Performance Analysis of Filesystem IO using HDF5 and ADIOS on a Cray XC30

Jason Wang
Christopher Harris
Andreas Wicenec

CUG2014, Lugano

- iVEC provides large scale scientific computing resources for researchers in Australia and beyond
- iVEC is an unincorporated joint venture between
  - CSIRO
  - Curtin University
  - Edith Cowan University
  - Murdoch University
  - The University of Western Australia
- and is supported by the Western Australian government

iVEC Pawsey Centre

- International Centre for Radio Astronomy Research
- Since its launch in September 2009 ICRAR has emerged as a major new international centre of excellence in astronomical science and technology.
- ICRAR is a joint venture between
  - The University of Western Australia
  - Curtin University
- and is supported by the Western Australian government

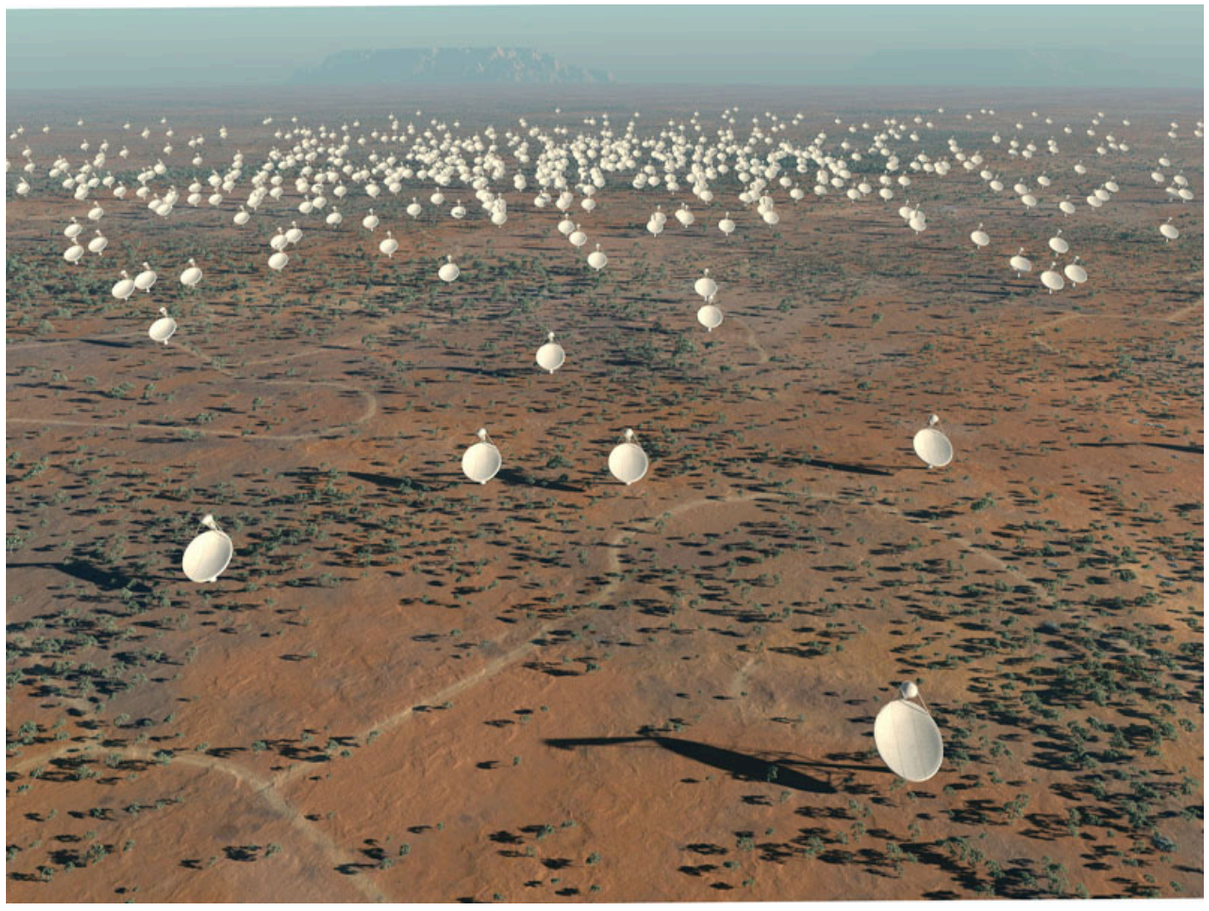
The Square Kilometer Array

- The largest radio telescope in the world
- €1.5 billion project
- 11 member countries
- Timeline
  - **2016** Phase 1 prototype systems deployed
  - **2018-2023** Phase 1 constructed
  - **2023-2030** Phase 2 constructed
- Currently conceptual design & preliminary benchmarks

- Compute Challenge:
  - 100 PFLOPS

- Data Challenge:
  - ExaBytes per day
  - 1 EB = 10^{18} Bytes

- iVEC and ICRAR both heavily involved
  - iVEC in COMP module of SKA Science Data Processor
  - ICRAR leading the DATA module of SKA Science Data Processor



Artist's impression of the SKA dishes. Credit: SKA Organisation

Signal Correlation in Radio Astronomy

$$R_{s_i s_j}(\omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s_i^*(t)s_j(\tau+t)e^{-j\omega\tau}dtd\tau.$$

$$= R_i^*(\omega)R_j(\omega)$$

Software implementation:

1. Unpack (low-byte integer -> 8-byte complex floating point)
2. FFT (Fast Fourier Transform)
3. CMAC (Conjugate Multiplication and Accumulation)

Data Rate

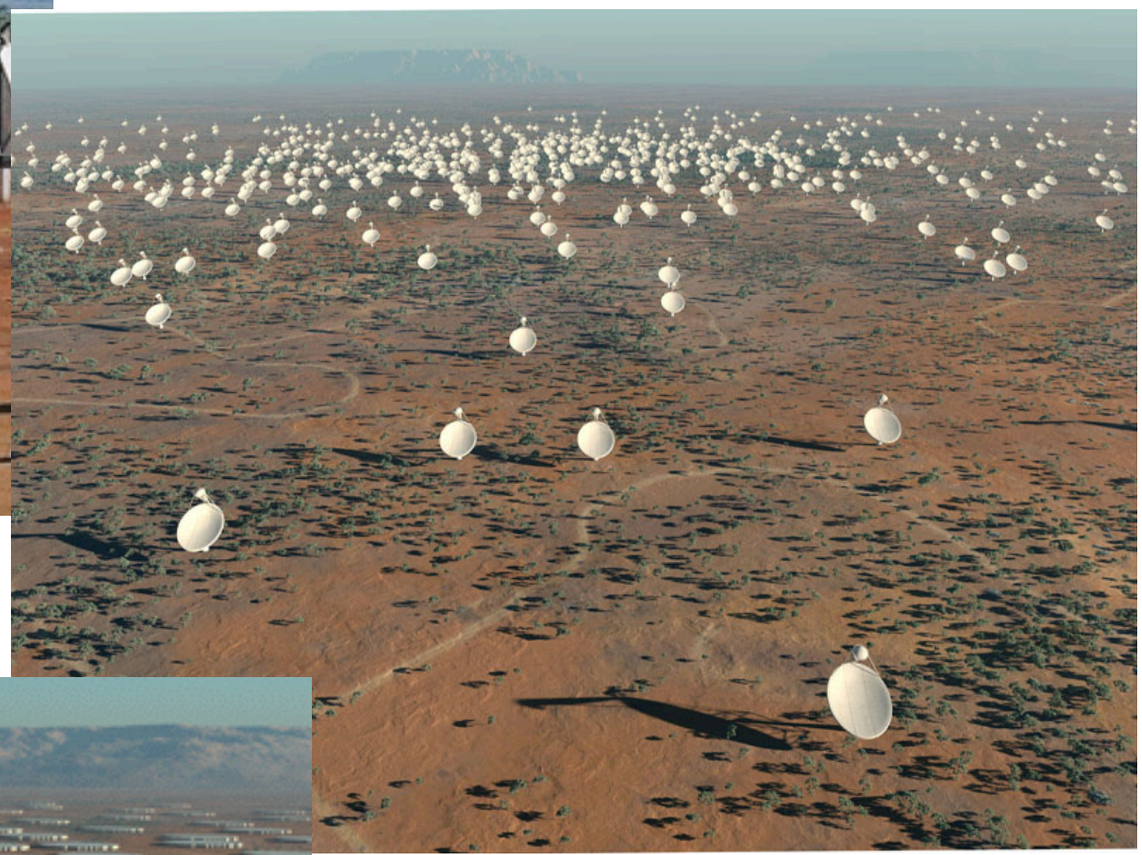
- N - number of input data streams (generally two data streams for polarisations from a single antenna)
- F - number of frequency channels per visibility
- I - number of integrated visibilities per second

$$R = 4FIN(N+1) \text{ Bytes/s}$$

SKA Problem Size

Artist's impression of the SKA dishes & Aperture Arrays. Credit: SKA Organisation

SKA Phase 1 Low Frequency:
900 Station x 300 Antennas/Station



SKA Phase 1 Mid: 254 Antennas

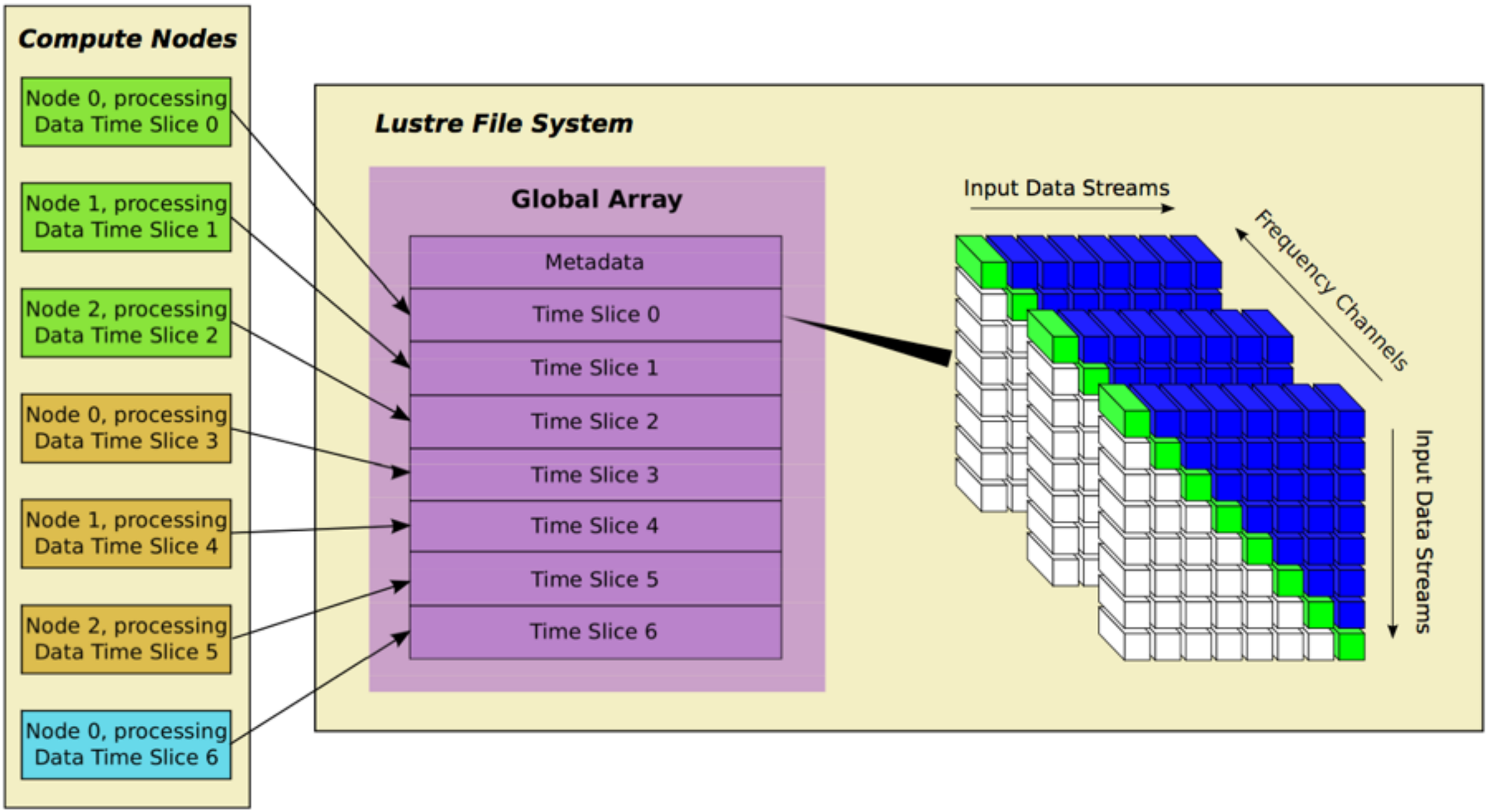
Phase II = Phase I x 10

Possible Solutions

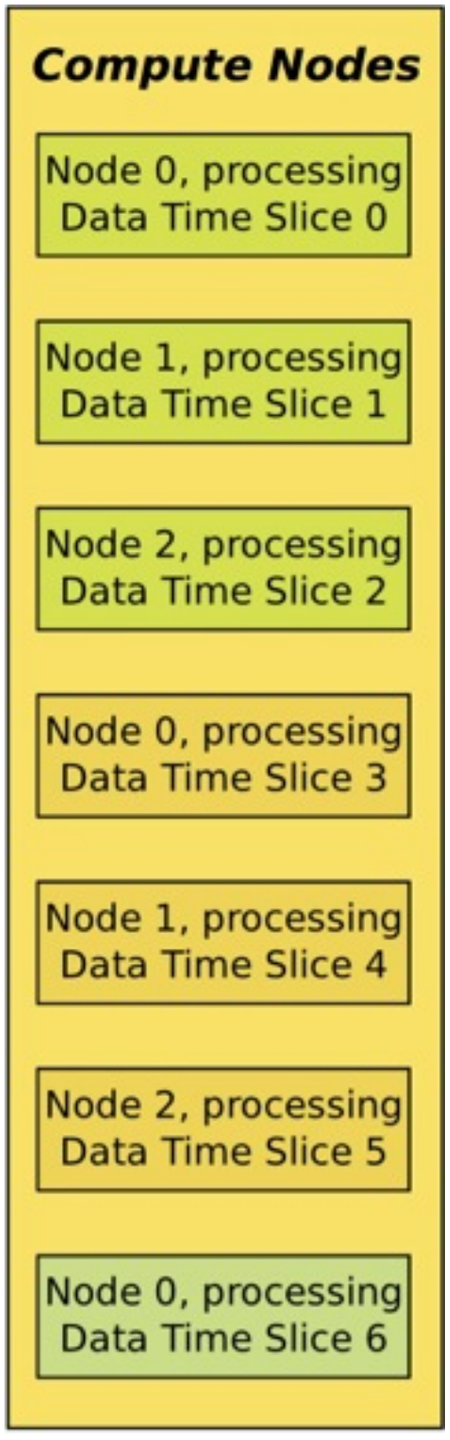
- Hardware (FPGA)
  - Power efficient and stable
  - Fixed work flow, usually without keeping visibility (correlator output) data
- Software (CPU/GPU clusters)
  - Division Model
    - Time division multiplex (DiFX correlator & our previous work)
    - Baseline division multiplex (Our previous work)
    - Frequency division multiplex (MWA telescope from ICRAR)
  - Output
    - Streaming to the next stage without writing to disk storage
    - Writing to local storage of compute nodes (MWA telescope)
    - **Writing to global storage / Lustre (this work)**

Murchison Widefield Array (MWA)
Photography by Paul Bourke and Jonathan Knispel.
Supported by WASP (UWA), iVEC, ICRAR, and CSIRO.

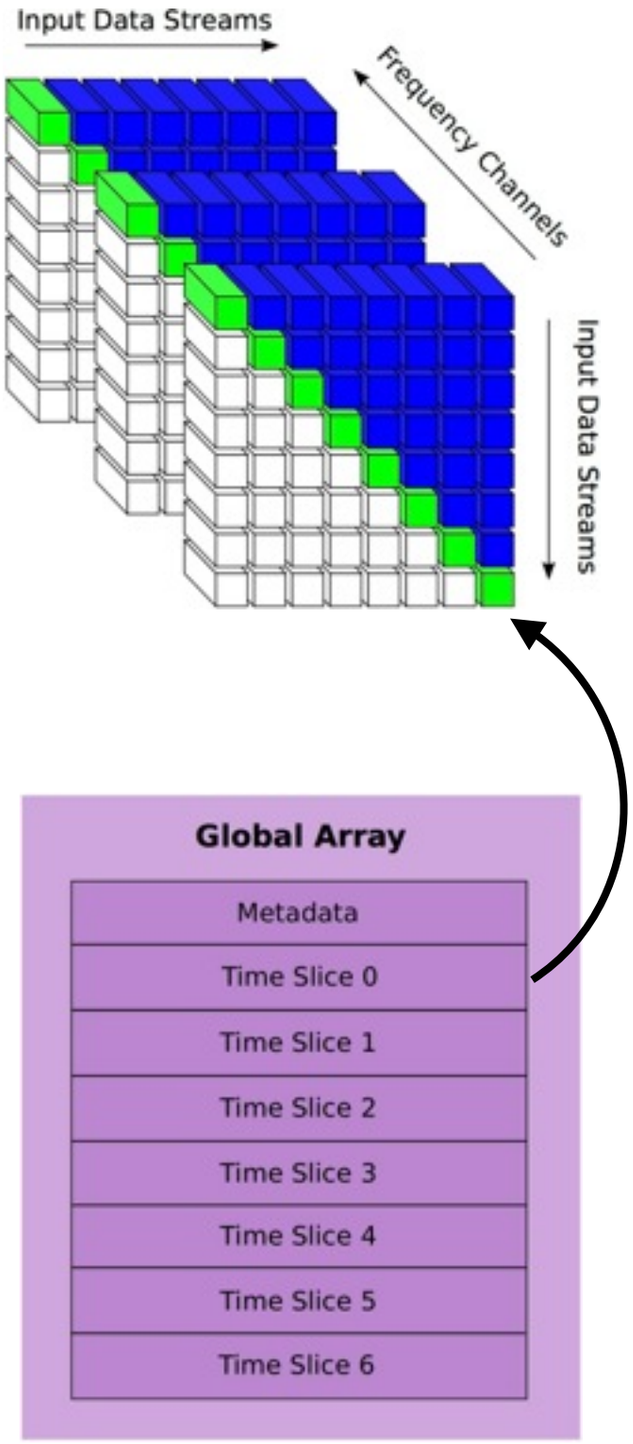
A time-division correlator writing output data to a Lustre filesystem



Testing Parameter Ranges



Parameter	Range
Number of Frequency Channels	128 - 1024
Number of Input Data Streams	100 - 400
Compute Nodes	20 - 90
Lustre Stripe Size	1 - 8
Number of Time Slices	100 - 400
Time Slice Size	5MB - 650MB
Global Array Size	500MB - 263GB



Testbed - iVEC Magnus

- Cray Cascade XC30, with Aries dragonfly interconnect
- Arrives in two stages, first stage in production since Jan 2014
- 3,328 processing cores
- 13.3 terabytes of memory
- 2 petabytes of storage
- 72 gigabits per second interconnect

Testbed - iVEC Magnus (Storage)

- Cray Sonexion 1600
- Two petabytes of storage via nine Scalable Storage Units (SSUs)
- 8 OSTs, each using a 8+2 RAID 6 configuration
- The specification of each SSU has a 5 GB per second bandwidth from the IOR benchmark, and thus the expected peak bandwidth is 45 GB per second.

Reference System - iVEC Fornax

- SGI system for data intensive processing
- Specialised architecture includes:
  - Graphics Processing Units (GPUs)
  - High memory per node
  - Local storage
  - Dual Infiniband interconnect
- In production since July 2012
- 1,152 processing cores (96 compute nodes)
- 43,008 graphics processing cores
- 7.1 terabytes of memory
- 480 terabytes of storage
- 80 gigabits per second interconnect

Reference System - iVEC Fornax (Storage)

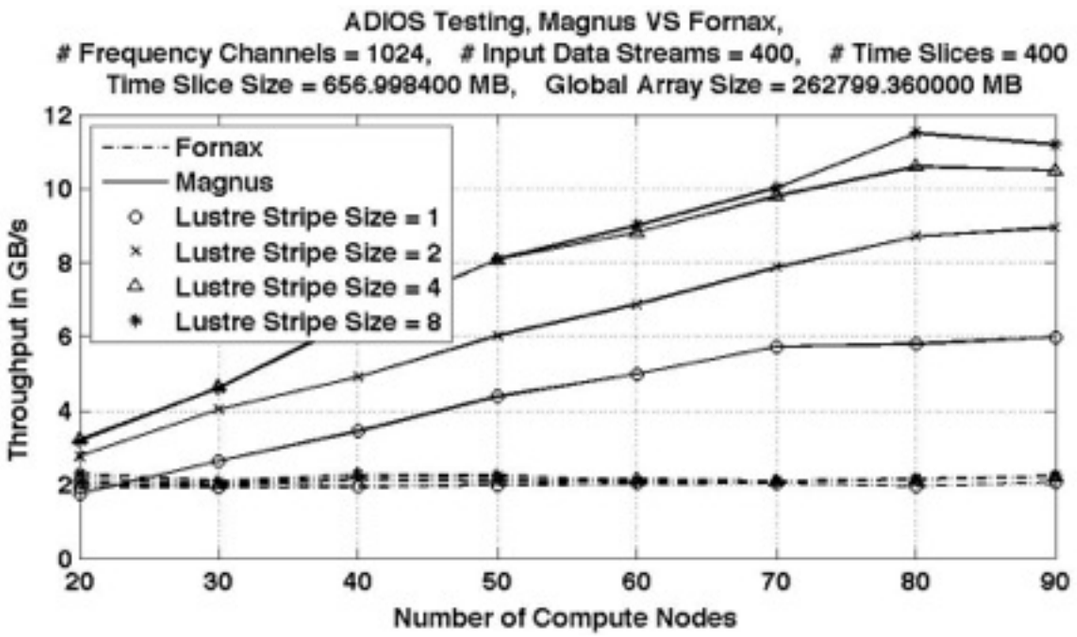
- SGI Infinite S16k, (re-badged DDN SFA 10k)
- 8 Object Storage Servers (OSSs) and 44 Object Storage Targets (OSTs), of which 32 are assigned to the scratch file system used in this testing.
- 16 4x QDR Infiniband connections to the switch connecting compute nodes,
- 8 4x QDR Infiniband connections between OSTs and OSSs
- Each OST consists of 10 Hitachi Deskstar 7K2000 hard drives arranged into a 8+2 RAID 6 configuration.
- Operational testing using the ost_survey Lustre benchmark achieved a mean bandwidth of 343 MB per second, and thus the expected bandwidth is approximately 11 GB per second

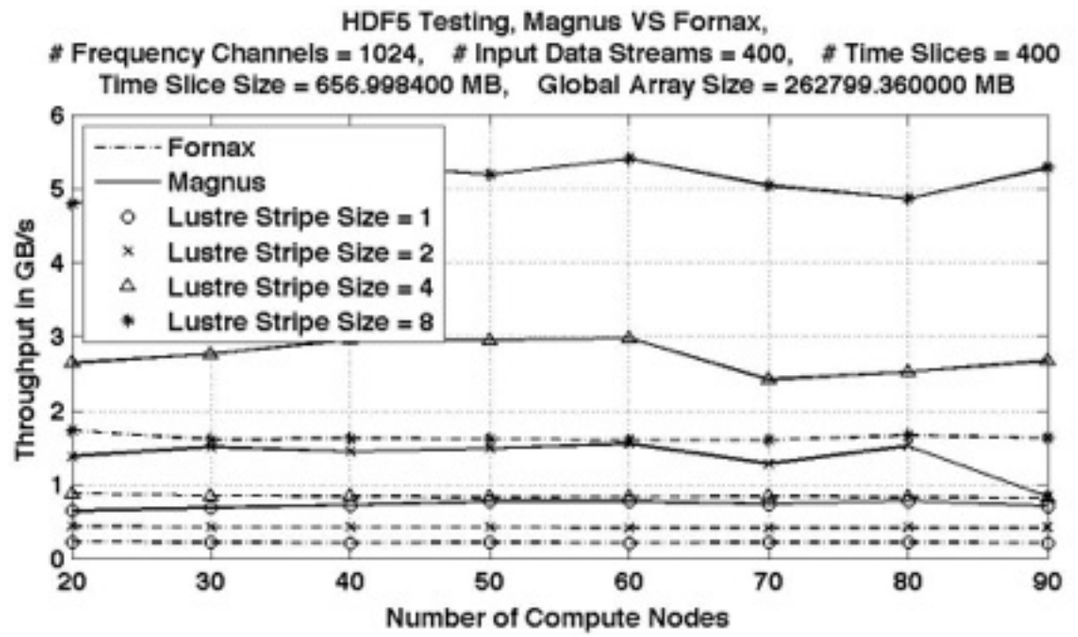
IO Library - ADIOS & HDF5

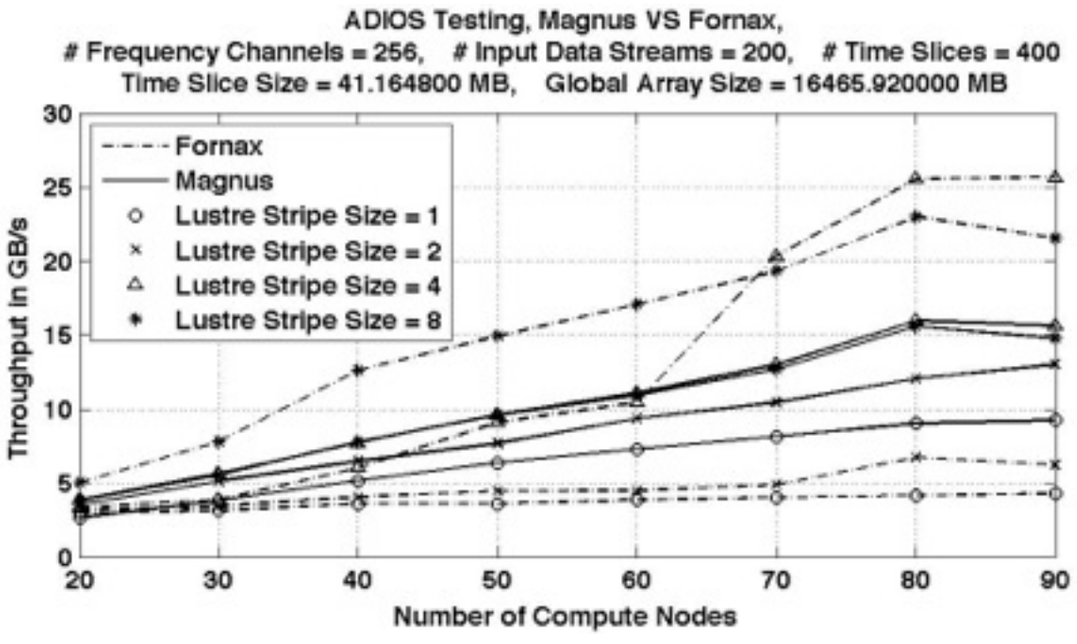
- Adaptive IO System (ADIOS)
  - An ORNL product
  - Designed for ultra-large scale parallel IO
  - Flexible & convenient XML configuration

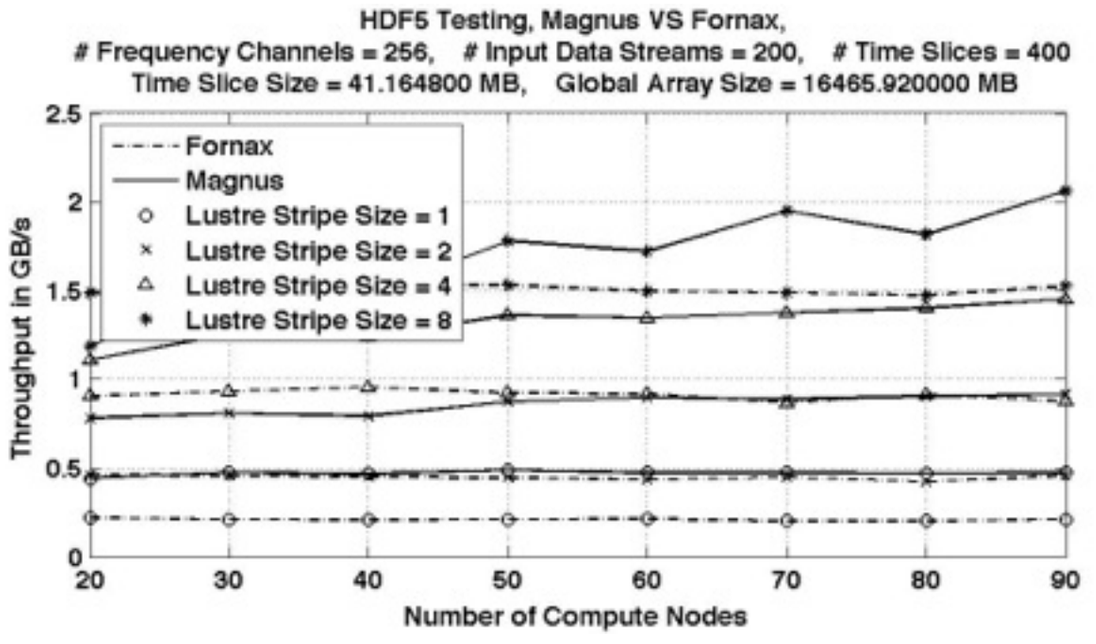
- Hierarchical Data Format 5 (HDF5)
  - Widely used format for scientific data

ADIOS & HDF5

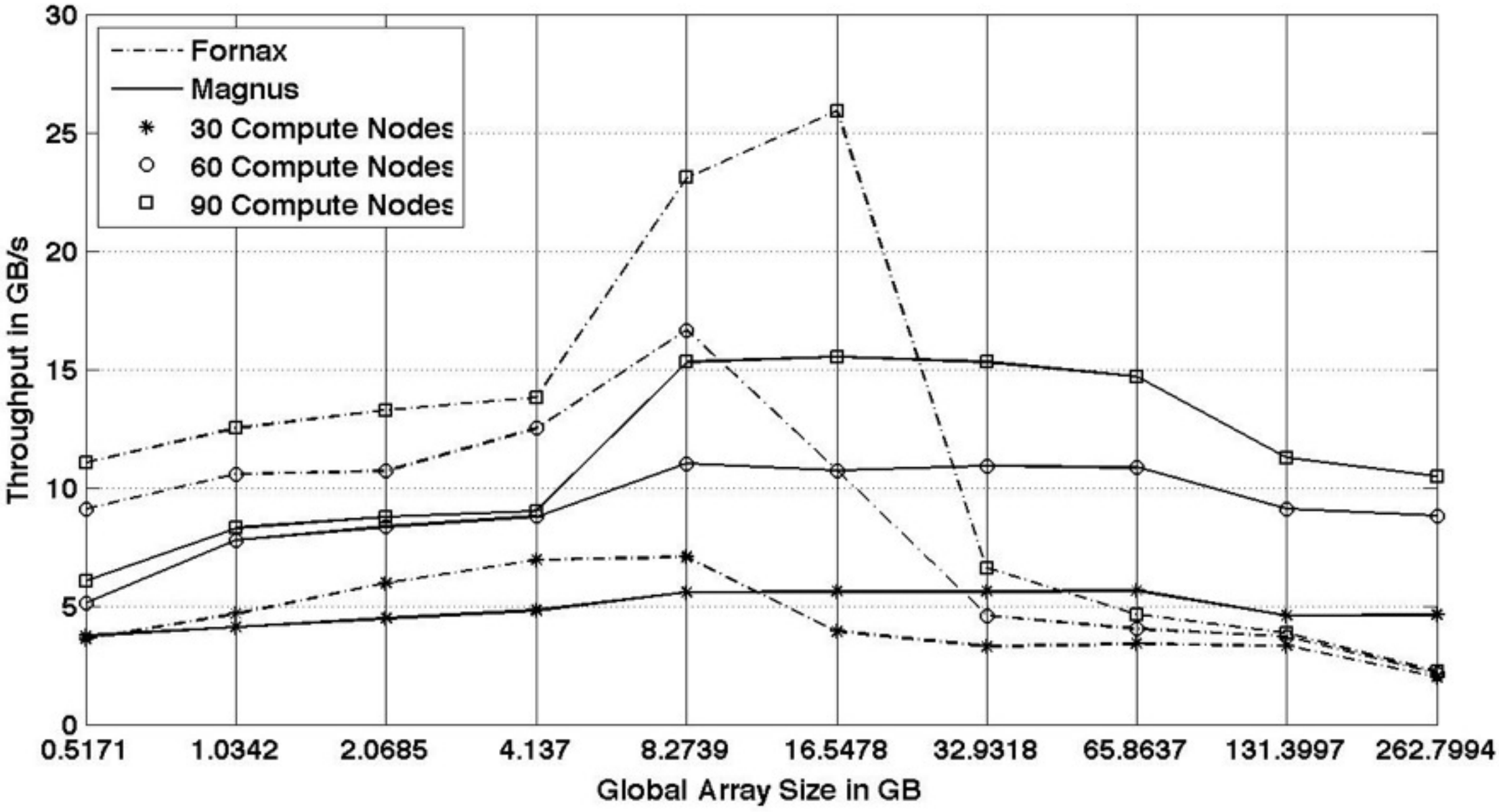

ADIOS Testing, Magnus VS Fornax,
Frequency Channels = 1024, # Input Data Streams = 400, # Time Slices = 400
Time Slice Size = 656.998400 MB, Global Array Size = 262799.360000 MB
Fornax
Magnus
Lustre Stripe Size = 1
Lustre Stripe Size = 2
Lustre Stripe Size = 4
Lustre Stripe Size = 8
Throughput in GB/s
Number of Compute Nodes


HDF5 Testing, Magnus VS Fornax,
Frequency Channels = 1024, # Input Data Streams = 400, # Time Slices = 400
Time Slice Size = 656.998400 MB, Global Array Size = 262799.360000 MB
Fornax
Magnus
Lustre Stripe Size = 1
Lustre Stripe Size = 2
Lustre Stripe Size = 4
Lustre Stripe Size = 8
Throughput in GB/s
Number of Compute Nodes

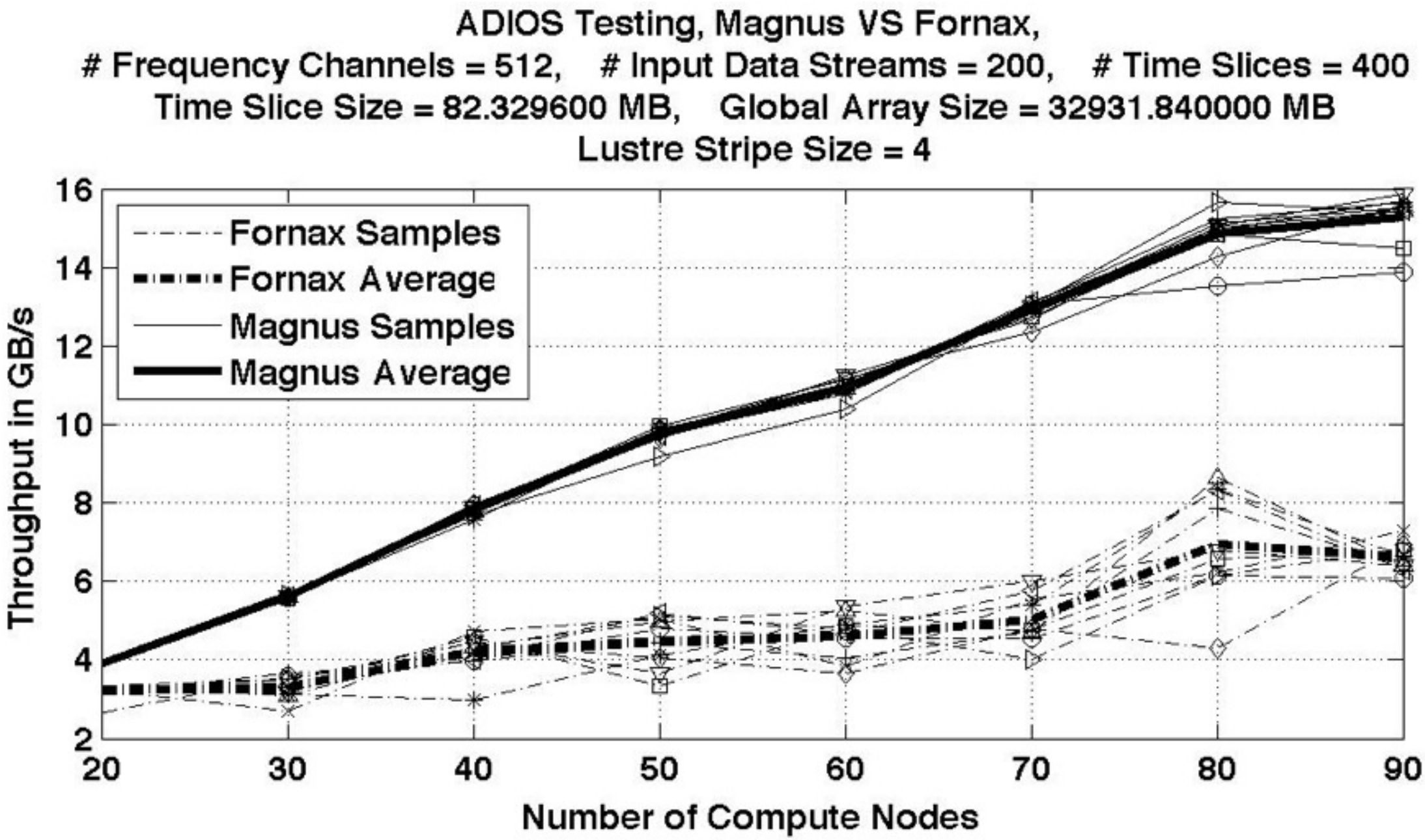

ADIOS Testing, Magnus VS Fornax,
Frequency Channels = 256, # Input Data Streams = 200, # Time Slices = 400
Time Slice Size = 41.164800 MB, Global Array Size = 16465.920000 MB
Fornax
Magnus
Lustre Stripe Size = 1
Lustre Stripe Size = 2
Lustre Stripe Size = 4
Lustre Stripe Size = 8
Throughput in GB/s
Number of Compute Nodes


HDF5 Testing, Magnus VS Fornax,
Frequency Channels = 256, # Input Data Streams = 200, # Time Slices = 400
Time Slice Size = 41.164800 MB, Global Array Size = 16465.920000 MB
Fornax
Magnus
Lustre Stripe Size = 1
Lustre Stripe Size = 2
Lustre Stripe Size = 4
Lustre Stripe Size = 8
Throughput in GB/s
Number of Compute Nodes

File Size Impact



Stability


ADIOS Testing, Magnus VS Fornax,
Frequency Channels = 512, # Input Data Streams = 200, # Time Slices = 400
Time Slice Size = 82.329600 MB, Global Array Size = 32931.840000 MB
Lustre Stripe Size = 4
Fornax Samples
Fornax Average
Magnus Samples
Magnus Average
Throughput in GB/s
Number of Compute Nodes
2
4
6
8
10
12
14
16
20
30
40
50
60
70
80
90

Conclusion & Future work

- This is an early-stage investigation into file formats & Lustre storage. The testing results are included as part of the data benchmarks for the SKA Science Data Processor.
- In future, we are looking at
  - Larger testing scale (Upgraded peta-scale Cray XC30 at iVEC)
  - Other storage backends (local storage, object storage such as Ceph)

Thank You!