



NCI
AUSTRALIA

Big data to the other side of the World

Global Big Data and Compute at 100G

- NCI
- Copernicus
- Data transfer challenges
- Investigations with ICM
- Conclusion

- 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



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

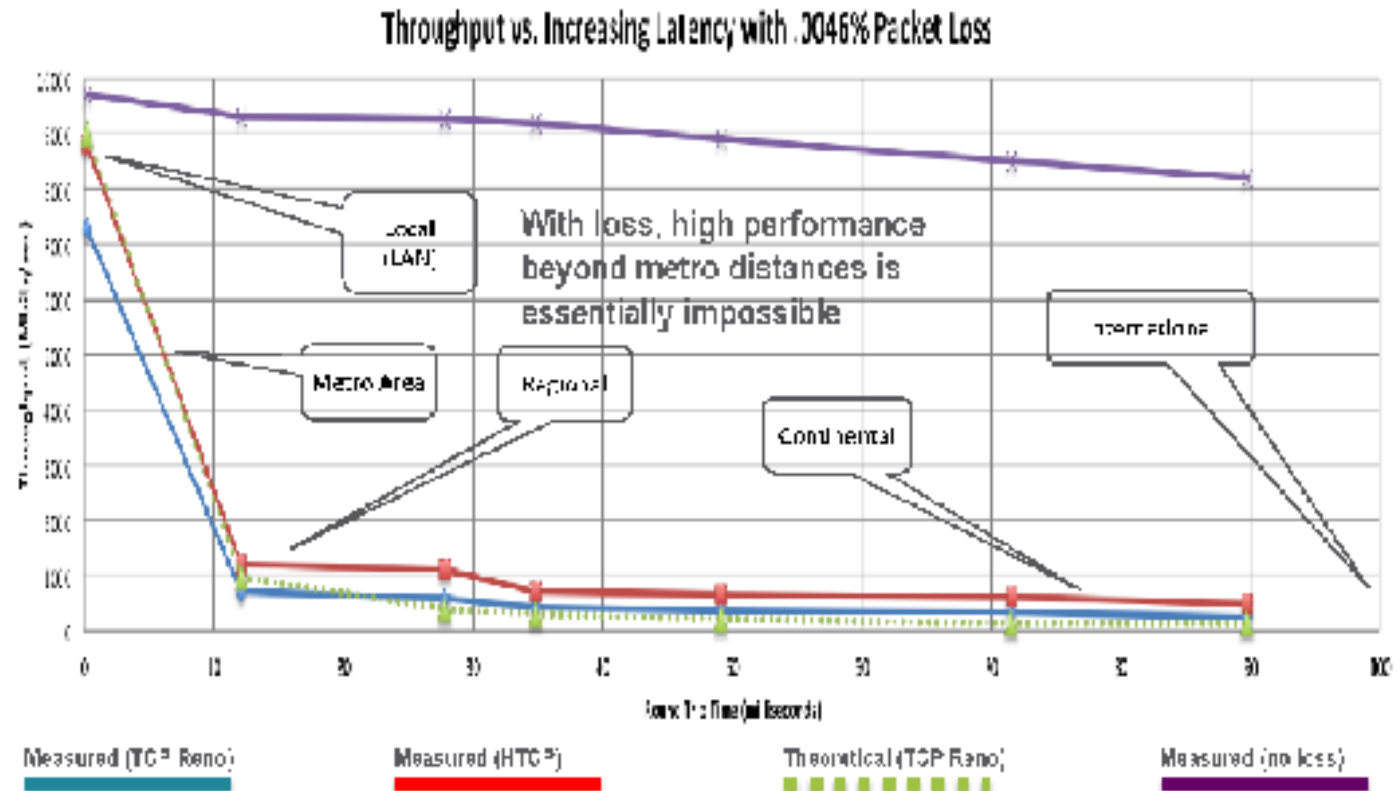
- “The fastest filesystems in the Southern Hemisphere”
 - Disk
 - Lustre File systems
 - 140GB/s sustained I/O throughput
 - 37.1Pb over multiple filesystems
 - 56G FDR InfiniBand interconnect
 - Tape
 - Dual site..... moving to two + cold archive

- 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

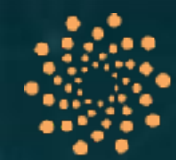
By default TCP/IP does not perform well over high bandwidth, high delay circuits.

A small amount of packet loss makes a huge difference in TCP performance

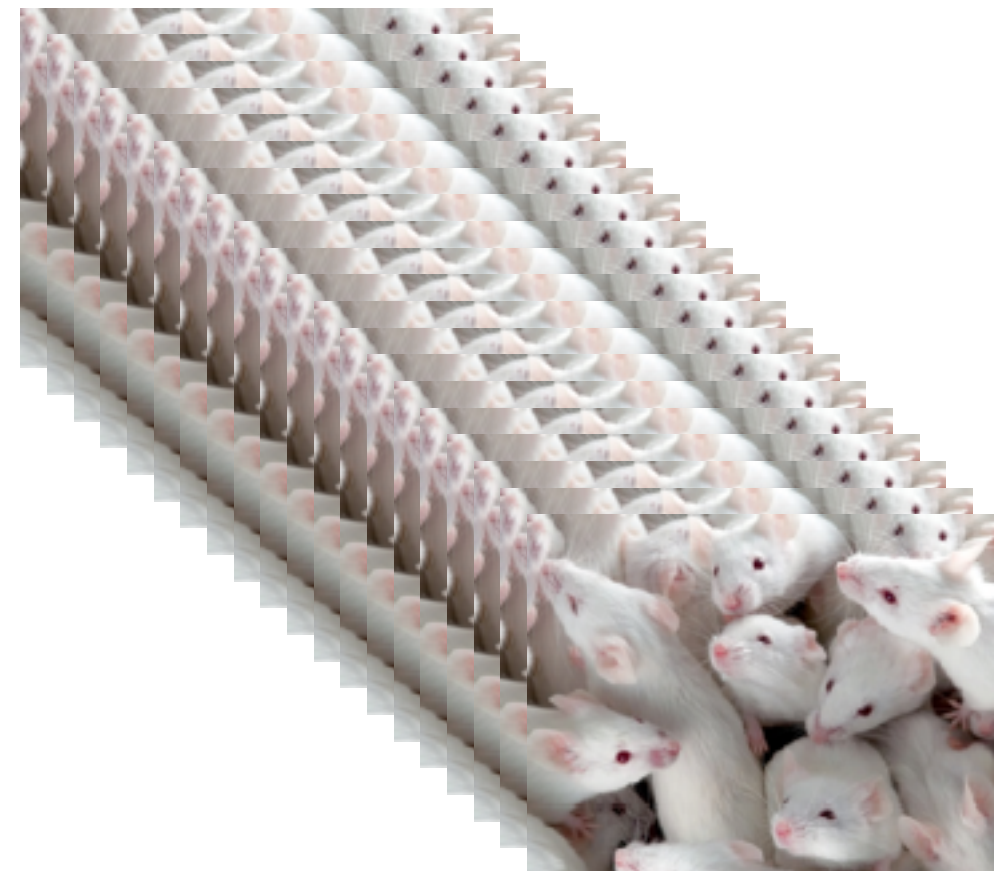
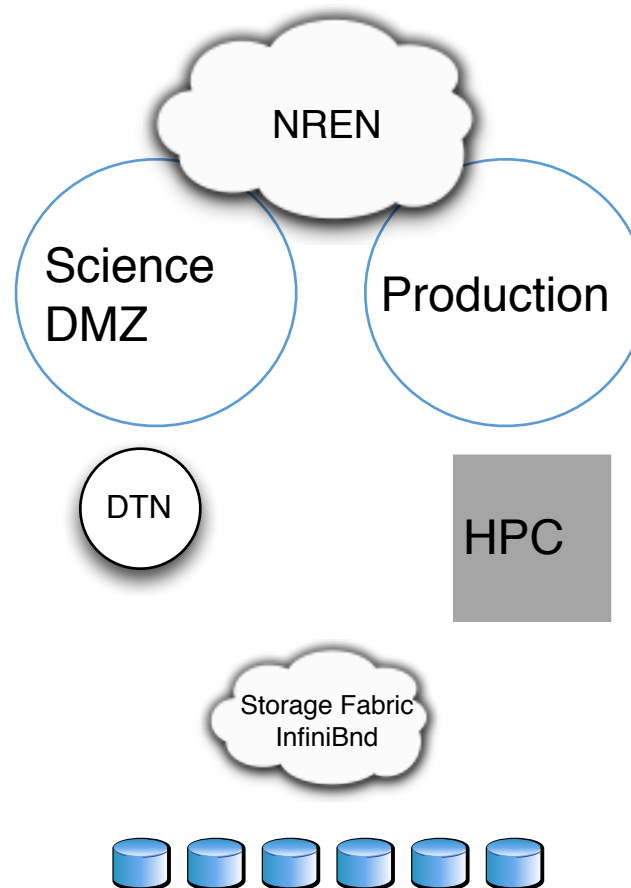


Source: <http://www.nci.org.au>

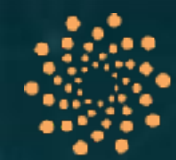
© 2001, Lixia Shen, NCI



- 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





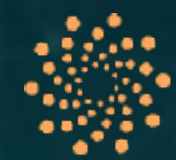


- Satellites Pairs

- **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

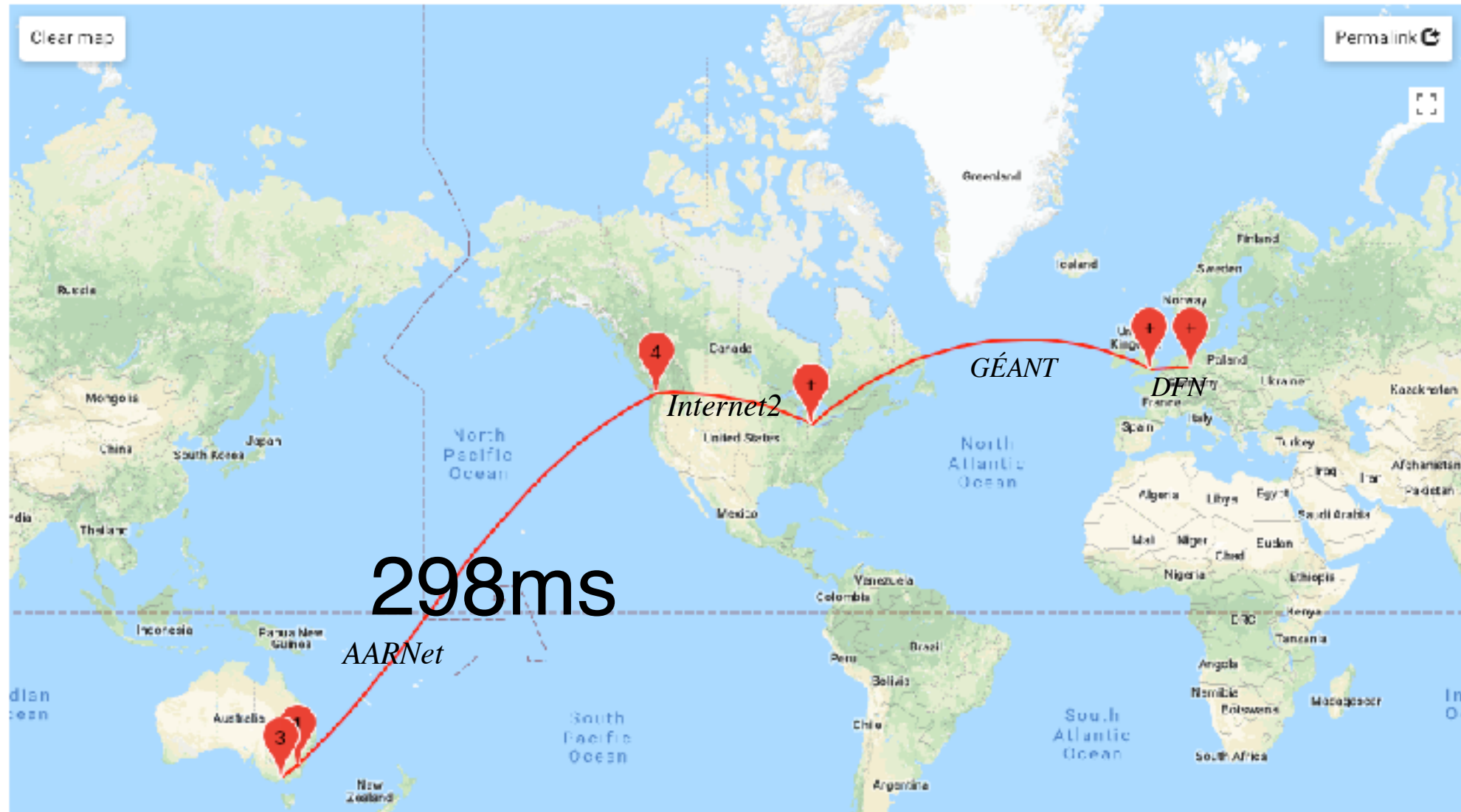






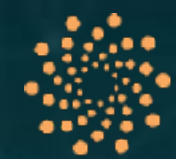
- Data is delivered as a set of small and large files
 - Thumbnails
 - ZIP files containing satellite images

NCI Network Path NCI to SCIHUB

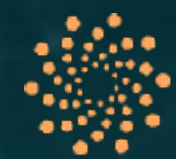


NCI Network Path NCI to ICM Poland





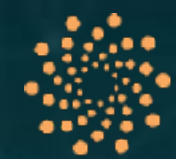
- Accelerating http using Squid
- Regional delay products and source tuning limits effective download rate
 - ICM Poland - Authenticated squid
 - Direct cache (AARNet4 10G)
 - Client at NCI requests download, squid at ICM manages first party transfer from ESA and pipelines the flow to NCI
 - 2x - 3x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



- 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

- 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 lstatlavg speed | %lpath/URI
- =====+=====+=====+====+=====+
- =====
- e215a9|OK | 3.7MiB/sl100|scihub_S3//
S3A_SL_2_LST_____20180531T201936_20180531T202236_2018
0531T212328_0179_031_385_0180_SVL_O_NR_003.zip
- Status Legend:

- NCI HTCP, ICM CUBIC

```
[ 4] local 203.0.19.11 port 35926 connected to 213.135.59.178 port 5201
[ ID] Interval          Transfer      Bandwidth      Retr  Cwnd
[ 4]  0.00-10.00  sec    6.81 MBytes   5.71 Mbits/sec    0   696 KBytes
[ 4] 10.00-20.00  sec   89.8 MBytes  75.4 Mbits/sec    0   7.41 MBytes
[ 4] 20.00-30.00  sec    221 MBytes   186 Mbits/sec  2109   4.64 MBytes
[ 4] 30.00-40.00  sec    188 MBytes   157 Mbits/sec    0   9.47 MBytes
[ 4] 40.00-50.00  sec    400 MBytes   336 Mbits/sec    0  19.6 MBytes
[ 4] 50.00-60.00  sec    778 MBytes   652 Mbits/sec    0  37.1 MBytes
[ 4] 60.00-70.00  sec    258 MBytes   216 Mbits/sec  11728   19.8 KBytes
[ 4] 70.00-80.00  sec    5.00 MBytes    4.19 Mbits/sec   554   540 KBytes
[ 4] 80.00-90.00  sec   63.8 MBytes   53.5 Mbits/sec    0   6.05 MBytes
[ 4] 90.00-100.00 sec    339 MBytes   284 Mbits/sec    0  17.4 MBytes
[ 4] 100.00-110.00 sec    695 MBytes   583 Mbits/sec    0  33.4 MBytes
[ 4] 110.00-120.00 sec    694 MBytes   582 Mbits/sec  13220  16.8 MBytes
- - - - -
[ ID] Interval          Transfer      Bandwidth      Retr
[ 4]  0.00-120.00  sec    3.65 GBytes   261 Mbits/sec  27611
[ 4]  0.00-120.00  sec    3.62 GBytes   259 Mbits/sec
```

- NCI HTCP, ICM HTCP

[4] local 203.0.19.11 port 36326 connected to 213.135.59.178 port 5201

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

[ID]	Interval		Transfer	Bandwidth	Retr	
[4]	0.00-120.00	sec	5.06 GBytes	362 Mbits/sec	22075	sender
[4]	0.00-120.00	sec	5.03 GBytes	360 Mbits/sec		receiver

[ID]	Interval		Transfer	Bandwidth		
[4]	0.00-10.00	sec	702 MBytes	589 Mbits/sec		
[4]	10.00-20.00	sec	1.10 GBytes	942 Mbits/sec		
[4]	20.00-30.00	sec	1.10 GBytes	942 Mbits/sec		
[4]	30.00-40.00	sec	1.10 GBytes	942 Mbits/sec		
[4]	40.00-50.00	sec	528 MBytes	443 Mbits/sec		
[4]	50.00-60.00	sec	461 MBytes	387 Mbits/sec		
[4]	60.00-70.00	sec	567 MBytes	475 Mbits/sec		
[4]	70.00-80.00	sec	804 MBytes	674 Mbits/sec		
[4]	80.00-90.00	sec	1.06 GBytes	909 Mbits/sec		
[4]	90.00-100.00	sec	1.10 GBytes	942 Mbits/sec		
[4]	100.00-110.00	sec	1.10 GBytes	941 Mbits/sec		
[4]	110.00-120.00	sec	1.10 GBytes	942 Mbits/sec		
- - - - -						
[ID]	Interval		Transfer	Bandwidth	Retr	
[4]	0.00-120.00	sec	11.8 GBytes	846 Mbits/sec	38185	sender
[4]	0.00-120.00	sec	10.7 GBytes	763 Mbits/sec		receiver

- 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)

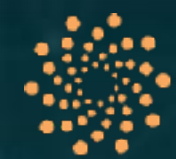
4 * Intel 10G Network Adapter

1 * Mellanox Connect-x5 (Single port) - 100G Ethernet

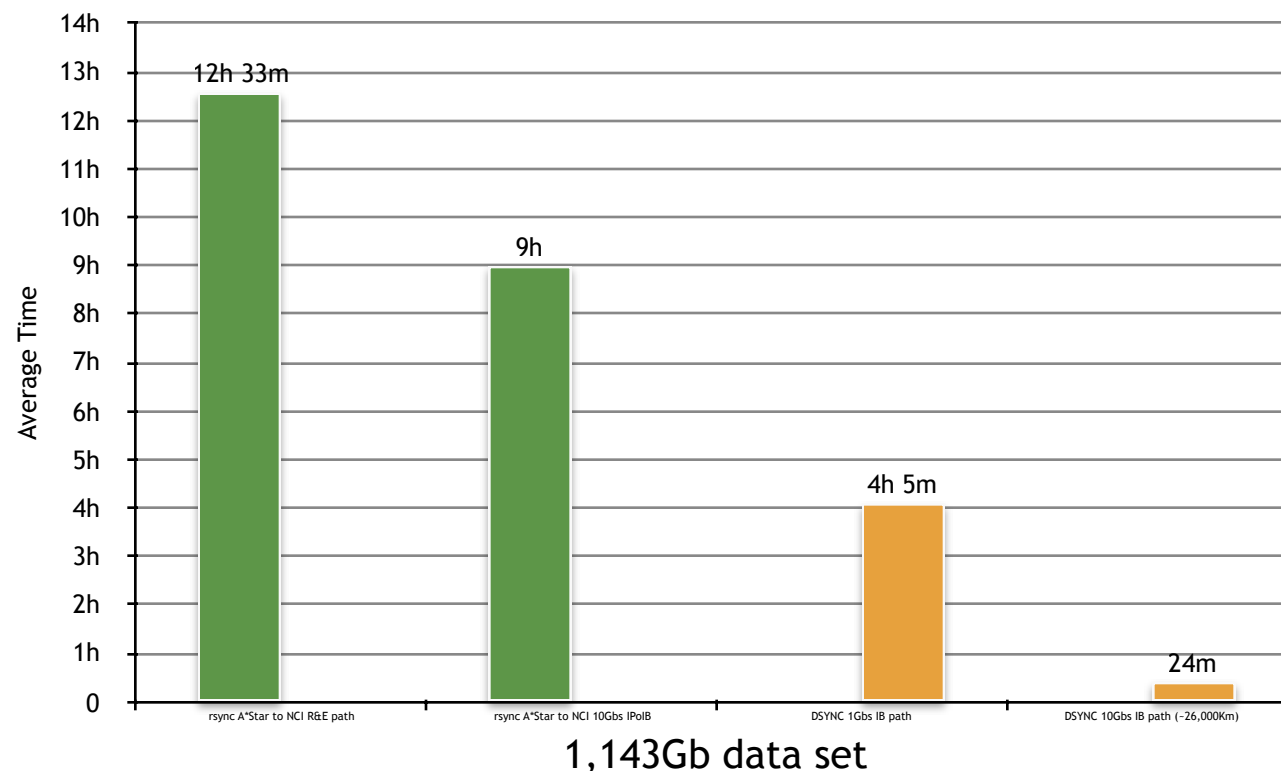
1 * Mellanox Connect-x5 (Dual port) - EDR InfiniBand

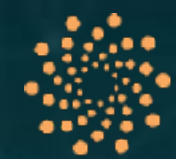
4Tb SSD

30Tb RAID5 HDD

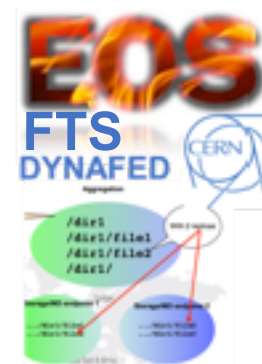


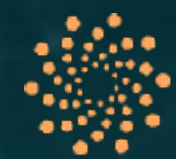
- Sentinel
 - Many http gets
- CMIP5
 - Globus - gridftp
 - Parallel streams
- InfiniCloud - DSYNC, BeeGFS
 - Continual 9.98G UDP flow



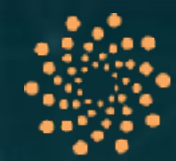


- Middleware layers - eXtreme Data Cloud





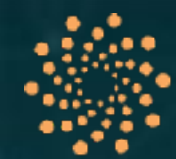
- Large data transfers need to be considered as more than a web based system
 - At the Design phase not as an afterthought
 - Better protocols
 - Easier troubleshooting
- Our NREN collaborations allows us to overcome many of these problems and I would like to thank my colleagues from ICM, Geant and DFN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



- For more information or remote performance testing please contact me

andrew.howard@anu.edu.au





NCI
AUSTRALIA

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

- 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:

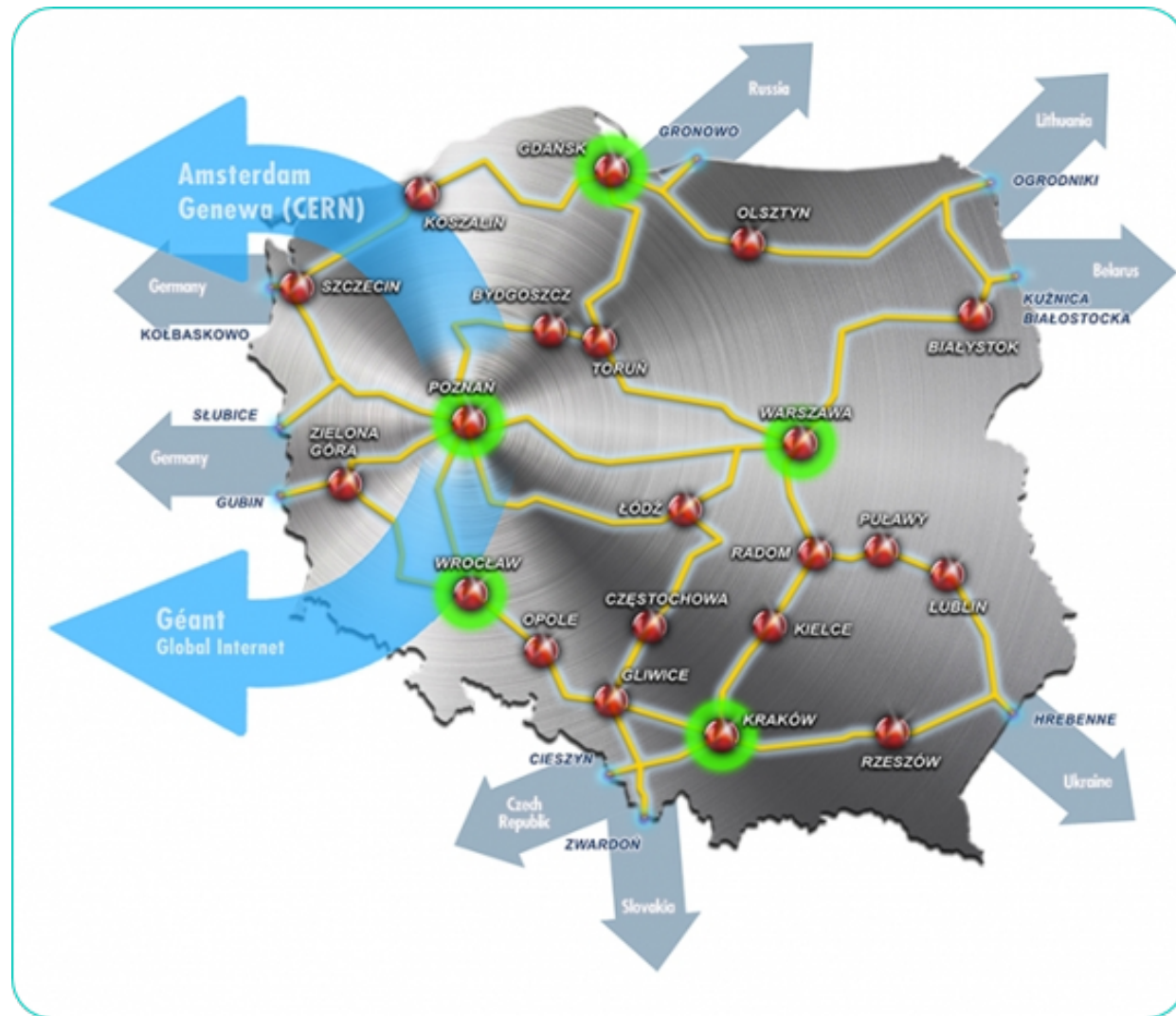
- 5 HPC systems (inc. Cray XC40), 2.1 PetaFlops peak performance
- dedicated cluster for DataScience

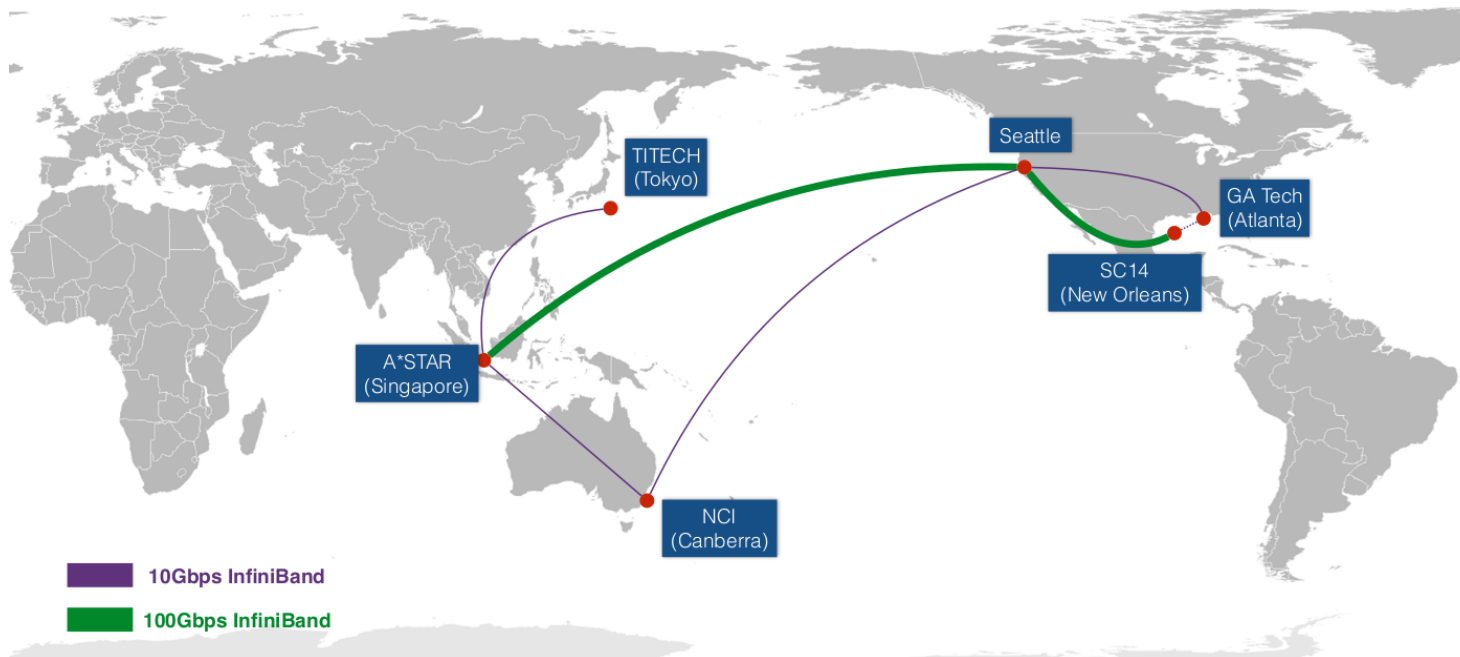


- ICM operates two Data Centers in Warsaw with variety of HPC systems
- Research and services: visualization, PCJ (Parallel Computing in Java), Virtual Library of Science, HPC resources for science
- meteorology (www.meteo.pl) - 150 000 000 visits per year
- Research works in HPC/networking area:
 - high bandwidth connectivity ($>1\text{Tb/s}$)
 - 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



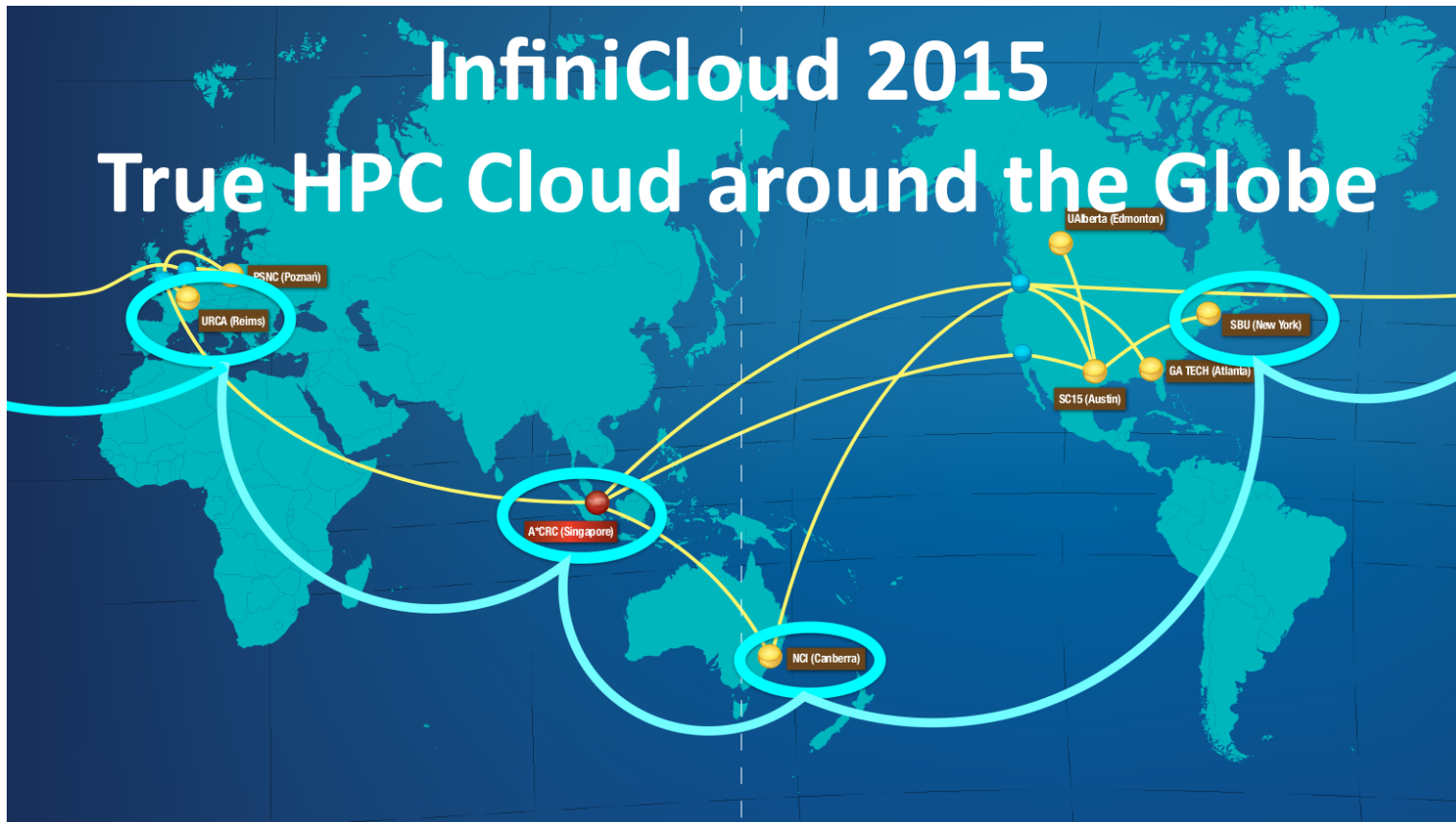


InfiniCortex

- 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 Strategies IB range extenders

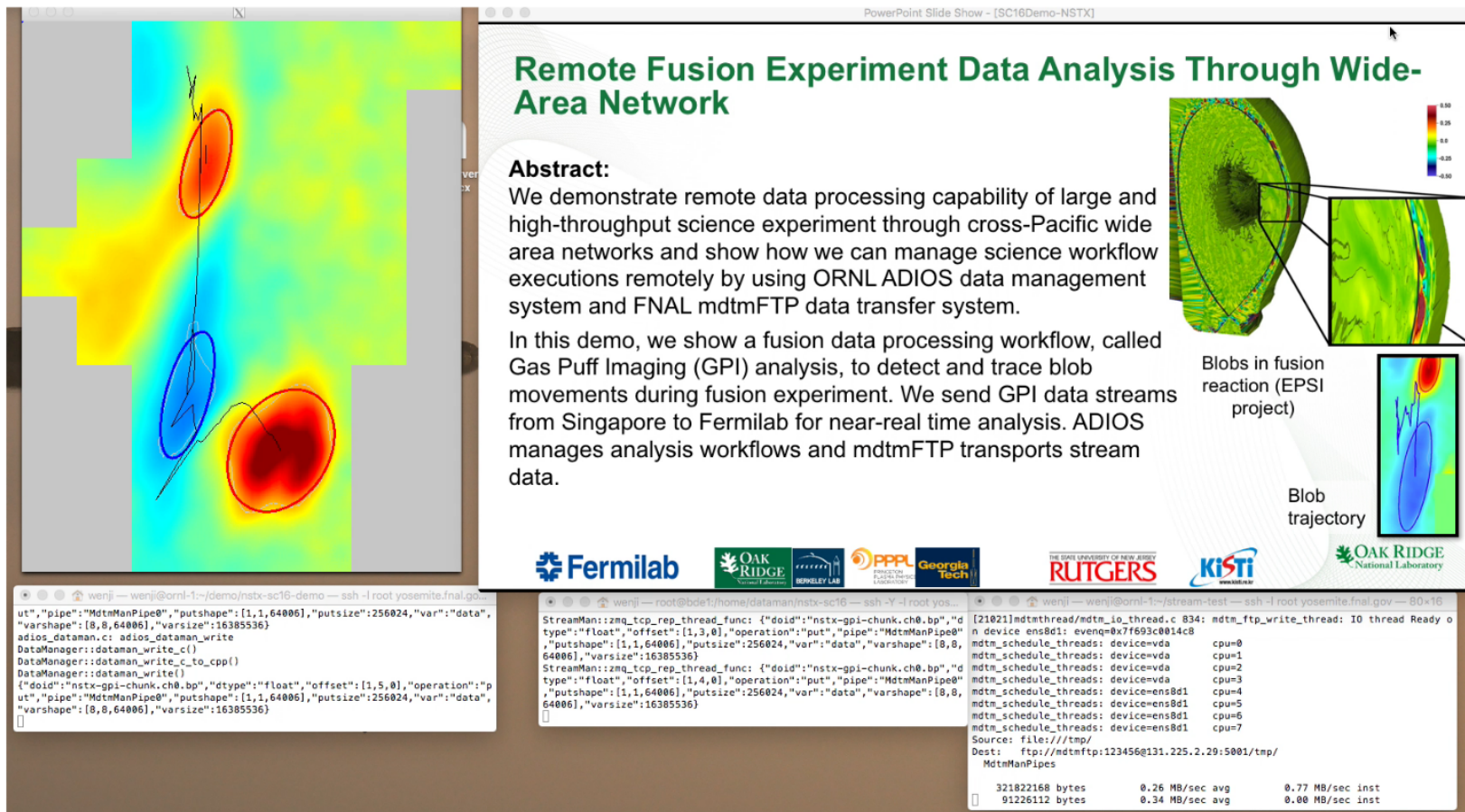
InfiniCloud 2015

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 Strategies)
- InfiniCloud - created by Jakub Chrzęszczczyk & 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 mdtmFTP 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 mdtmFTP transports stream data.

The slide includes three visualizations: a large heatmap on the left showing blob movements with red and blue regions; a 3D visualization of a fusion reaction (EPSI project) in the top right; and a 2D visualization of a blob trajectory in the bottom right. A color scale on the right ranges from -0.50 to 0.50.

Logos at the bottom include Fermilab, Oak Ridge National Laboratory, PPPL, Georgia Tech, Rutgers, KISTI, and Oak Ridge National Laboratory.

Terminal windows at the bottom show command-line interactions and data transfer statistics:


```

[21021]mdtmthread/mdtm_io_thread.c 834: mdtm_ftp_write_thread: IO thread Ready o
n device ens8d1: evenq=0x7f693c0014c8
mdtm_schedule_threads: device=vda      cpu=0
mdtm_schedule_threads: device=vda      cpu=1
mdtm_schedule_threads: device=vda      cpu=2
mdtm_schedule_threads: device=vda      cpu=3
mdtm_schedule_threads: device=ens8d1    cpu=4
mdtm_schedule_threads: device=ens8d1    cpu=5
mdtm_schedule_threads: device=ens8d1    cpu=6
mdtm_schedule_threads: device=ens8d1    cpu=7
Source: file:///tmp/
Dest:  ftp://mdtmftp:123456@131.225.2.29:5001/tmp/
MdtmManPipes
321822168 bytes      0.26 MB/sec avg      0.77 MB/sec inst
91226112 bytes       0.34 MB/sec avg      0.00 MB/sec inst
  
```

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



DTN INFRASTRUCTURE @ ICM

DTN node specification:

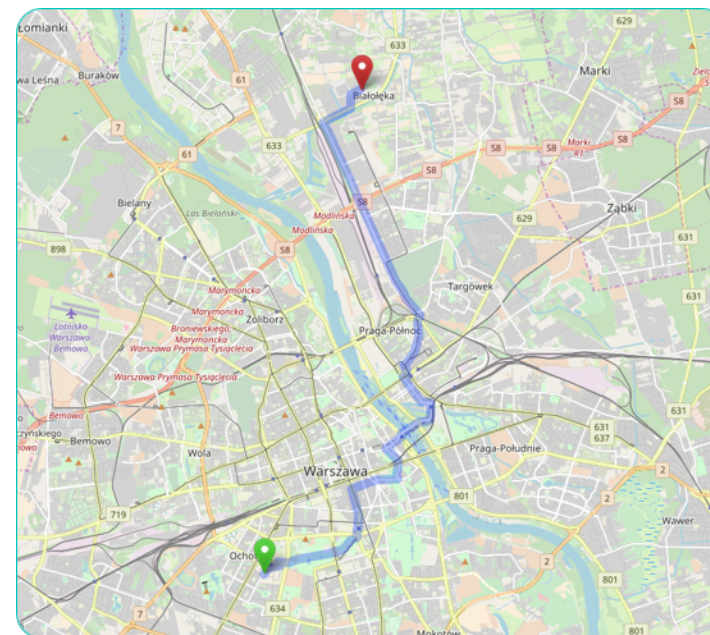
- 2x Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz
- 1x PCIe x16 for 100Gbit Ethernet with Mellanox ConnectX-4 (MT27700) network transfer ~98Gbit/s with MTU=9000
- 2x PCIe x8 for storage InfiniBand network with Mellanox ConnectX-4 (MT27700)

Storage specification:

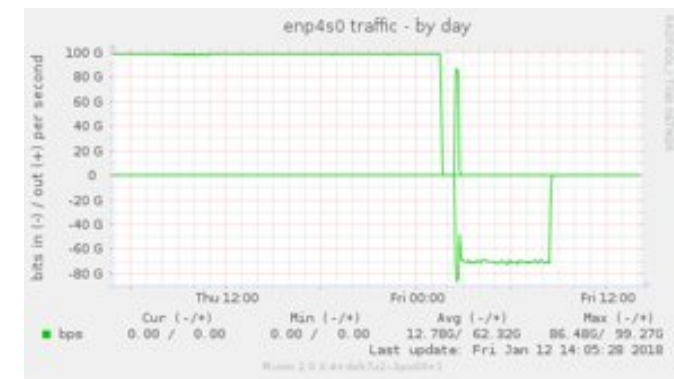
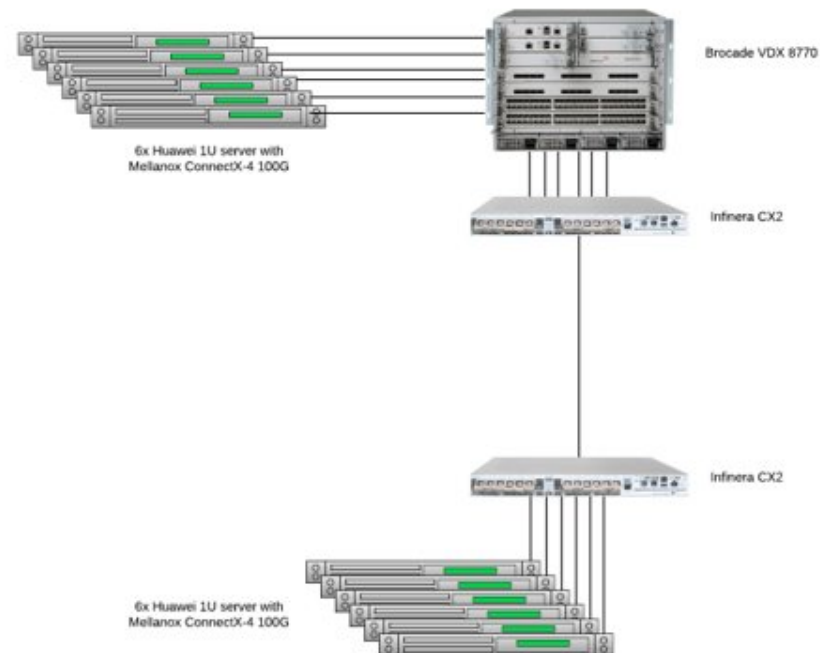
- Lustre 2.10.1
- 32 nodes:
 - 4 HDD in RAID5
 - single port InfiniBand card Mellanox ConnectX-3 (MT27500)
- 560TB fast storage space

- ### High Speed Research Network
-
- The map illustrates the High Speed Research Network, showing connections between various research centers and networks. The legend indicates that red lines represent 100Gbps connections and yellow lines represent 10Gbps connections.
- 100Gbps Connections (Red Lines):**
- GEANT/PSN/RIKEN to ICM
 - GEANT/TEIN-CC to SOE
 - SOE to Singapore
 - Singapore to A*STAR and NSCC
 - Singapore to Hong Kong
 - Hong Kong to Japan
 - Japan to SINET
 - SINET to TransPAC / PACIFIC WAVE
 - TransPAC / PACIFIC WAVE to GRPnet/StarLight/MREN
 - GRPnet/StarLight/MREN to ICAIR
 - ICAIR to ANA 300
 - ANA 300 to SC17 (Denver)
 - SC17 (Denver) to SCinet
 - SCinet to 100Gbps co-shared with INTERNET2
 - 100Gbps co-shared with INTERNET2 to AARNet-X
 - AARNet-X to NCI
 - NCI to 100Gbps co-shared with NICT
 - 100Gbps co-shared with NICT to Japan
- 10Gbps Connections (Yellow Lines):**
- ICM to GEANT/TEIN-CC
- Research Centers and Networks:**
- GEANT/PSN/RIKEN
 - ICM
 - GEANT/TEIN-CC
 - SOE
 - Singapore
 - A*STAR
 - NSCC
 - Hong Kong
 - Japan
 - SINET
 - TransPAC / PACIFIC WAVE
 - GRPnet/StarLight/MREN
 - ICAIR
 - ANA 300
 - SC17 (Denver)
 - SCinet
 - 100Gbps co-shared with INTERNET2
 - AARNet-X
 - NCI
 - 100Gbps co-shared with NICT

- 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



- 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 Tb/s HPC DC Interconnect

Infinera

Mellanox

ICM Team

Marek Michalewicz - PI

Robert Paciorek

Marcin Semeniuk

Sebastian Tymkow

Mirosław Nazaruk

Maciej Szpindler

And others

