Big data to the other side of the World
Global Big Data and Compute at 100G
Overview

• NCI
• Copernicus
• Data transfer challenges
• Investigations with ICM
• Conclusion
Australia’s National Computational Infrastructure

• One of the two research HPC facilities in Australia
• Bringing together big data and compute
• National and Regional data custodian for many large data collections
  • CMIP
  • Asia regional Copernicus hub
• Refresh and upgrade of compute resources in progress for implementation mid 2019
- **Raijin**
  - 84,656 cores (Intel Xeon Sandy Bridge 2.6 GHz, Broadwell 2.6 GHz) in 4416 compute nodes
  - 120 NVIDIA Tesla K80 GPUs in 30 nodes and 8 NVIDIA Tesla P100 GPUs in 2 nodes
  - 32 Intel Xeon Phi (64 core Knights Landing, 1.3 GHz) in 32 compute nodes
  - 4 IBM POWER8 nodes (64 cores running at 4.02GHz)
  - 300 Terabytes of main memory
  - Hybrid FDR/EDR Mellanox Infiniband full fat tree interconnect (up to 100 Gb/sec)
  - 8 Petabytes of high-performance operational storage capacity
  - FDR (56G) and EDR (100G) InfiniBand interconnect
  - Full Fat Tree
Cloud

- Tenjin
  - 33.5 Teraflop Private Cloud
  - OpenStack
  - 56G Ethernet
- Nectar
  - National research cloud
  - OpenStack
  - 56G Ethernet (Fat Tree)
- VMware
  - 10G Ethernet switched
Storage

• “The fastest filesystems in the Southern Hemisphere"
  • Disk
  • Lustre File systems
    • 140GB/s sustained I/O throughput
    • 37.1Pb over multiple filesystems
    • 56G FDR InfiniBand interconnect
  • Tape
    • Dual site….. moving to two + cold archive
Networks

- 10G, 56G and 100G Ethernet
- 40G (QDR), 56G (FDR) and 100G (EDR) InfiniBand
- Data Centre
- Science DMZ
- National (AARNet)
- International
- Monitoring, Operations and Design
- Big data transfers
• High bandwidth, medium (domestic) to high (international) latency networks
• Traditional transfer tools perform at one to two orders of magnitude below available bandwidth
  • The greater the distance the lower the performance
• Science DMZ handles the “last mile” problem of firewalls, traffic shapers and other network restricting devices but is not the “magic bullet” for big data movement
By default TCP/IP does not perform well over high bandwidth, high delay circuits.
Problem Space

• Most remote data servers use protocols designed in the 90’s (http)
  • Tools are optimised and tested within a regional environment
    • Europe, Americas, Asia
    • Oceania (Australia and New Zealand) are large data consumers
  • As a TCP/IP based protocol they suffer from TCP related tuning issues
    • Window Sizes
    • Congestion control
    • Highly sensitive to even small <1% packet loss with severe performance degradation
  • All sites must be tuned
  • DTN to DTN is performant but is a small subset of the transfer requirement
  • These do not scale well on a global scale
• Commercial response by FANG (Facebook, Amazon, Netflix, Google) is global distributed CDN network to bring data closer to the user
Mice and Elephant Flows

- NREN
- Science DMZ
- Production
- DTN
- HPC
- Storage Fabric
  - InfiniBnd
- InfiniBnd
- Mice
- Elephant
Mice and Elephant Flows
Copernicus

- **Satellites Pairs**
  - **Sentinel-1**: polar-orbiting, all-weather, day-and-night radar imaging mission for land and ocean services
  - **Sentinel-2**: polar-orbiting, multispectral high-resolution imaging mission for land monitoring
  - **Sentinel-3**: multi-instrument mission to measure sea-surface topography, sea- and land-surface temperature, ocean colour and land colour with high-end accuracy and reliability
  - **Sentinel-5 Precursor** – also known as Sentinel-5P: forerunner of Sentinel-5 to provide timely data on a multitude of trace gases and aerosols affecting air quality and climate
  - **Sentinel-4**: payload devoted to atmospheric monitoring that will be embarked upon a Meteosat Third Generation-Sounder (MTG-S) satellite in geostationary orbit
  - **Sentinel-5**: payload will monitor the atmosphere from polar orbit aboard a MetOp Second Generation satellite
  - **Sentinel-6**: radar altimeter to measure global sea-surface height
Copernicus Big data in small pieces

• Data is delivered as a set of small and large files
  • Thumbnails
  • ZIP files containing satellite images
Network Path NCI to ICM Poland

370ms

AARNet

NORDUnet
Caching

- Accelerating http using Squid
- Regional delay products and source tuning limits effective download rate
  - ICM Poland - Authenticated squid
  - Direct cache (AARNet4 10G)
    - Client at NCI requests download, squid at ICM manages first party transfer from ESA and pipelines the flow to NCI
    - 2x - 3x transfer rate increase over same network
  - Tired Cache (AARNet-X 100G)
    - Selectively control the data routing to 100G service
AARNet-X Design

• Mellanox SN2100 (16 port 100G) switches
  • x86 CPU
  • Spectrum chipset
• Network OS
  • Cumulus Linux
• 200G national core carried over DWDM OTN network
• L2 and L3 transport
• Low cost of 100G edge ports
Sentinel Download method

- Object based
  - Mix of small (40Kb png), medium (50-70Mb zip) and large (1-2Gb zip)
- **https** GET used for transfer protocol
- Periodic (15 minute) queries to ESA hubs
  - Returns a list of downloadable products
- Each hub (S1,S2,S4,S4,S5) has different available download quotas and are tuned for European delay products and 1500 MTU.
- Republished product sets may appear without notice
- Highly parallel
• 10G AU
  • BPD 351ms * 10,000,000,000/8 = 442.5 M
• 100G AU
  • BPD 354ms * 100,000,000,000/8 = 4.4 G
• Direct from Australia
  • ~2.5Mbs
• Using ICM proxy
  • ~5-7Mbs
  • 2-3x improvement
Download Results:

- gid   I stat I avg speed I % I path/URI

- e215a9 I OK I 3.7MiB/s I 100 I scihub_S3/S3A_SL_2_LST____20180531T201936_20180531T202236_20180531T212328_0179_031_385_0180_SVL_O_NR_003.zip

Status Legend:
### Effect of Congestion control Mismatch

- **NCI HTCP, ICM CUBIC**

<table>
<thead>
<tr>
<th>ID</th>
<th>Interval</th>
<th>Transfer</th>
<th>Bandwidth</th>
<th>Retr</th>
<th>Cwnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.00-10.00 sec</td>
<td>6.81 MBytes</td>
<td>5.71 Mbits/sec</td>
<td>0</td>
<td>696 KBytes</td>
</tr>
<tr>
<td>4</td>
<td>10.00-20.00 sec</td>
<td>89.8 MBytes</td>
<td>75.4 Mbits/sec</td>
<td>0</td>
<td>7.41 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>20.00-30.00 sec</td>
<td>221 MBytes</td>
<td>186 Mbits/sec</td>
<td>2109</td>
<td>4.64 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>30.00-40.00 sec</td>
<td>188 MBytes</td>
<td>157 Mbits/sec</td>
<td>0</td>
<td>9.47 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>40.00-50.00 sec</td>
<td>400 MBytes</td>
<td>336 Mbits/sec</td>
<td>0</td>
<td>19.6 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>50.00-60.00 sec</td>
<td>778 MBytes</td>
<td>652 Mbits/sec</td>
<td>0</td>
<td>37.1 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>60.00-70.00 sec</td>
<td>258 MBytes</td>
<td>216 Mbits/sec</td>
<td>11728</td>
<td>19.8 KBytes</td>
</tr>
<tr>
<td>4</td>
<td>70.00-80.00 sec</td>
<td>5.00 MBytes</td>
<td>4.19 Mbits/sec</td>
<td>554</td>
<td>540 KBytes</td>
</tr>
<tr>
<td>4</td>
<td>80.00-90.00 sec</td>
<td>63.8 MBytes</td>
<td>53.5 Mbits/sec</td>
<td>0</td>
<td>6.05 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>90.00-100.00 sec</td>
<td>339 MBytes</td>
<td>284 Mbits/sec</td>
<td>0</td>
<td>17.4 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>100.00-110.00 sec</td>
<td>695 MBytes</td>
<td>583 Mbits/sec</td>
<td>0</td>
<td>33.4 MBytes</td>
</tr>
<tr>
<td>4</td>
<td>110.00-120.00 sec</td>
<td>694 MBytes</td>
<td>582 Mbits/sec</td>
<td>13220</td>
<td>16.8 MBytes</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>ID</th>
<th>Interval</th>
<th>Transfer</th>
<th>Bandwidth</th>
<th>Retr</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.00-120.00 sec</td>
<td>3.65 GBytes</td>
<td>261 Mbits/sec</td>
<td>27611 sender</td>
</tr>
<tr>
<td>4</td>
<td>0.00-120.00 sec</td>
<td>3.62 GBytes</td>
<td>259 Mbits/sec</td>
<td>sender</td>
</tr>
</tbody>
</table>

---

*nci.org.au*
### NCI HTCP, ICM HTCP

<table>
<thead>
<tr>
<th>[ 4]</th>
<th>Interval</th>
<th>Transfer</th>
<th>Bandwidth</th>
<th>Retr</th>
<th>CWND</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00-10.00 sec</td>
<td>7.63 MBytes</td>
<td>6.40 Mbits/sec</td>
<td>0</td>
<td>679 KBytes</td>
<td></td>
</tr>
<tr>
<td>10.00-20.00 sec</td>
<td>87.1 MBytes</td>
<td>73.0 Mbits/sec</td>
<td>0</td>
<td>7.02 MBytes</td>
<td></td>
</tr>
<tr>
<td>20.00-30.00 sec</td>
<td>291 MBytes</td>
<td>244 Mbits/sec</td>
<td>98</td>
<td>9.13 MBytes</td>
<td></td>
</tr>
<tr>
<td>30.00-40.00 sec</td>
<td>336 MBytes</td>
<td>282 Mbits/sec</td>
<td>0</td>
<td>15.5 MBytes</td>
<td></td>
</tr>
<tr>
<td>40.00-50.00 sec</td>
<td>630 MBytes</td>
<td>528 Mbits/sec</td>
<td>0</td>
<td>29.5 MBytes</td>
<td></td>
</tr>
<tr>
<td>50.00-60.00 sec</td>
<td>392 MBytes</td>
<td>329 Mbits/sec</td>
<td>3753</td>
<td>9.55 MBytes</td>
<td></td>
</tr>
<tr>
<td>60.00-70.00 sec</td>
<td>360 MBytes</td>
<td>302 Mbits/sec</td>
<td>0</td>
<td>16.9 MBytes</td>
<td></td>
</tr>
<tr>
<td>70.00-80.00 sec</td>
<td>675 MBytes</td>
<td>566 Mbits/sec</td>
<td>0</td>
<td>32.1 MBytes</td>
<td></td>
</tr>
<tr>
<td>80.00-90.00 sec</td>
<td>615 MBytes</td>
<td>516 Mbits/sec</td>
<td>3201</td>
<td>10.8 MBytes</td>
<td></td>
</tr>
<tr>
<td>90.00-100.00 sec</td>
<td>380 MBytes</td>
<td>319 Mbits/sec</td>
<td>0</td>
<td>16.9 MBytes</td>
<td></td>
</tr>
<tr>
<td>100.00-110.00 sec</td>
<td>655 MBytes</td>
<td>549 Mbits/sec</td>
<td>0</td>
<td>30.8 MBytes</td>
<td></td>
</tr>
<tr>
<td>110.00-120.00 sec</td>
<td>750 MBytes</td>
<td>629 Mbits/sec</td>
<td>15023</td>
<td>16.6 MBytes</td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>[ 4]</th>
<th>Interval</th>
<th>Transfer</th>
<th>Bandwidth</th>
<th>Retr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00-120.00 sec</td>
<td>5.06 GBytes</td>
<td>362 Mbits/sec</td>
<td>22075</td>
<td>sender</td>
</tr>
<tr>
<td>0.00-120.00 sec</td>
<td>5.03 GBytes</td>
<td>360 Mbits/sec</td>
<td></td>
<td>receiver</td>
</tr>
<tr>
<td>ID</td>
<td>Interval</td>
<td>Transfer</td>
<td>Bandwidth</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-------------</td>
<td>------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00-10.00 sec</td>
<td>702 MBytes</td>
<td>589 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10.00-20.00 sec</td>
<td>1.10 GBytes</td>
<td>942 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20.00-30.00 sec</td>
<td>1.10 GBytes</td>
<td>942 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>30.00-40.00 sec</td>
<td>1.10 GBytes</td>
<td>942 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>40.00-50.00 sec</td>
<td>528 MBytes</td>
<td>443 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>50.00-60.00 sec</td>
<td>461 MBytes</td>
<td>387 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>60.00-70.00 sec</td>
<td>567 MBytes</td>
<td>475 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>70.00-80.00 sec</td>
<td>804 MBytes</td>
<td>674 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>80.00-90.00 sec</td>
<td>1.06 GBytes</td>
<td>909 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>90.00-100.00 sec</td>
<td>1.10 GBytes</td>
<td>942 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>100.00-110.00 sec</td>
<td>1.10 GBytes</td>
<td>941 Mbits/sec</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>110.00-120.00 sec</td>
<td>1.10 GBytes</td>
<td>942 Mbits/sec</td>
<td></td>
</tr>
</tbody>
</table>

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

<table>
<thead>
<tr>
<th>ID</th>
<th>Interval</th>
<th>Transfer</th>
<th>Bandwidth</th>
<th>Retr</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.00-120.00 sec</td>
<td>11.8 GBytes</td>
<td>846 Mbits/sec</td>
<td>38185 sender</td>
</tr>
<tr>
<td>4</td>
<td>0.00-120.00 sec</td>
<td>10.7 GBytes</td>
<td>763 Mbits/sec</td>
<td>receiver</td>
</tr>
</tbody>
</table>
We need to provide our HPC centres and researchers with a friction free data transfer system

- Easy to use
- Secure using a Federated Access system

The network and tools should have the data in the right location at the right time

Able to effectively use different storage tiers

- SSD
- Spinning Disk
- Tape

The researcher creates a Data Intent definition

- Data Source
- Data Target
  - Transfer priority (High, Medium, Low)
  - Storage performance (SSD, Disk, Tape)
- optional Network intersection
• PowerEdge R740

2 * Skylake - Intel(R) Xeon(R) Gold 5122 CPU @ 3.60GHz (fam: 06, model: 55, stepping: 04)
4 * Intel 10G Network Adapter
1 * Mellanox Connect-x5 (Single port) - 100G Ethernet
1 * Mellanox Connect-x5 (Dual port) - EDR InfiniBand
4Tb SSD
30Tb RAID5 HDD
Existing Data Transfer Patterns

- Sentinel
  - Many http gets
- CMIP5
  - Globus - gridftp
    - Parallel streams
- InfiniCloud - DSYNC, BeeGFS
  - Continual 9.98G UDP flow

![Graph showing data transfer times]

- Rsync A*Star to NCI R&E path: 12h 33m
- Rsync A*Star to NCI 10Gbs IPoIB: 9h
- DSYNC 1Gbs IB path: 4h 5m
- DSYNC 10Gbs IB path (~26,000Km): 24m

1,143Gb data set
Looking forward

• Middleware layers - eXtreme Data Cloud
Conclusion

• Large data transfers need to be considered as more than a web based system
  • At the Design phase not as an afterthought
    • Better protocols
    • Easier troubleshooting
• Our NREN collaborations allows us to overcome many of these problems and I would like to thank my colleagues from ICM, Geant and DFN for their kind assistance
• For more information or remote performance testing please contact me

    andrew.howard@anu.edu.au
Acknowledgements
InfiniCortex 2.0: Convergence and integration of global scale research networking; big data generation, flow, storage and processing

Marek Michalewicz, Jaroslaw Skomial
ICM, University of Warsaw

TNC18, June 2018, Trondheim
Overview

- About ICM
- InfiniCortex & ICM
- DTN & ICM
- Inter-Data Center 1,2Tb/s testbed
- Future work
ICM, University of Warsaw

Interdisciplinary Centre for Mathematical and Computational Modelling

People at ICM:
• researchers,
• software developers,
• IT administration,
• HPC experts

Hardware:
• 5 HPC systems (inc. Cray XC40), 2.1 PetaFlops peak performance
• dedicated cluster for DataScience
ICM, University of Warsaw

- ICM operates two Data Centers in Warsaw with variety of HPC systems
- Research and services: visualization, PCJ (Parallel Computing in Java), Virtual Library of Science, HPC resources for science
- Meteorology (www.meteo.pl) - 150 000 000 visits per year
- Research works in HPC/networking area:
  - high bandwidth connectivity (>1Tb/s)
  - Long distance data transfers
  - Data Transfer Nodes
  - Remote HPC systems integration
• 5 HPC Centers and MAN networks
• Multiple international Connections
• Operated by Poznan Supercomputing and Networking Center
• Over 7500 km of fiber cables
• Owned infrastructure
InfiniCortex

- Initiated in 2014 and run subsequently in 2015 and 2016
- initiated in Singapore by A*STAR Computational Research Centre under direction of Dr Marek Michalewicz
- purpose: create concurrent Galaxy of Supercomputers connected across the globe with RDMA
- long range InfiniBand using Obsidian Strategics IB range extenders
InfiniCloud 2015

True HPC Cloud around the Globe

InfiniCortex

- closed the ring around the Globe with fully InfiniBand connectivity
- sub-nets with NEW IB routers (CrossBow from Obsidian Strategics)
- InfiniCloud - created by Jakub Chrzęszczyk & Andrew Howard
Remote Fusion Experiment Data Analysis Through Wide-Area Network

Abstract:
We demonstrate remote data processing capability of large and high-throughput science experiment through cross-Pacific wide area networks and show how we can manage science workflow executions remotely by using ORNL ADIOS data management system and FNAL mtdmFTP data transfer system.

In this demo, we show a fusion data processing workflow, called Gas Puff Imaging (GPI) analysis, to detect and trace blob movements during fusion experiment. We send GPI data streams from Singapore to Fermilab for near-real time analysis. ADIOS manages analysis workflows and mtdmFTP transports stream data.

InfiniCloud

- From 2014 6-8 applications shown each year with various partners at SuperComputing
- ICM joined InfiniCortex and participated SC demos since 2015
- Demo at SC16, Salt Lake City
- Work BY C.S. Chang and others
• DTN nodes operating at various bandwidth (10-100Gb/s)

• 100Gb/s Data Transfer Node demo @SC17

• Collaboration on transfer performance
• 100Gb/s link from ICM to SC17 site
• 10Gb/s link from ICM to Singapore
• Successful cooperation of multiple NRENs
• Applications:
  • 1. Remote Medical Image Processing
  • 2. ESA Copernicus Satellite data transfer
• 1.2 Tb/s (12 * 100Gb/s)
• Pair of Infinera CX2 devices
• Only two fibers
• 20 km
• Most recent Photonic Integrated Circuit (Infinera)
• Only 1U size
• 12 servers with 100Gb/s interface
• Integrated with ICM 100Gb/s infrastructure
• 12 servers with 100Gb/s interface
Future work

- Evolution of DTN infrastructure at ICM:
  - More 100G nodes connected to storage network
  - Collaborative work on transfer performance
- Implementations
- Asia Pacific Research Platform – European Node in Poland
- HPC Centers in Poland - PIONIER-based DTN network
**DTN**
- Poznan Supercomputing and Networking Center
- A*CRC, Singapore
- NCI, Australia
- Geant
- Northwestern University, Starlight

**1,2 Tbps HPC DC Interconnect**
- Infinera
- Mellanox

**ICM Team**
- Marek Michalewicz - PI
- Robert Paciorek
- Marcin Semeniuk
- Sebastian Tymkow
- Mirosław Nazaruk
- Maciej Szpindler
- And others
Thank you