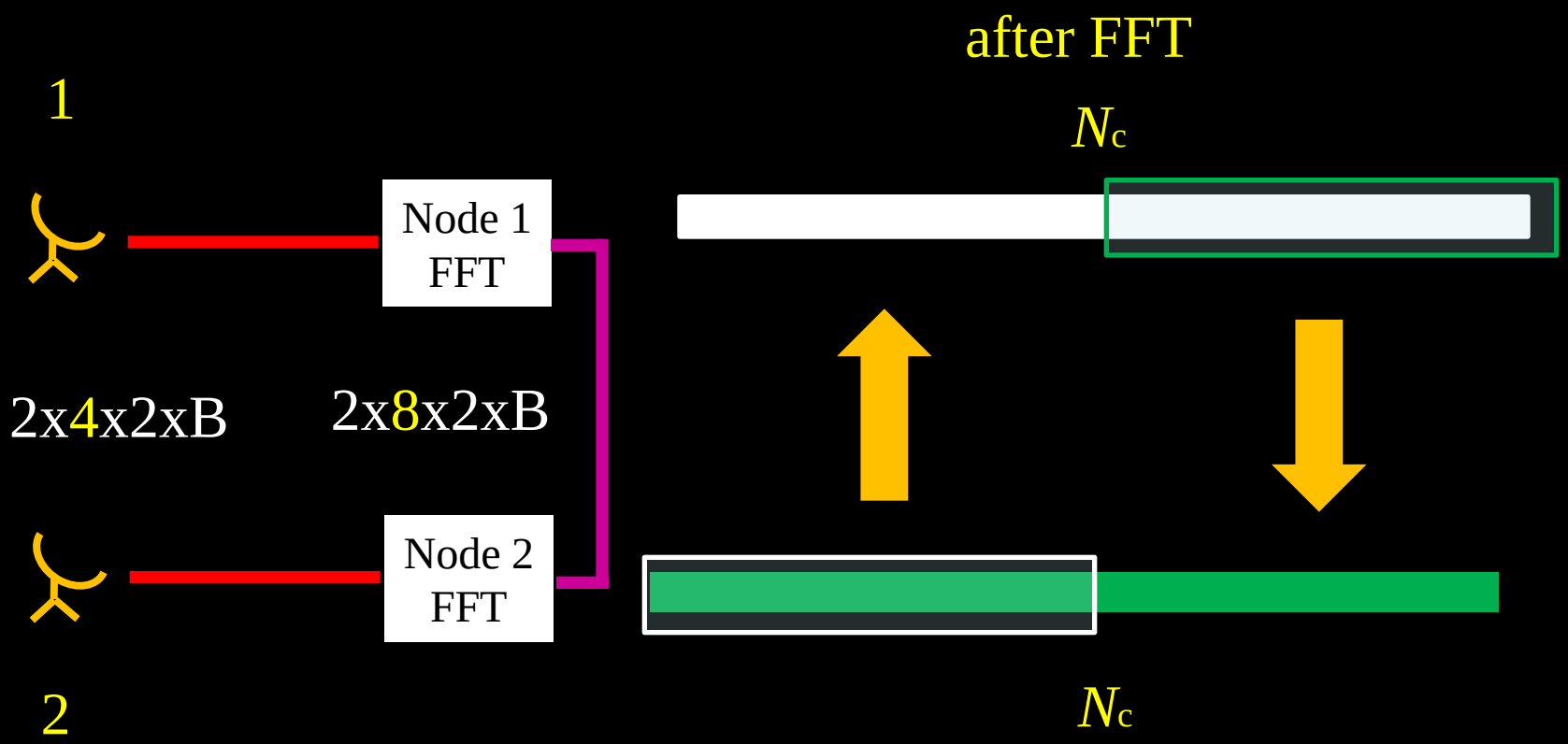
  < 1% (28 bytes for headers / 65,507 bytes data length)
- Oversample overhead

  ~ 20% (memo 130)

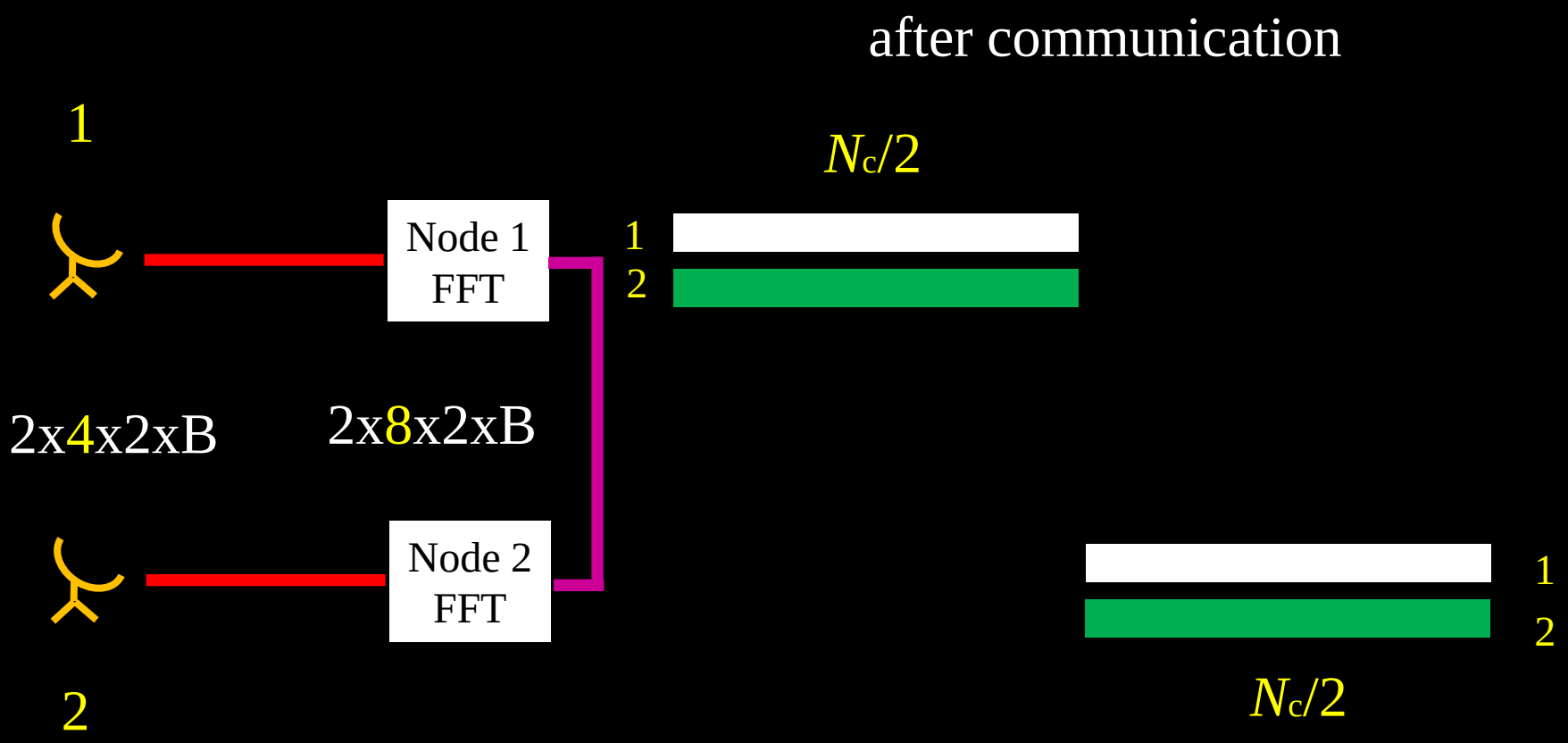
CoDR of a Software Correlator for the dish array



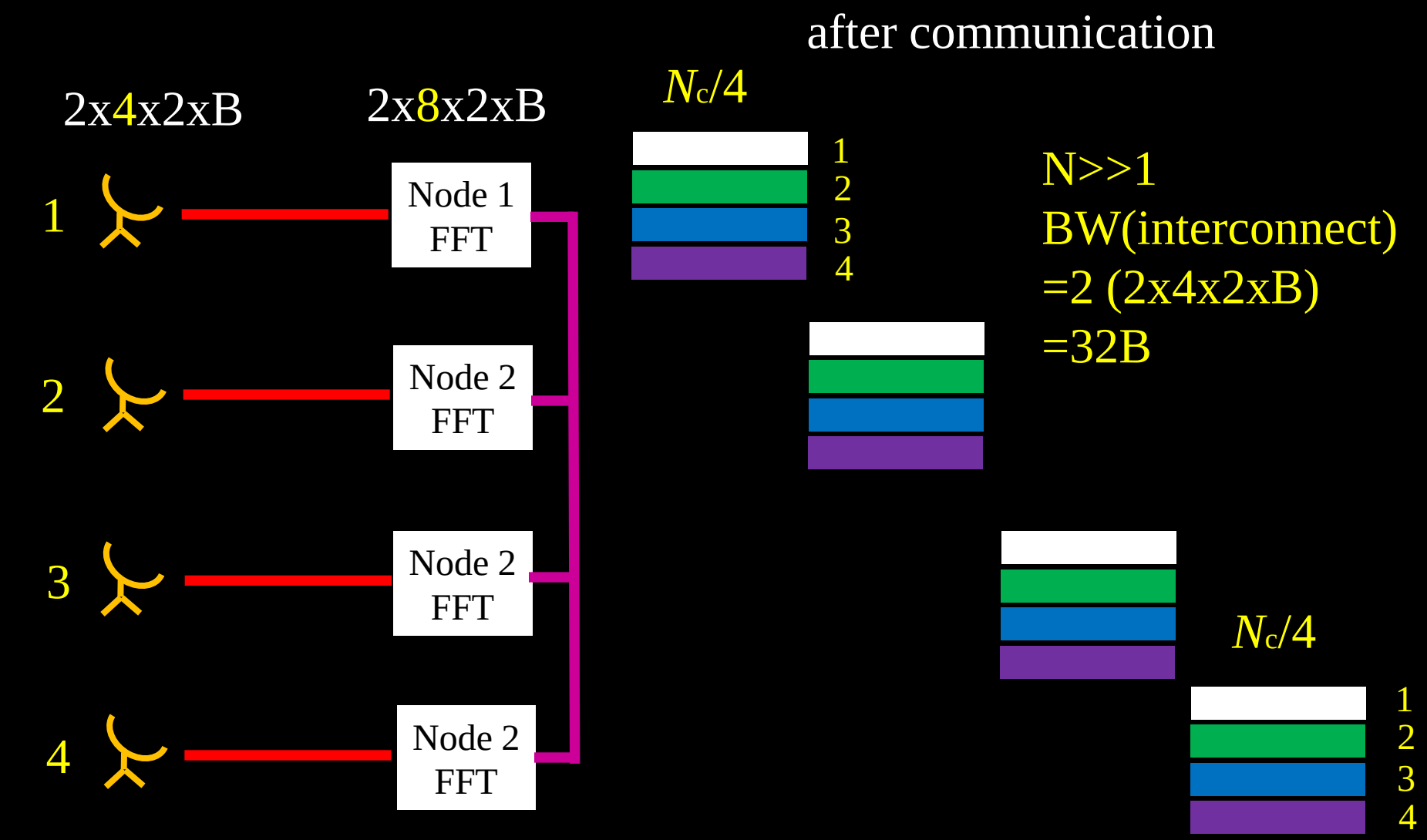
Communication between Computer Nodes I



Communication between Computer Nodes II



All-to-All Communication between Computer Nodes III



Data Rate per Station

- Pure data:

  2 (pol) x 4 (bit/sample) x 2 (Nyquist) x 0.38GHz (BW) x 160 (beams) =972.8 Gb/s

- Encoding overhead

  20% (8b/10b; PCIe 2.0) → 1.5% (128b/130b; PCIe 3.0)

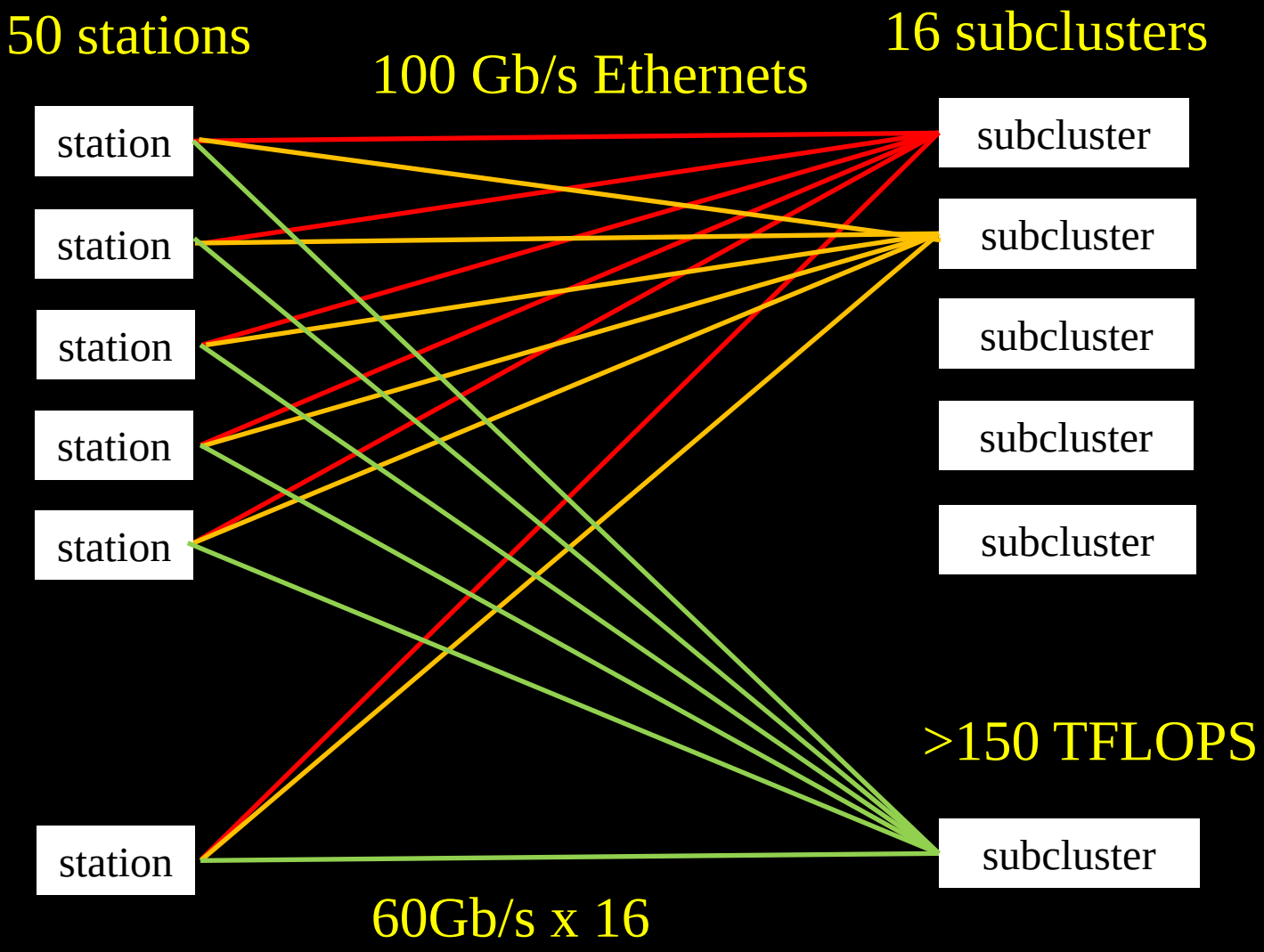
- UDP (User Data protocol) overhead

  < 1% (28 bytes for headers / 65,507 bytes data length)

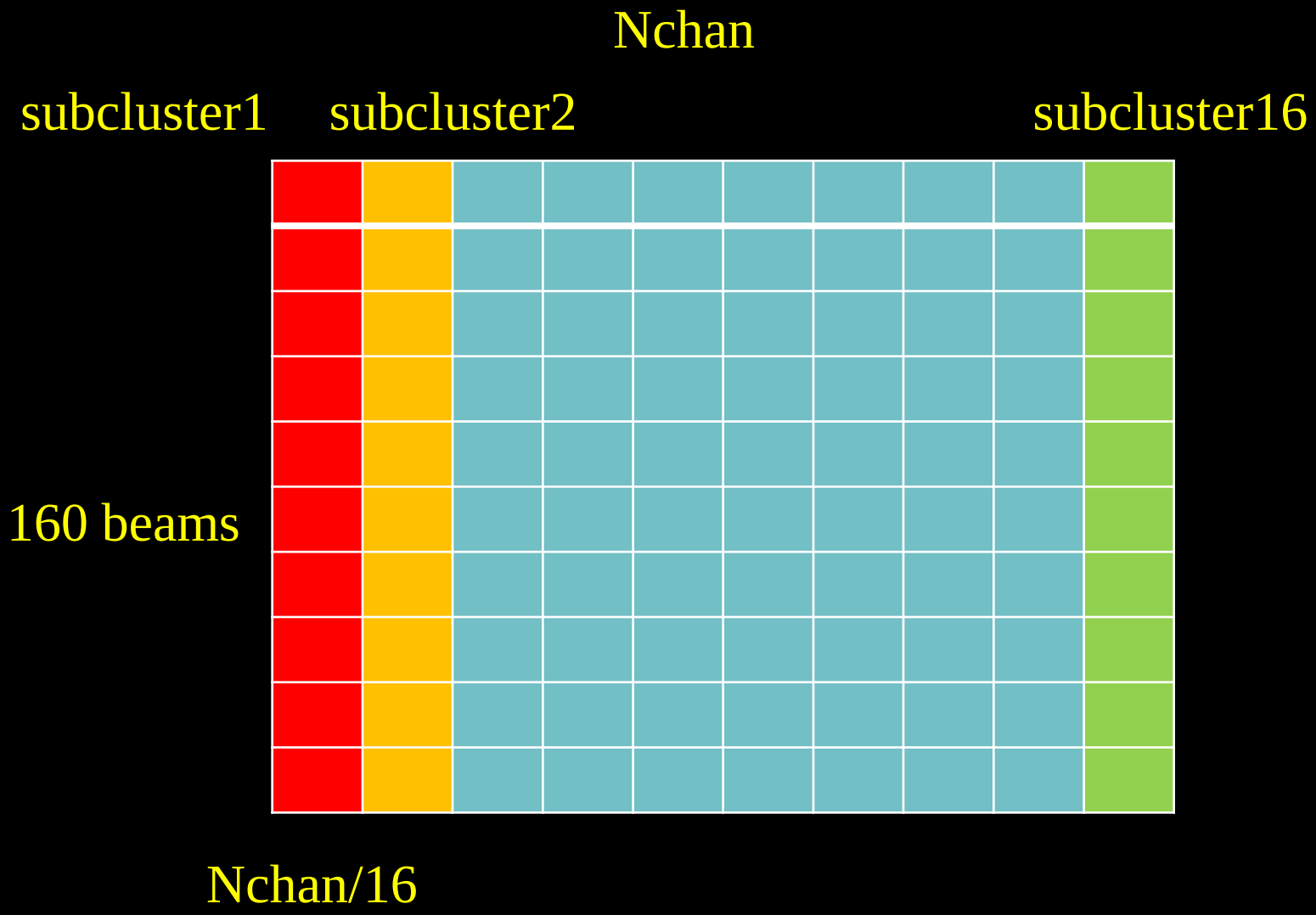
- Oversample overhead

  ~ 20% (memo 130)

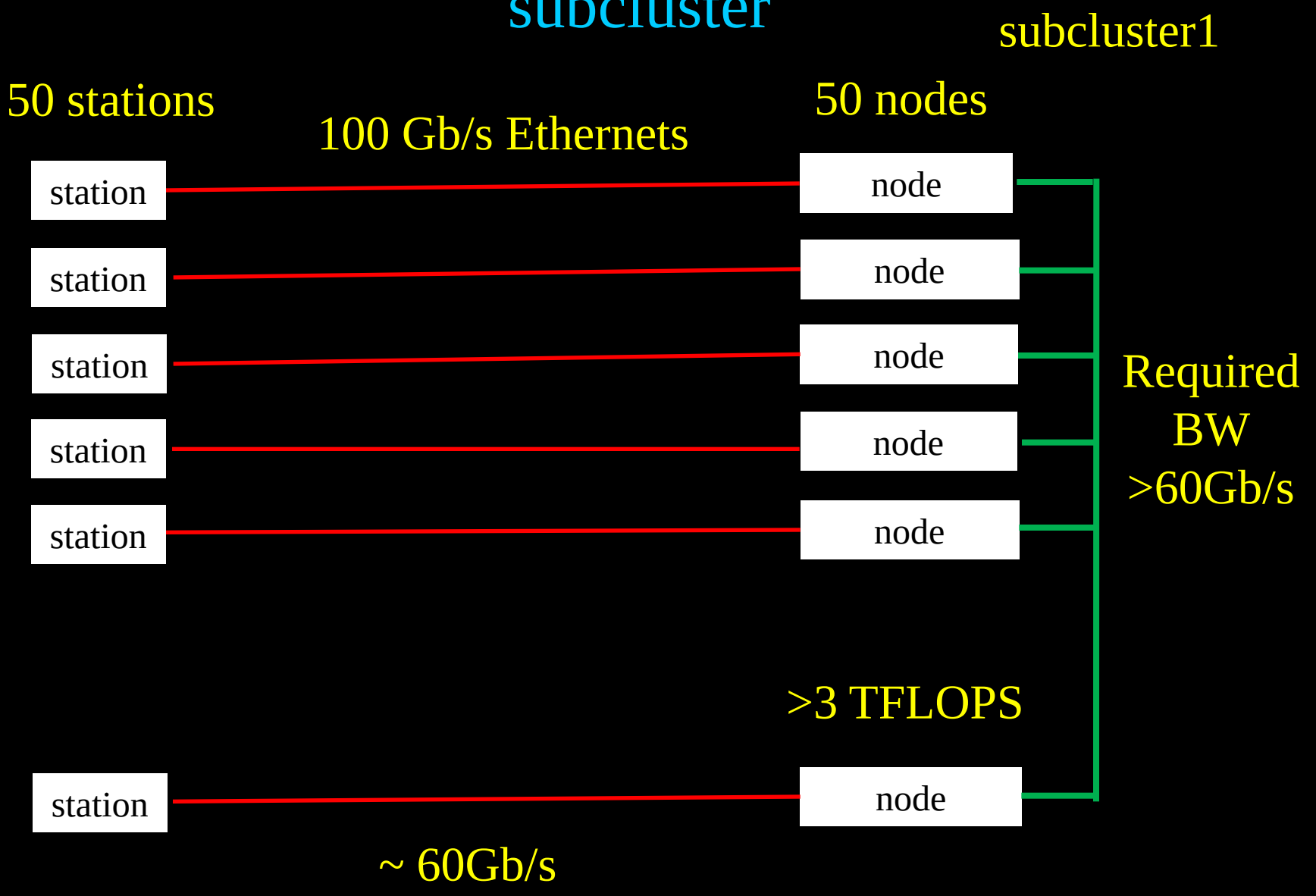
CoDR of a Software Correlator for the sparse AA



Channelized data in a station



Connectivity between stations and nodes in a subcluster



Cost and Power Estimates of SCs

	# of nodes	Cost per node [kEuros]	Cost of IB per port [kEuros]	Power per node [kW]	Total cost [M Euros]	Total power [MW]
Dish Array	250	5	1	1.0	1.5	0.25
Sparse AA	800	5	1	1.0	4.8	0.80
Total	1050				6.3	1.05

AI (Arithmetic Intensity)

- Definition: number of operations (flops) per byte
- AI = 8flops/16bytes (R_i,R_j) = 0.5

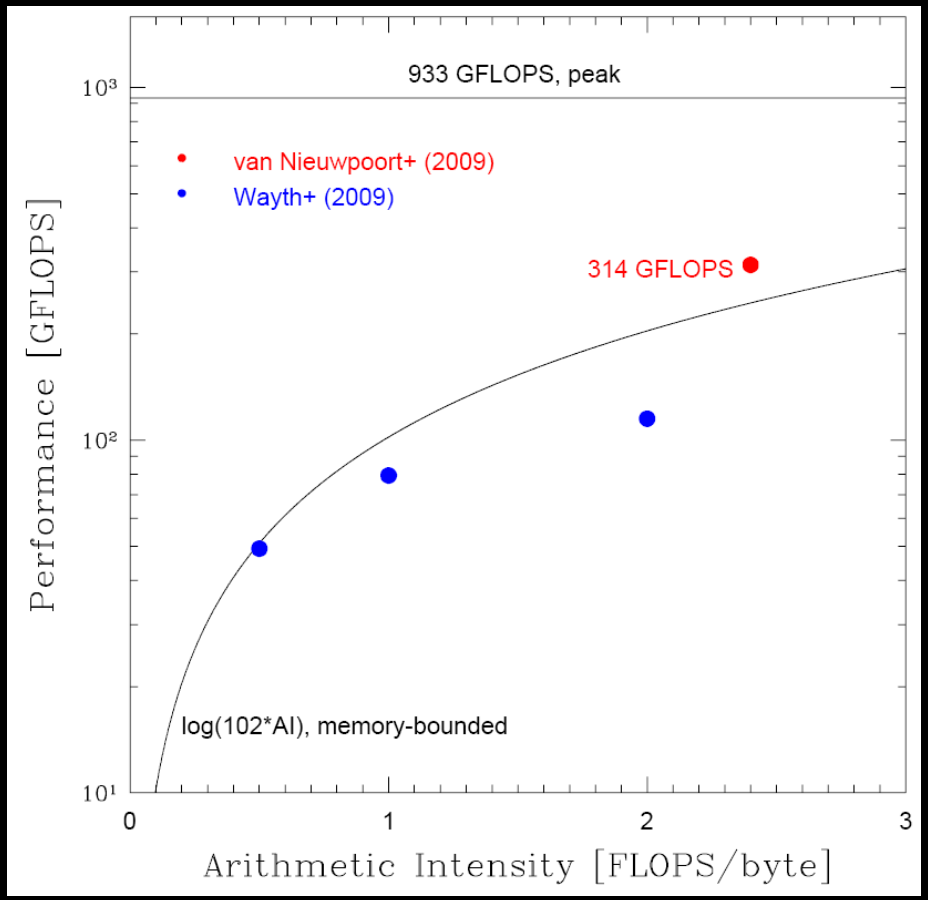
$$+R_i R_j^* = +(a_i + ib_i)(a_j - ib_j) = +(a_i a_j + b_i b_j) + i(b_i a_j - a_i b_j)$$

  AI = 32 flops/32bytes (R_i, L_i, R_j, L_j) = 1.0 for 1x1 tile

  AI = 2.4 for 3x2 tiles

- Since AIs are small numbers, correlation calculations are bounded by the memory bandwidth.
- Performance: AI x memory BW (=102GB/s)

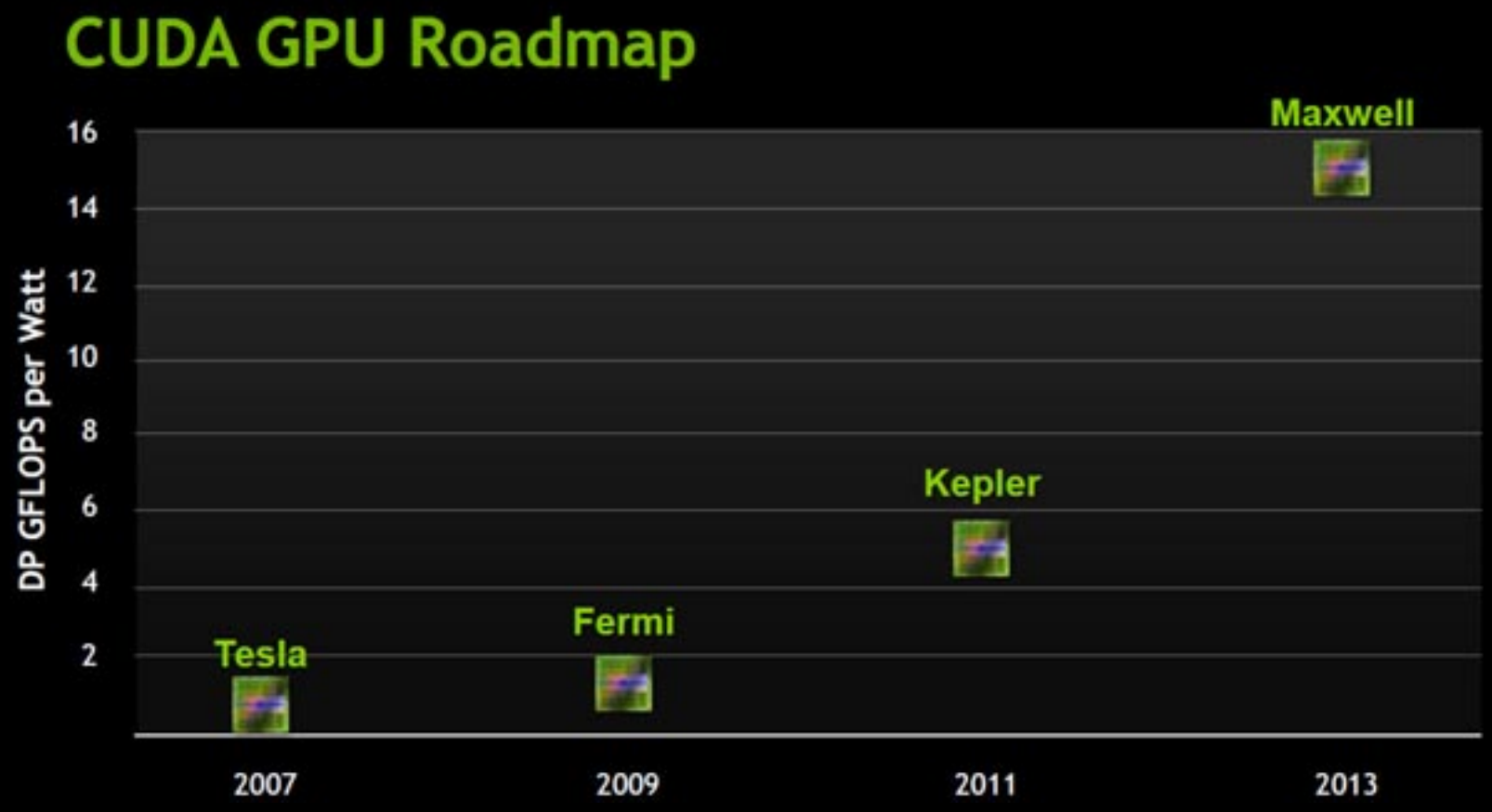
Wayth+ (2009), van Nieuwpoort+ (2009)



Performance of Tesla C1060 as a function of AI

- Performance is, indeed, memory-bounded.
- Maximum performance is about 1/3 of the peak performance.

Tesla Roadmap

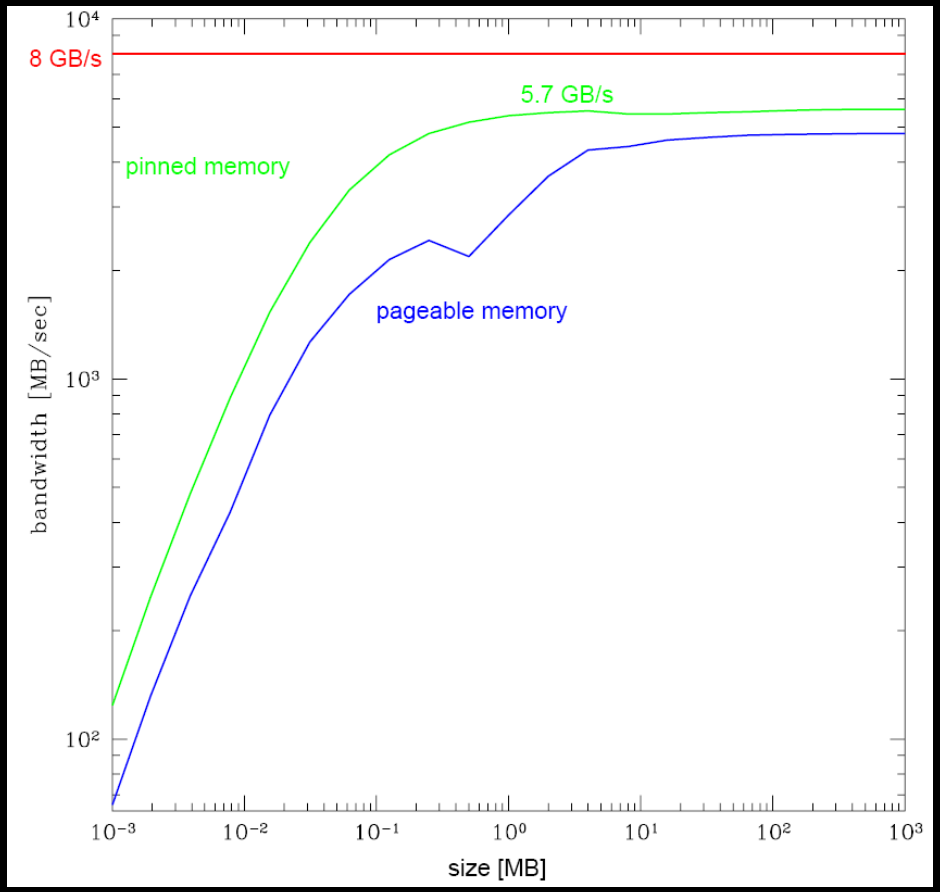


Conclusions

- Connectivity:
  - one 100GE connection / dish
  - 16 100GE connections / station
- Performance
  - One cluster with 1 PFLOPS for the dish array
  - 16 clusters, each cluster with 150 GFLOPS, for the sparse AA
- Cost and Power
  - 6 M Euros, 1 MW

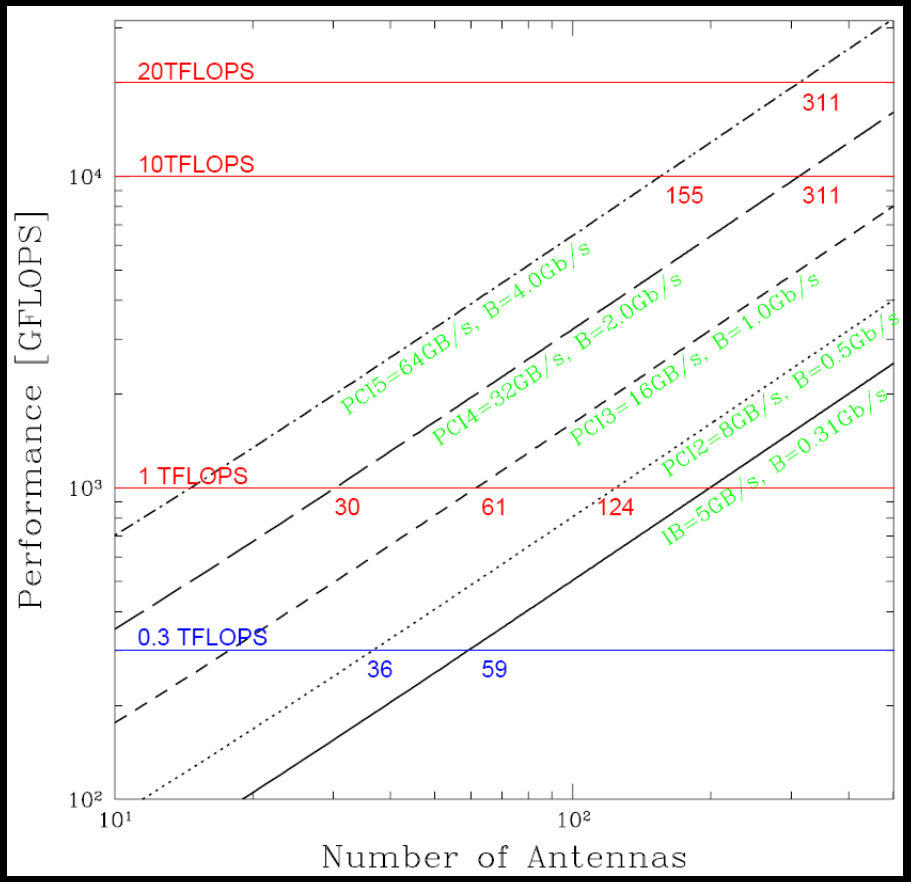
AI for host-device and host-host

- AI = N+1 FLOP/byte
  N x 16 bytes (R,L) = 16 N byte
  4x8xN(N+1)/2 FLOP
- PCI bus bandwidth
  PCI-e 2.0: 8.0GB/s (15 Jan. 2007)
  PCI-e 3.0: 16.0GB/s (2Q 2010, 2011)
  PCI-e 4.0: 32.0GB/s (201?)
  PCI-e 5.0: 64.0GB/s (201?)
- Performance [GFLOPS] =
  PCI BW [GB/s] x AI [FLOP/B]



Measured bandwidth of host-to-device (Tesla C1060)

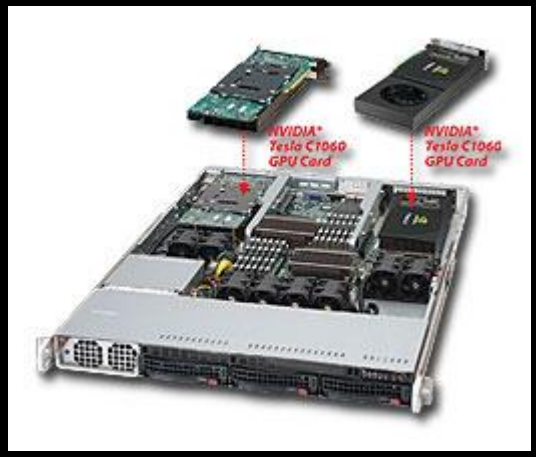
~70% of PCI -e2 bandwidth



Expected performance bounded by the BWs of PCI bus and interconnect

Power usage and Costing

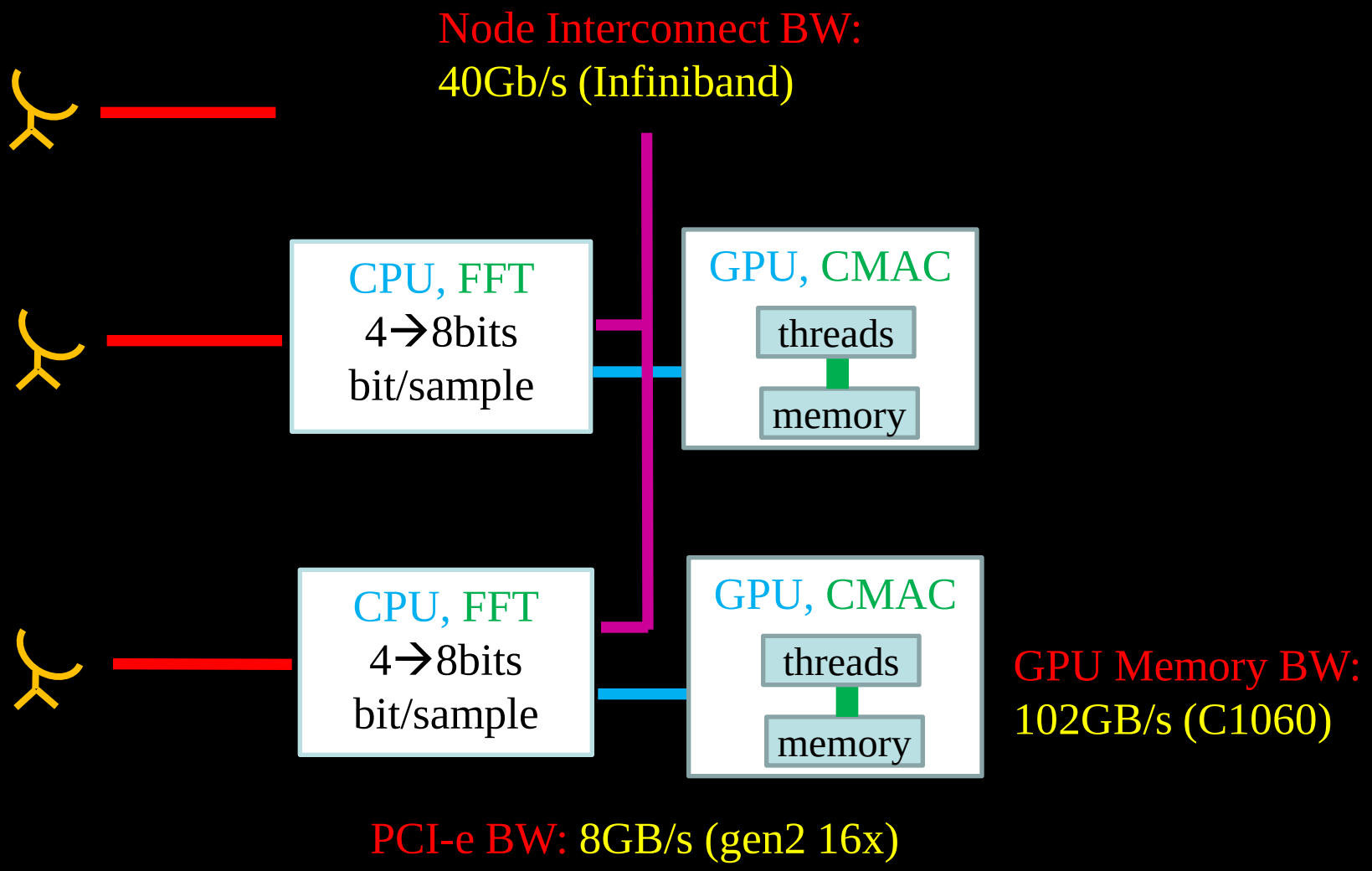
- Computer nodes
  - 1.4 KW, 4 K Euro for each server including 2x0.236 KW (2 GPUs)
  - 0.4 MW, 1.2 M Euro for 300 servers
- Network Switches
  - 3.8 KW for IB (40Gb) 328 ports



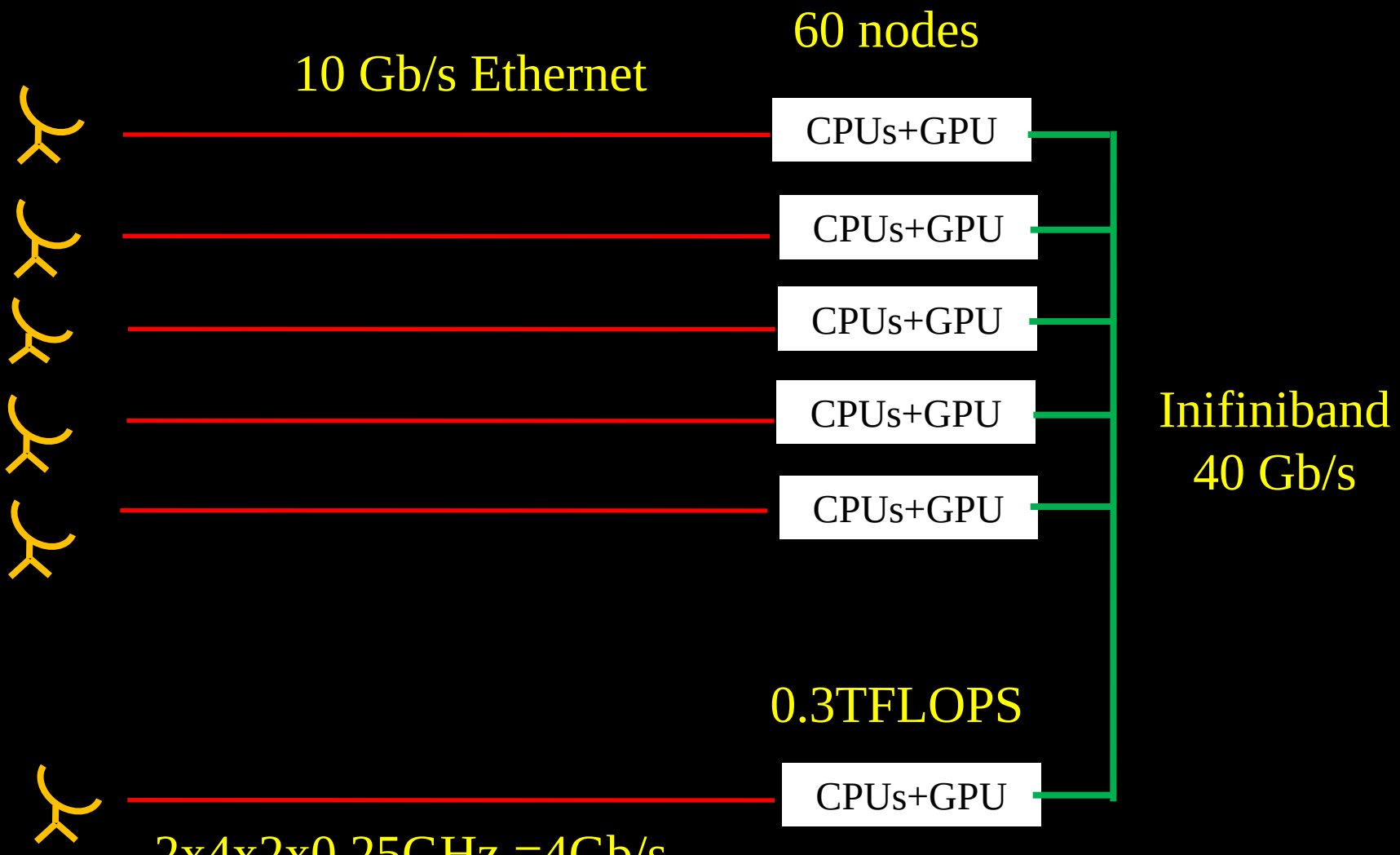
Technology Development in 2010

- 2Q 2010, (2011): PCI-e 3rd Generation
- 2Q (April) 2010: Nvida Fermi (512 cores, L1,L2 cache)
- (March 29) 2010: AMD 12 core Opteron 6100 processor
- (March 30) 2010: Intel 8 core Nehaem-EX Xeon processor

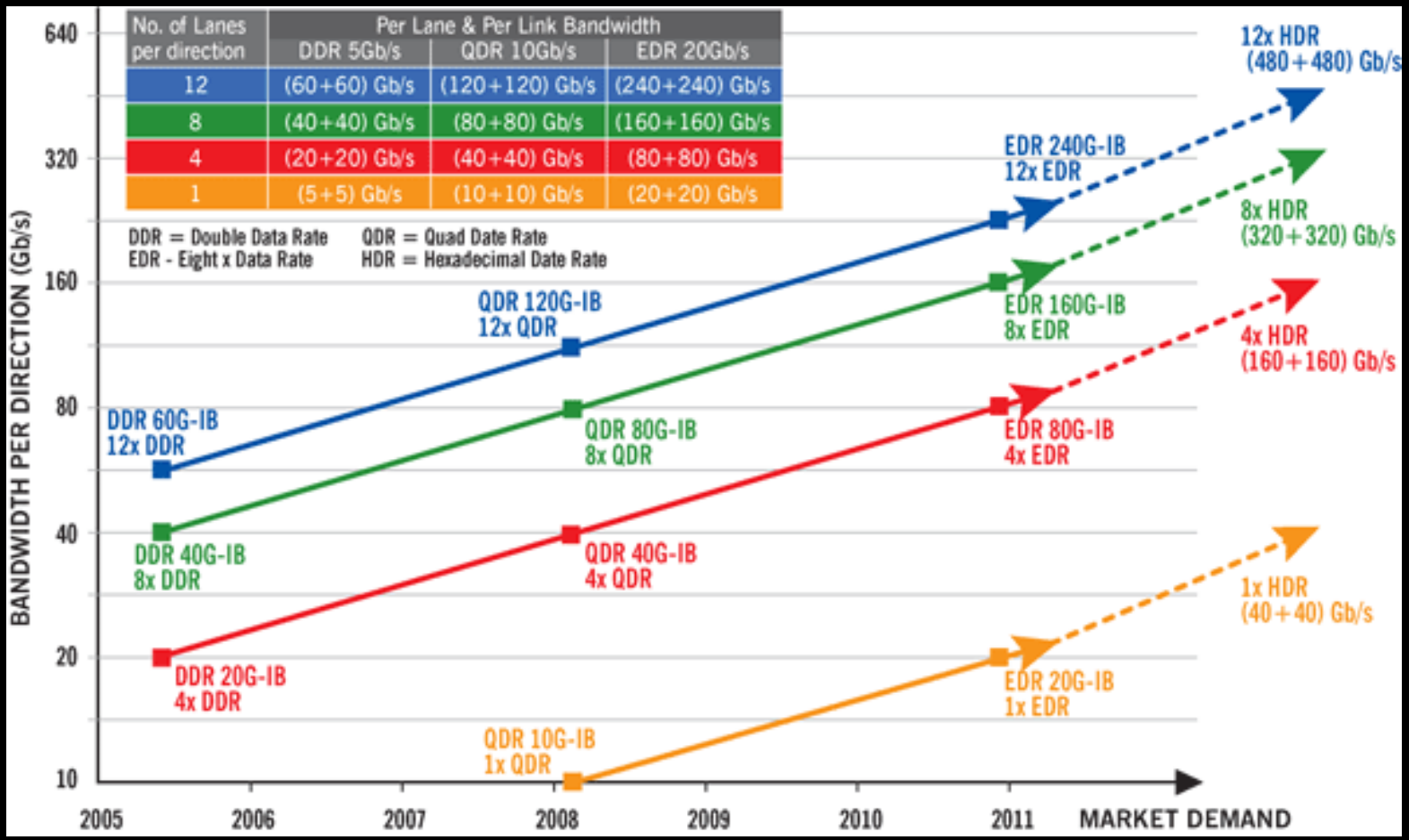
Data Flows


Node Interconnect BW:
40Gb/s (Infiniband)
CPU, FFT
4→8bits
bit/sample
GPU, CMAC
threads
memory
CPU, FFT
4→8bits
bit/sample
GPU, CMAC
threads
memory
GPU Memory BW:
102GB/s (C1060)
PCI-e BW: 8GB/s (gen2 16x)

Software Correlator with current technology



Infiniband Roadmap (IBTA)



Software Correlator for Phase I SKA (‘2018)