The "DiFX Baseband Data Processing system" in the Multi-Gigabit Era

John Morgan (ICRAR, AU) Helge Rottmann (MPIfR, DE)
Chris Phillips (CSIRO, AU) Adam Deller (NRAO, US)
Mauro Nanni (IRA/INAF, IT)

eVLBI Workshop
Perth
Monday 18 October 2010

International Centre for
Radio Astronomy Research

Overview

DiFX

- Baseband data is read from
  - Mark5 module
  - Disk file
  - Network socket
- Farmed out to cluster using MPI
- Converted to 32 bit floats
- Vector arithmetic (FFTs etc.) done using intel IPP library
- Results collated and written to disk

Parallelisation in DiFX

Correlation is readily parallelisable

- Splitting antennas is not usually a good idea
  - every antenna must be correlated with every other
- Splitting in time is OK
  - results must be averaged afterwards
- Splitting in frequency would also be possible
  - assuming more than one baseband channel[1]

At the moment DiFX only splits in time

- The next generation of DiFX will likely support multiple datastreams per antenna

[1] although c.f. WIDAR FXF architecture

How does the cluster interconnect limit us?

We need to get enough data onto each compute node to keep the CPU busy

- For current machines (8 CPU cores) this requires ~300 Mbit/s per machine
  - exact value depends on correlation parameters and efficiency of the code
- Most current clusters are built on 1 Gbps ethernet
- As multi-core technology develops we will require faster interconnects

Correlation is more demanding on the interconnect than many other scientific computing tasks due to

- Large data volumes
- Small number of operations per bit

In the past this has hampered the use of more exotic processors

Performance of DiFX

Phillips & Deller (2009) have already carried out extensive performance testing on a cluster with a 1 Gbit interconnect

- Fake baseband data was generated, eliminating reading the data from disk as a bottleneck
- Bottleneck was the interconnect for > 8 compute nodes (6 stations)

Several Clusters running DiFX have faster interconnects

- IRA, Bologna (10 Gbps Ethernet)
- MPIfR, Bonn (infiniband 12 Gbps)

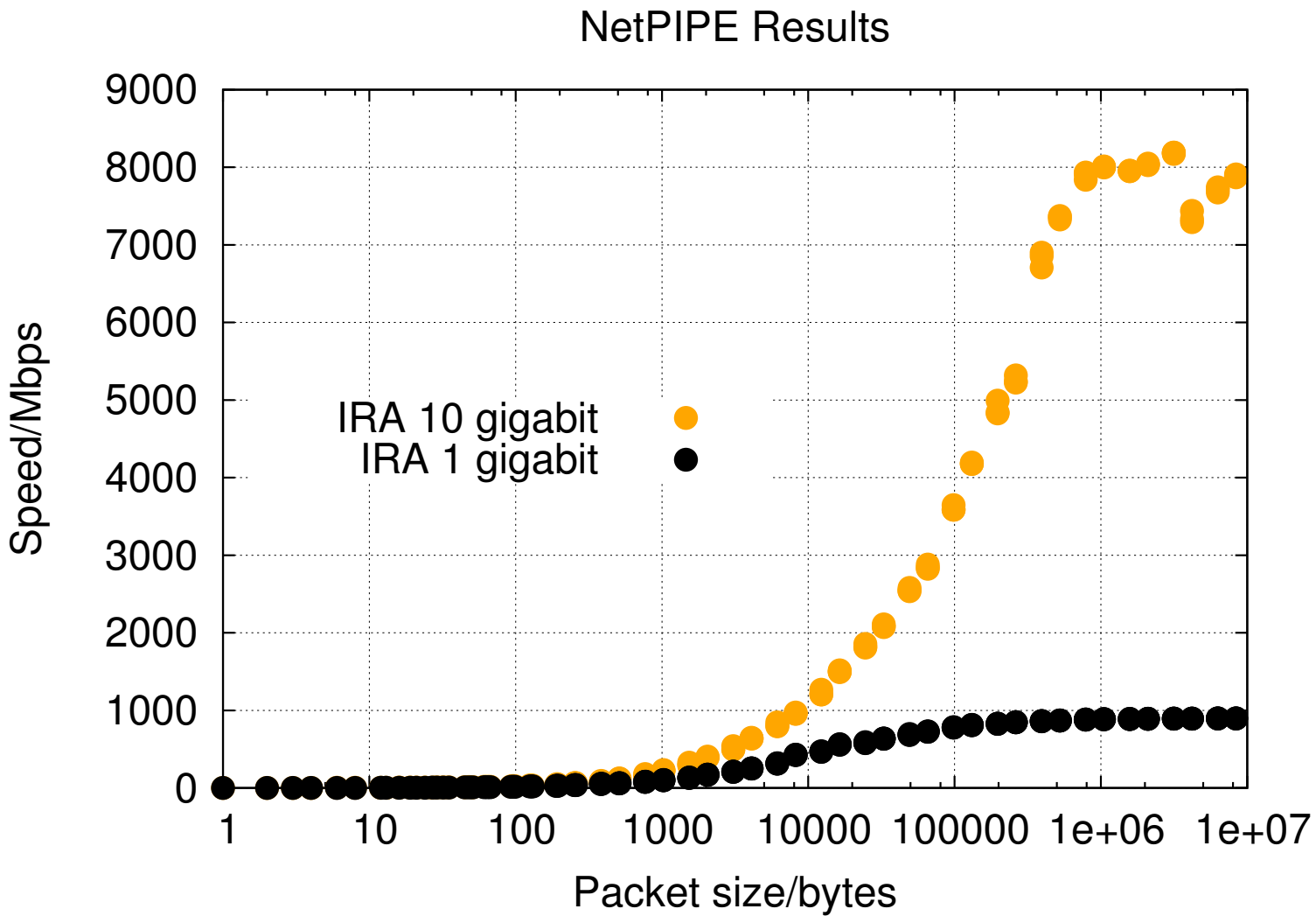
First

- Compare these interconnects in general terms

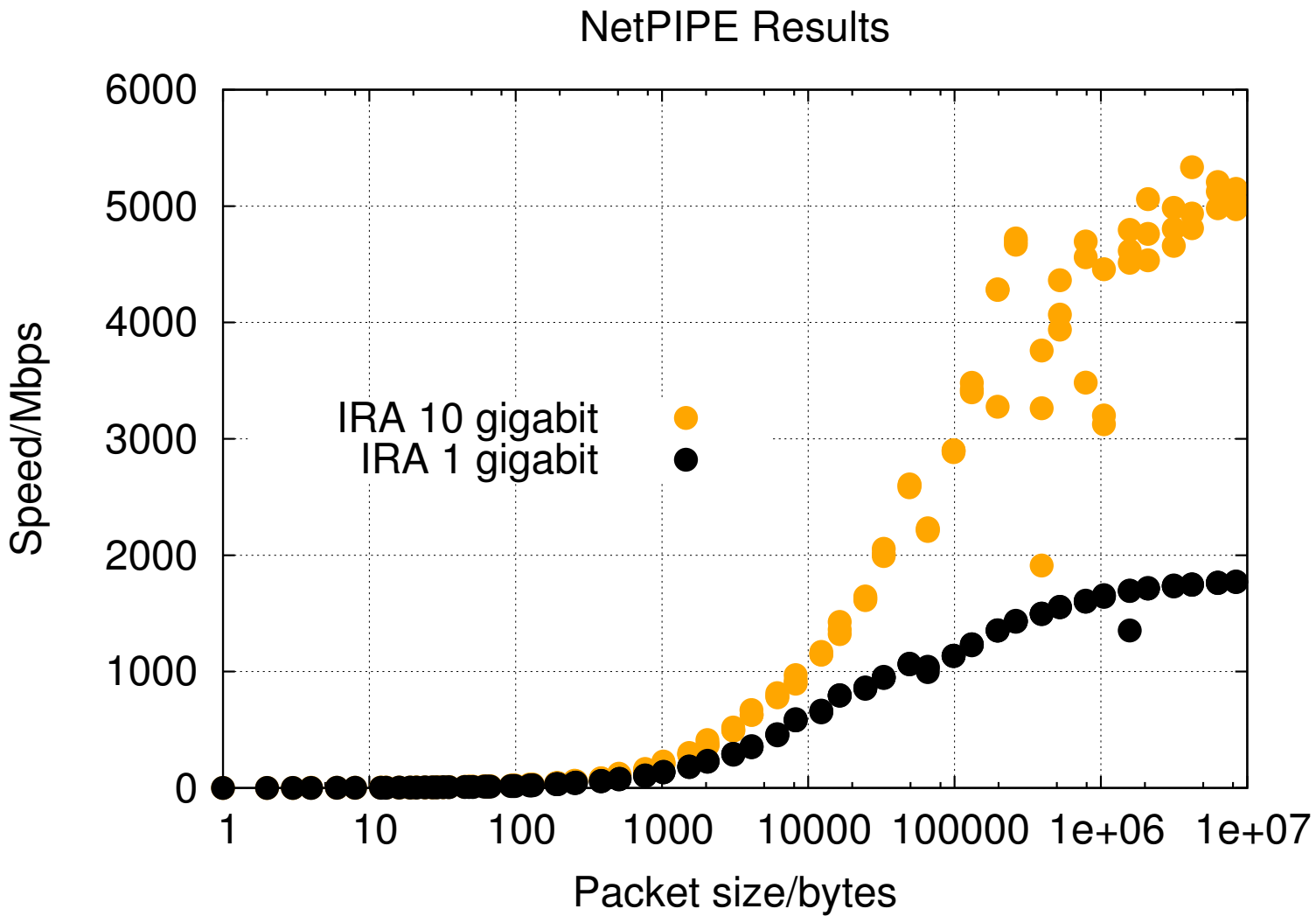
Then:

- Repeat the DiFX performance analysis on these clusters

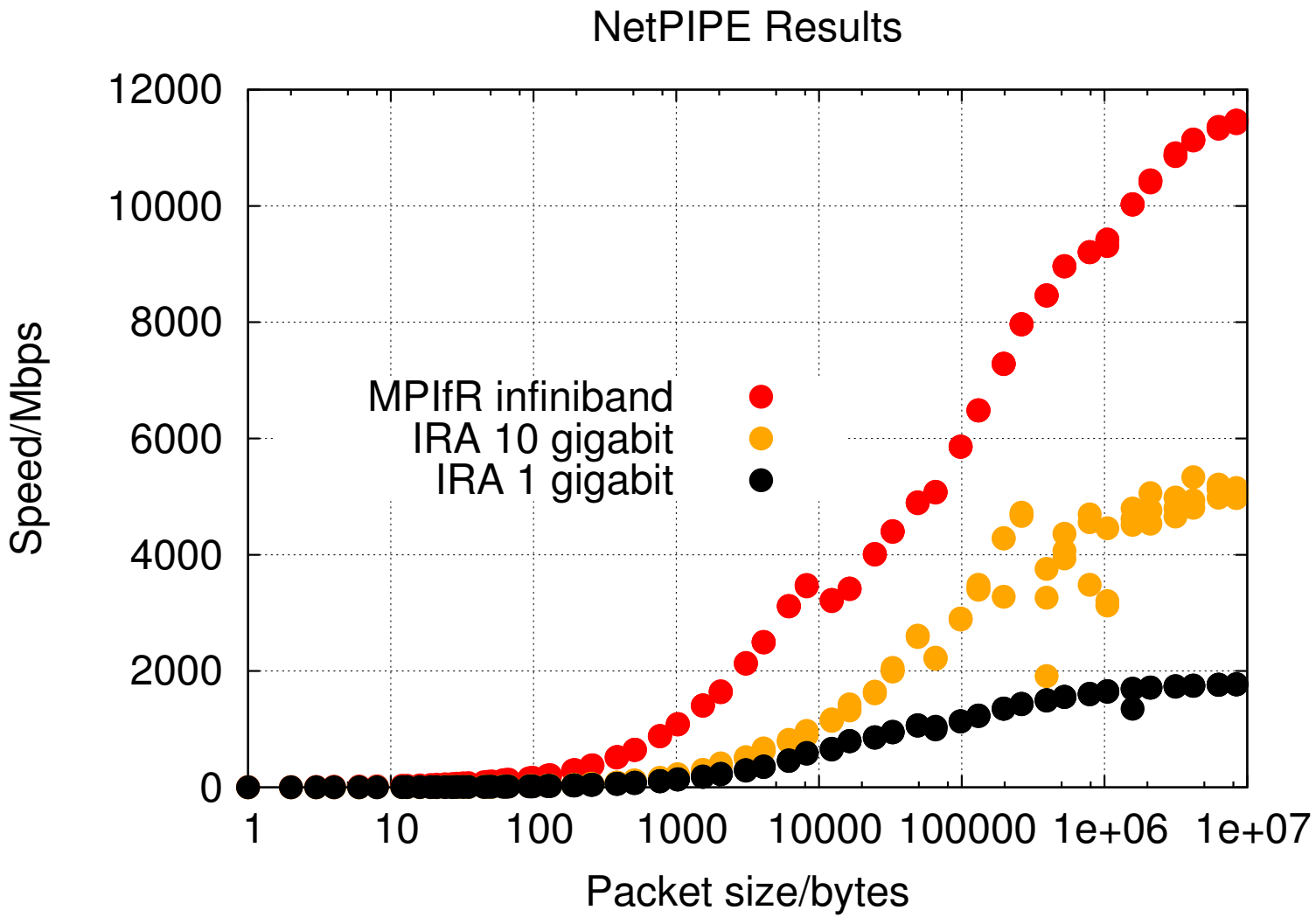
Speed of a TCP connection between two nodes



Speed of an MPI connection between two nodes



Speed of an MPI connection between two nodes



Overview

Network	Latency (microseconds)	Throughput (return time) (Mbps)
1 GbitE	44.8	1770
10 GbitE	33.8	5330
Infiniband	3.5	11500

- Not all high-speed interconnects are equal
- TCP performance doesn't necessarily reflect MPI performance
  - though if it doesn't you should try to fix it!
- Large packets are required for the fastest data transfer rates

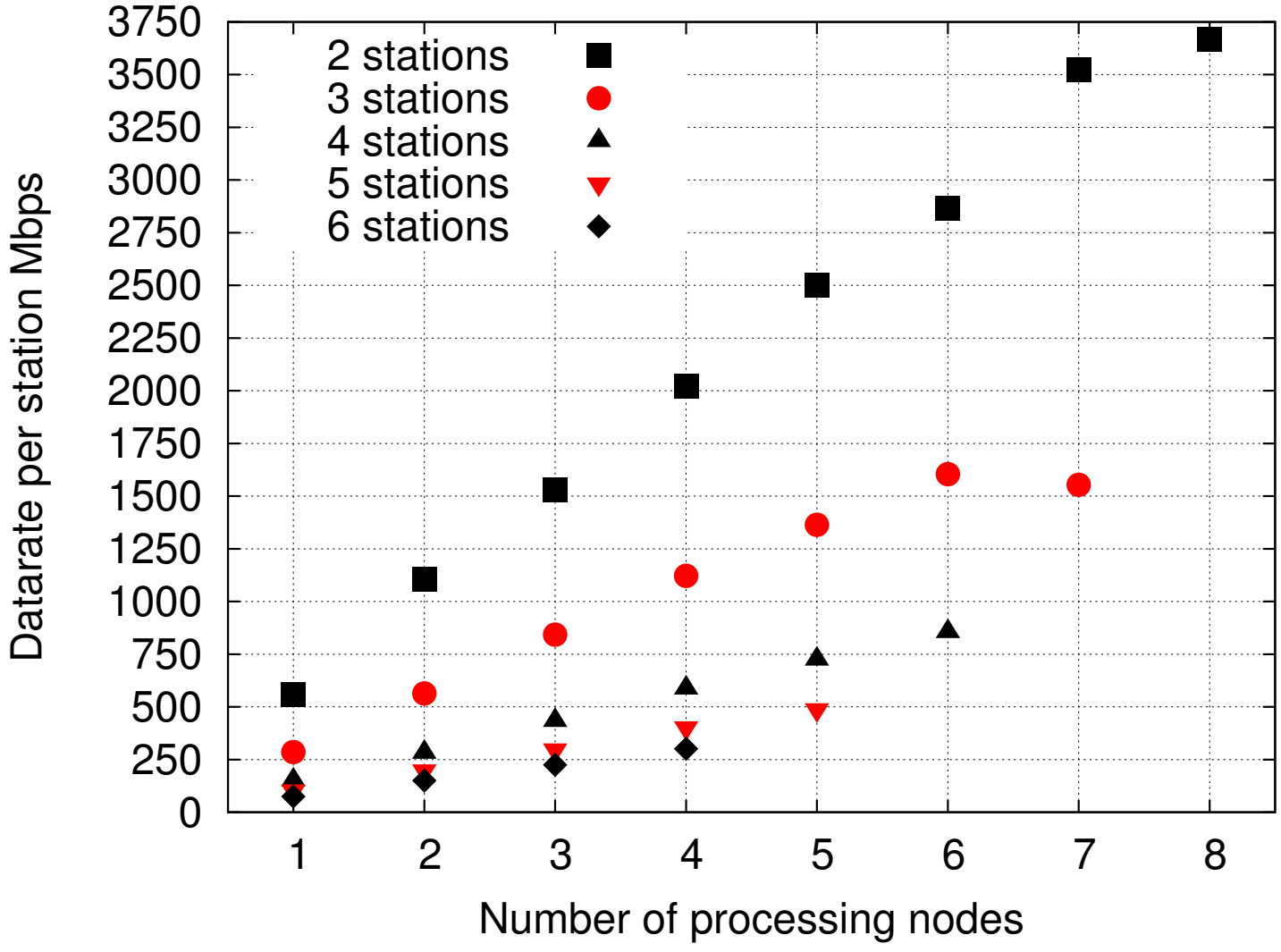
Setup

- Baseband data are read in chunks of 128MB
- Each datastream buffers 12 of these reads (1536 MB)
- 12 MB chunks from each datastream are sent to each core
- These chunks are sub-divided by 7 (in time), each CPU core processes one sub-chunk
- The 7 results are averaged in time (and averaged by a factor of 8 in frequency) and set back to the FXManager
- 40 of these packets are (time) averaged before writing out to disk

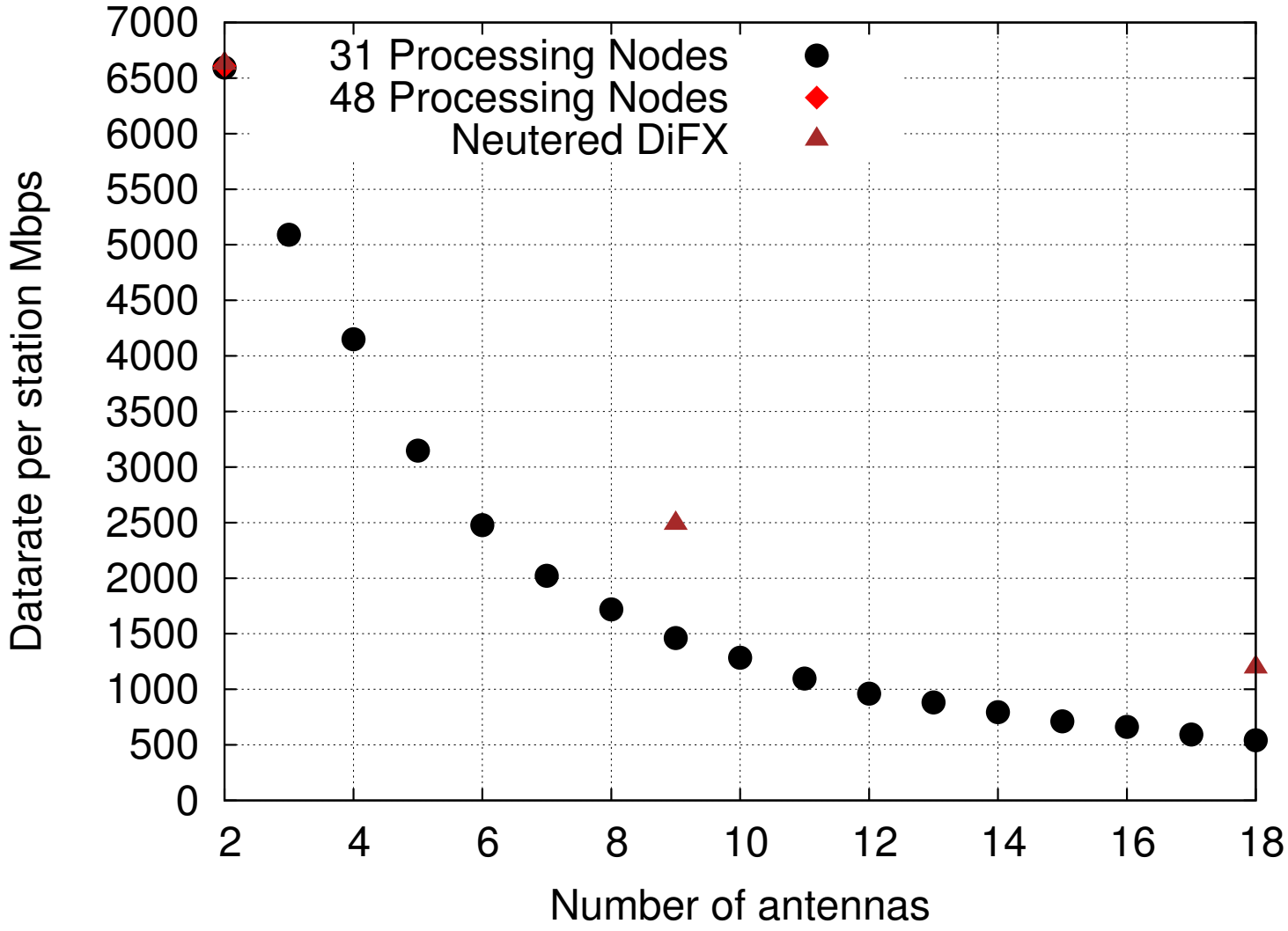
Nodes

- One node is the 'FXManager'
- One node per datastream (antenna) for reading the baseband data
- Remaining nodes are used for processing
- The 10 Gbps cluster had only 11 nodes available
- The infiniband cluster has over 50

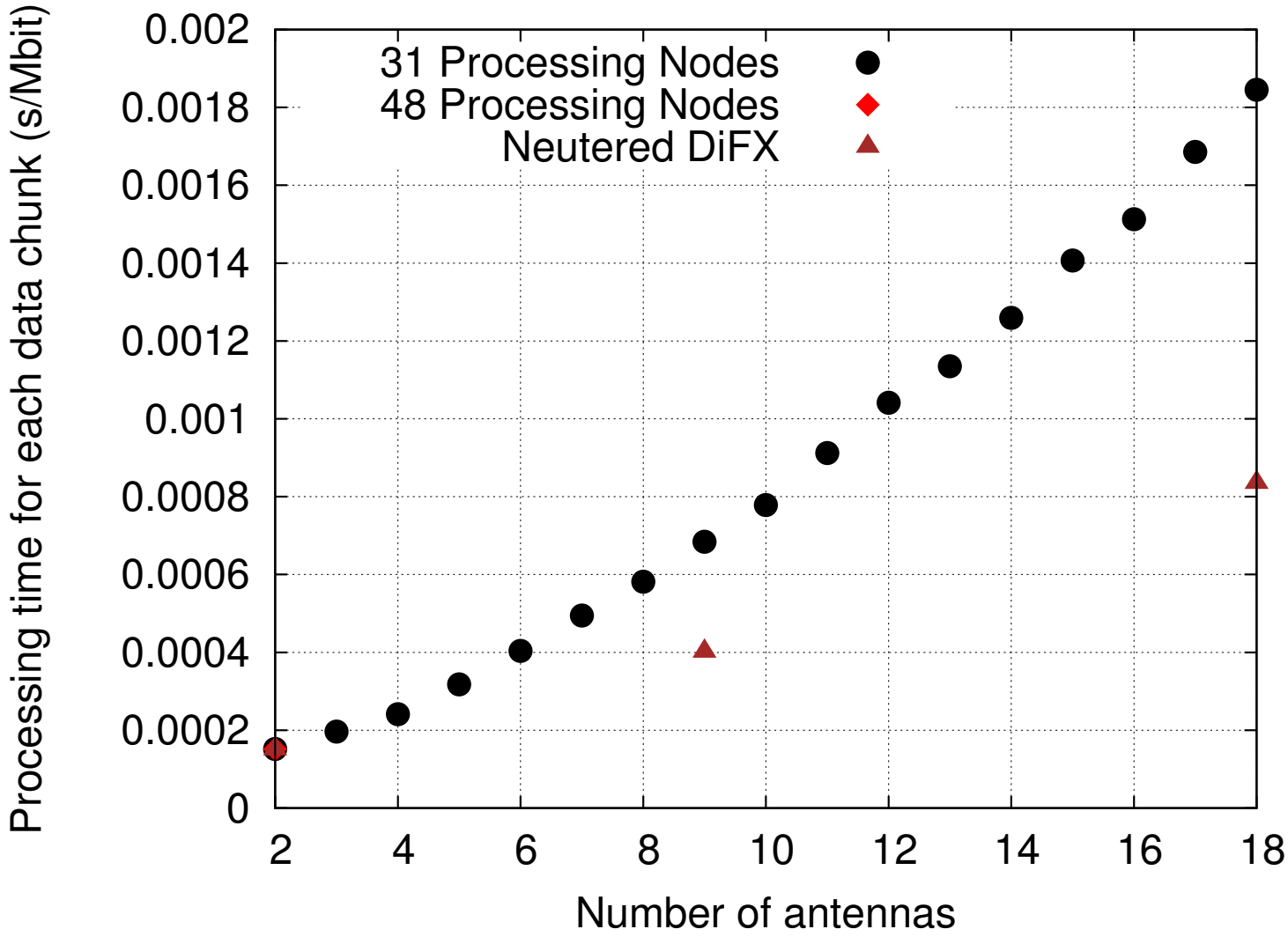
Speed of DiFX 10 Gbps Ethernet



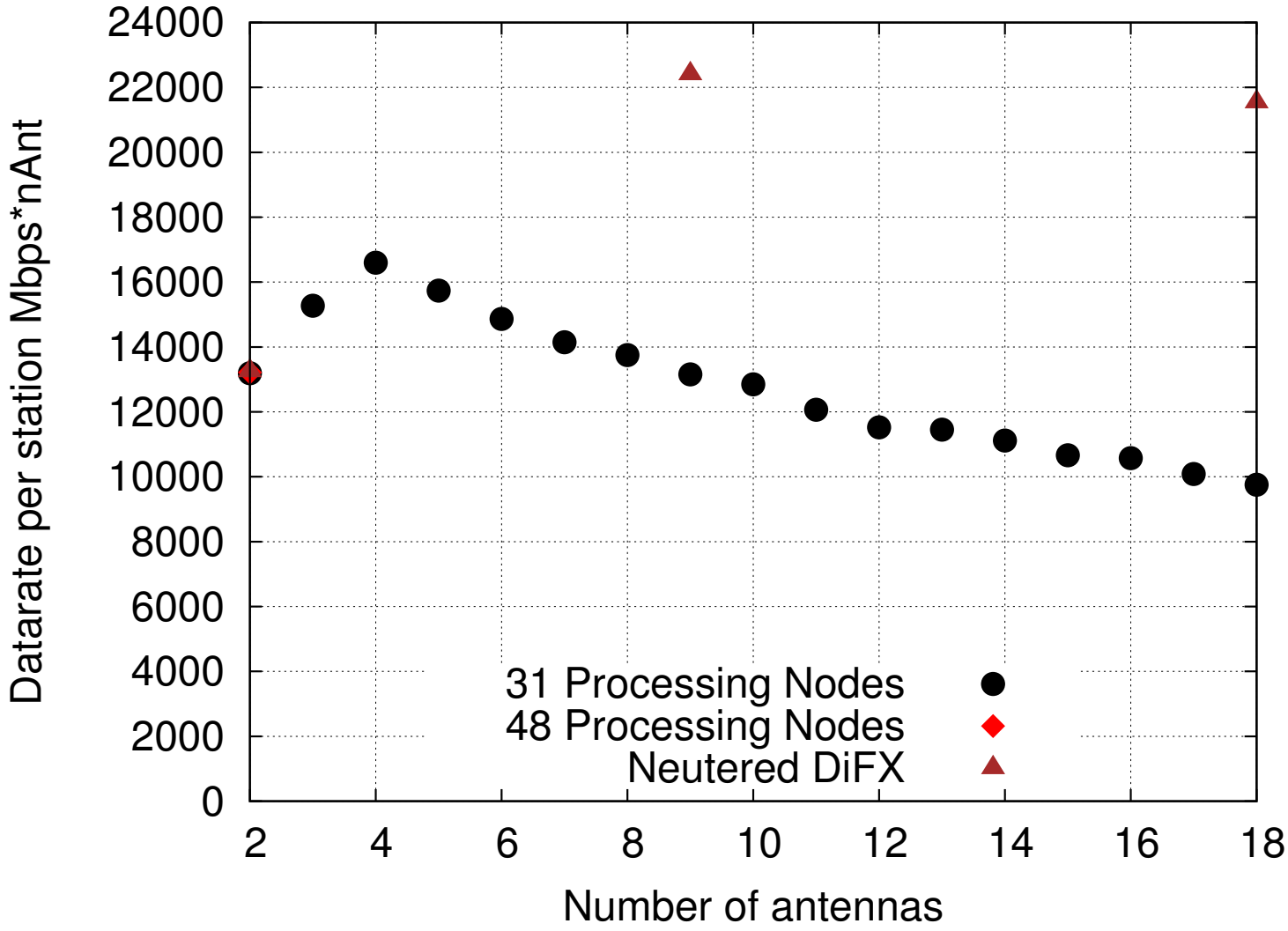
Speed of DiFX infiniband



Time taken to process a given observation period



Total data from datastreams to cores



Conclusions

- Different high-speed interconnects have different advantages.
- High-speed interconnects can supply over 6× as many nodes as 1 Gbit ethernet without bandwidth multiplexing
- It is very important to understand the limitations of high speed interconnects if you intend to use them at close to maximum speed
  - high speed is only reachable with large sends
- The DiFX software correlation platform will remain robust as data rates → 10 Gbps
  - multiple datastreams per antenna planned for a future release
  - other limitations on very high sampling rates to be overcome

DiFX as a Baseband Data Processing System

This is also a test of the Software (Correlator) baseband data processing architecture

- Farm out baseband data using MPI
- Convert to 32 bit floats
- Process using IPP on commodity computing hardware

The use of this isn't only restricted to Correlation.

Think of DiFX as an highly flexible, fully programmable back-end

(See also Mauro Nanni's talk)

DiFX as a Baseband Data Processing System

Useful not only for what can be done now

- Cross correlation/Autocorrelation with almost arbitrary resolution
  - frequency
  - time
- Pulsar Binning
- Phase cal. tone extraction

Or what may be developed in the future

- Online RFI flagging
- Phased array mode

But anything you want it to do!

- FPGA/ASIC back-end prototyping
- SETI
- More complex pulsar analysis
- One-off experiments