



AENEAS WP4

Analysis of Global SKA Data Transport and Optimal European Storage Topologies



Introduction

Richard Hughes-Jones



WP4 Objectives

Produce a design and best practice recommendations for the network, data transfer and storage technologies required for an ESRC.

- Investigate and demonstrate the data transfer and storage techniques required to confirm the viability of a distributed computing and network architecture for a European Science Data Centre.
- Collaborate with South Africa and Australia to address the challenges of moving large data volumes. PoC trials between NRENs & Radio Astronomy end sites. Support the work of WP3.
- Moving up the stack, Study of: data access protocols, data transfer protocols, replica and transfer management, data moving applications. Support for WP3.
- Develop an architecture and cost models for European & Global connectivity for ESRC.
- Looking forward: Strengthen collaboration between SKA community & infrastructure providers formation of SKA-NREN Forum

WP4 Tasks

- Task 4.1: Evaluation of existing data transfer protocols, storage sub-systems and applications
Partners: Chalmers (lead), GÉANT Ltd, Jülich, UMAN Stakeholders: CSIRO, SANReN
- Task 4.2: Inventory of the storage and network capabilities of existing and planned European Facilities for SKA
Partners: INAF (lead), GÉANT Ltd Stakeholders: ASTRON
- Task 4.3: Optimized design and cost model for a distributed ESRC data topology with world connectivity
Partners: GÉANT Ltd (lead), UMAN Stakeholders: AARNet, CSIRO, SANReN
- Task 4.4: Proof of Concept Activities supporting the design of data access and transport within Europe and from the Host countries to Europe
Partners: GÉANT Ltd (lead), Chalmers, UMAN Stakeholders: AARNet, CSIRO, SANReN



WP4 Deliverables

Deliverable	Title	Due date	Status
D 4.1	Best practice recommendations Data moving applications, protocols and storage	M 14 Mar 18	Complete Submitted and approved
D 4.2	Site Catalogue storage and networking	M 18 Aug 18	Indicative deliverable Submitted and approved
D 4.2	Site Catalogue storage and networking		Updated version submitted
D 4.3	Architecture and cost model for European ESDN network	M 20 Jul 19	Complete Submitted and approved
D 4.4	Architecture and cost model for World-wide network for SKA	M 32 Sep 19	Complete Ready for Submission
D 4.5	Data Transport Tests and Recommendations	M 34 Nov 19	Long-haul tests continue Draft outline done

WP4 Milestones



Milestone number	Milestone name	Related WP(s)	Due date	Means of verification	Status
11	List of possible regional site locations	WP2 WP4	M 9	List of possible sites established	Done
27	Demonstration of moving data from observatory sites (SA) to ESRC	WP3 WP4	M 19 Aug18	Demonstration completed	IWP4 made tests between NRENs WP3 moved data
30	Joint Milestone (WP4) on data replica manager	WP3 WP4	M 21 Oct18	Internal memo	Document written
31	Specifications for SKA Replica Manager	WP3 WP4	M 21 Oct18	Specification document written	Document written
33	Joint Milestone (WP4) on demonstration of moving data from observatory sites (AUS) to ESRC	WP3 WP4	M 24 Jan 19	Demonstration completed	Tests & demo made
35	Data transfer test Australian site to European site	WP4	M 27 Apr19	Technical note written	Some tests made Reported at CHEP2019
36	Report on Data Transport ESRC within Europe	WP3 WP4	M 28 May19	Technical note written	??
37	radio astronomy data over global routes from Australia to Europe	WP3 WP4	M 30 Jul19	WP3 Technical note written	Tests have been made
40	Joint Milestone (WP4) on demonstration of moving data within ESRC	WP3 WP4	M 31 Aug19	Demonstration completed	Tests & demo made Reported at CHEP2019



Task 4.1

D4.1 "Best practice recommendations
Data moving applications, Protocols
and Storage"

Simon Casey and Jimmy Cullen



Deliverable 4.1: overview

- Reasoning
 - SKA needs to transport large amounts of data across long distances
 - Long distance networks & high bandwidths not best friends
- Methodology
 - Examine existing data moving & network testing tools
 - Investigate performance of these both locally & over wider networks
 - Consider how LOFAR manages data transport to its Long Term Archive showed need for automated File Transfer System
 - Produce recommendations for optimising network transfers



Deliverable 4.1: software tools used

- Data moving
 - GridFTP
 - Developed and formally defined within the GridFTP working group of the Open Grid Forum.
 - Multiple implementations available, e.g. Globus Toolkit
 - Provide greater reliability & higher performance than regular FTP
 - We note the issue with lack of funding for community support.
 - FDT
 - Designed for efficient data transfers
 - Uses standard TCP, written in JAVA, controlled via CLI.
 - Xrootd (used in later work)

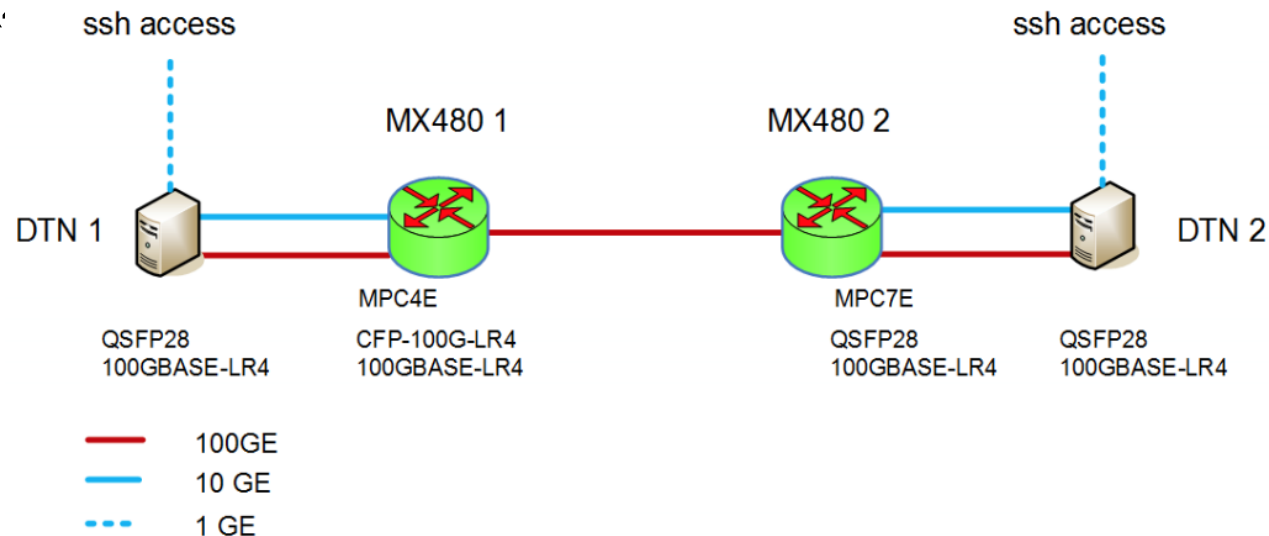


Deliverable 4.1: software tools used

- Network testing
 - UDPmon
 - Tests capabilities of end hosts and networks using UDP transfers
 - Detailed statistics of network transfers
 - Useful to pinpoint where performance issues lie
 - Iperf/iperf3
 - Can test both TCP and UDP
 - Multithreaded
 - Can be used to test several parallel streams
 - libVMA
 - Library to provide “kernel bypass” to standard TCP & UDP applications

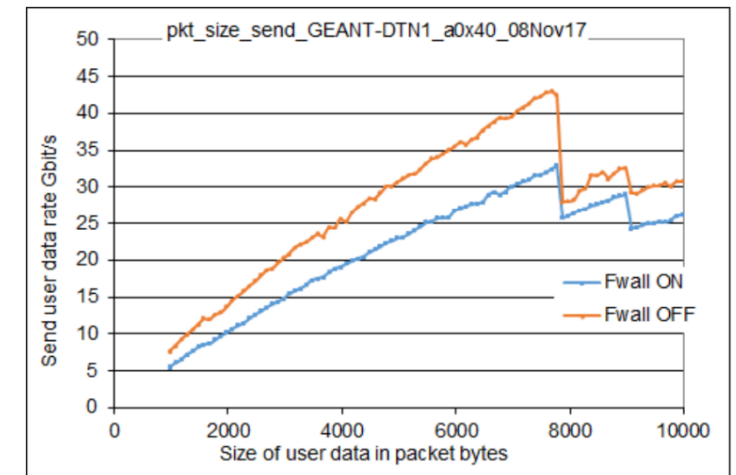
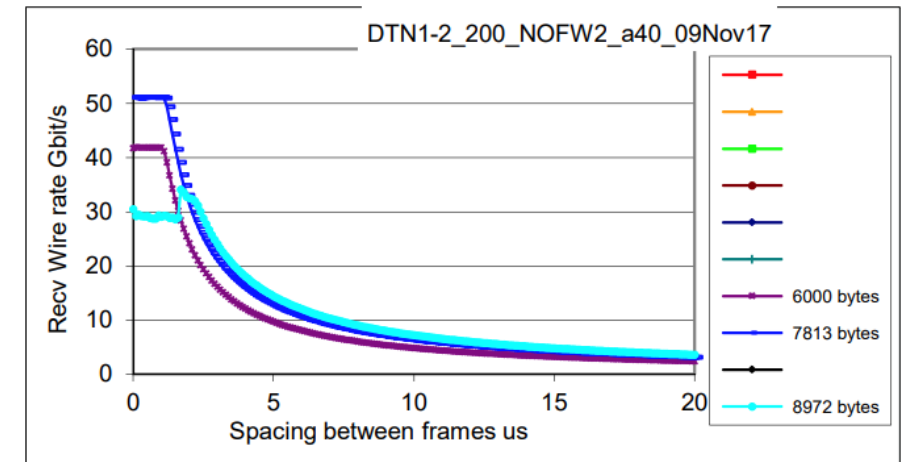
Deliverable 4.1: local network testing

- Two DTN machines in GEANT Network Lab
 - Fedora 23 Linux operating system
 - Mellanox ConnectX-3 (10G) & Mellanox ConnectX-4 / ConnectX-5 (100G)
- Each connected to a Juniper MX480 router via 10 GbE and 100 GbE
- MX480s connected via QSFP28 100GBASE-LR optical interfaces
 - MTU 9192

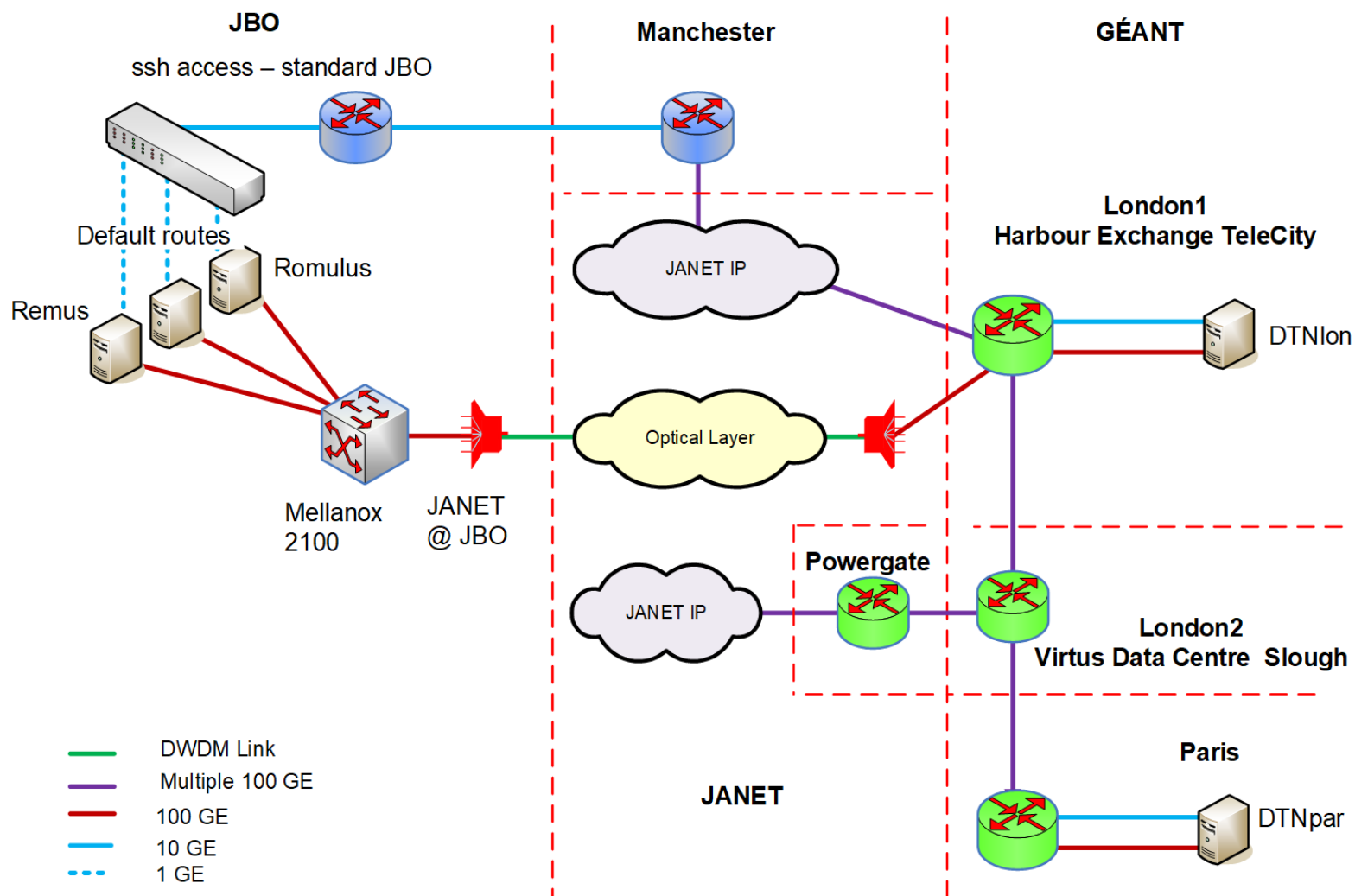


Deliverable 4.1: local network testing

- 100 Gbit UDP measurements
 - Maximum 50 Gbit/s single stream
 - 7813 byte packet size, 1.2 μ s packet spacing
 - Full jumbo frames (8972 byte packet size) give 30 Gbit/s
 - Expect bandwidth to increase with packet size
 - Possible artefact of Fedora 23 kernel
- Host PC firewall decreases maximum throughput \sim 20-30 %
- With libVMA achieved 89.8 Gbit/s sustained over 96 hours
 - 8972 byte packets, 0.8 μ s packet spacing, single stream
 - Very occasional packet loss due to receive NIC buffer overflow

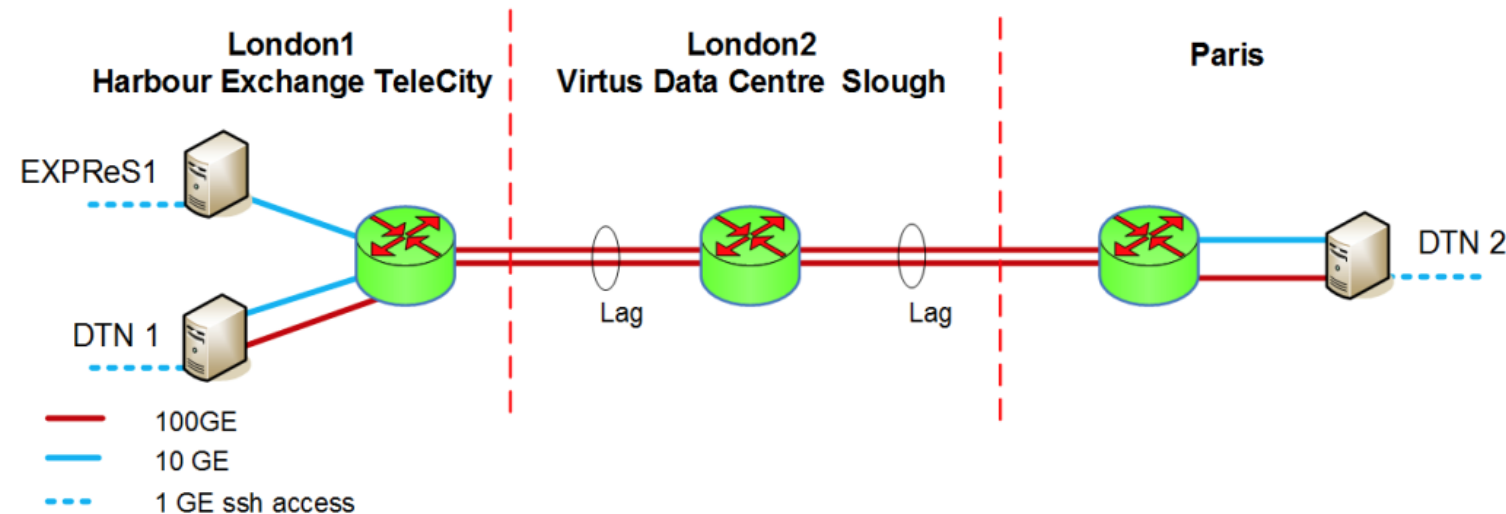


AENEAS DTNs & Network Topology Jodrell Bank to GÉANT



Deliverable 4.1: pan-European network testing

- Testing performed over GÉANT network between London & Paris POPs
 - RTT 7.5ms
- Similar DTN nodes as used for local testing
- UDP achieved stable 35 Gbit/s with firewall enabled, no packet loss
 - Same result as in local testing
 - Link uncongested



Deliverable 4.1: pan-European network testing

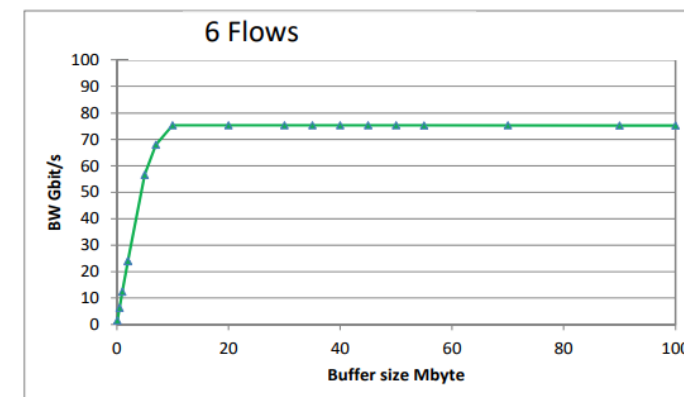
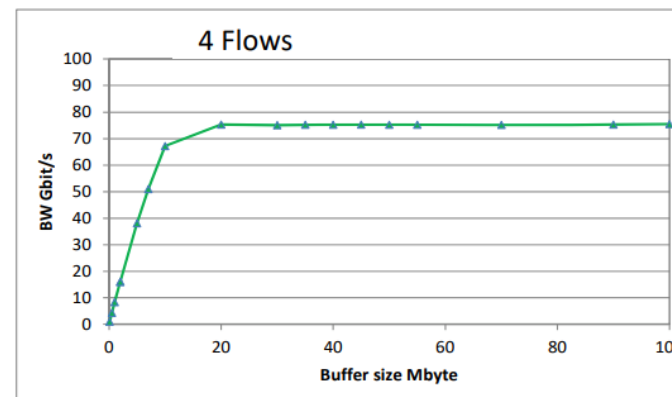
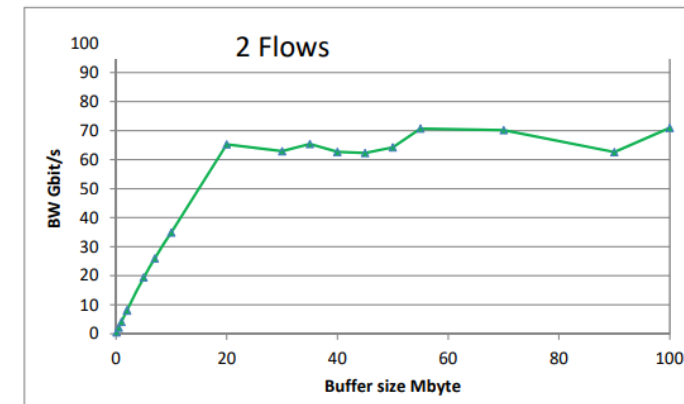
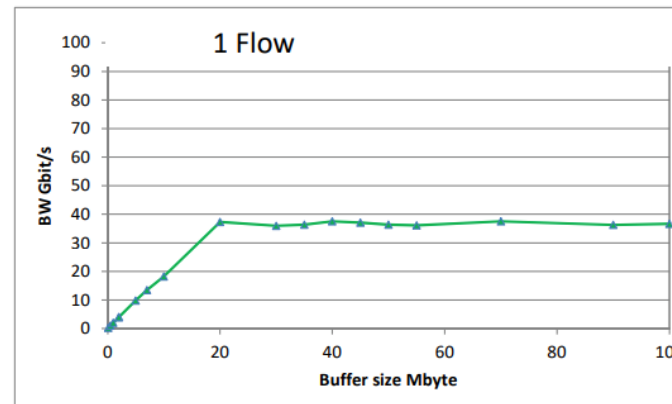
- TCP testing with iperf3 achieved 37 Gbit/s with TCP buffer of 35 MB

- Very close to the expected Bandwidth

- Delay Product of 34.7 MB for
37 Gbit/s with RTT 7.5 ms

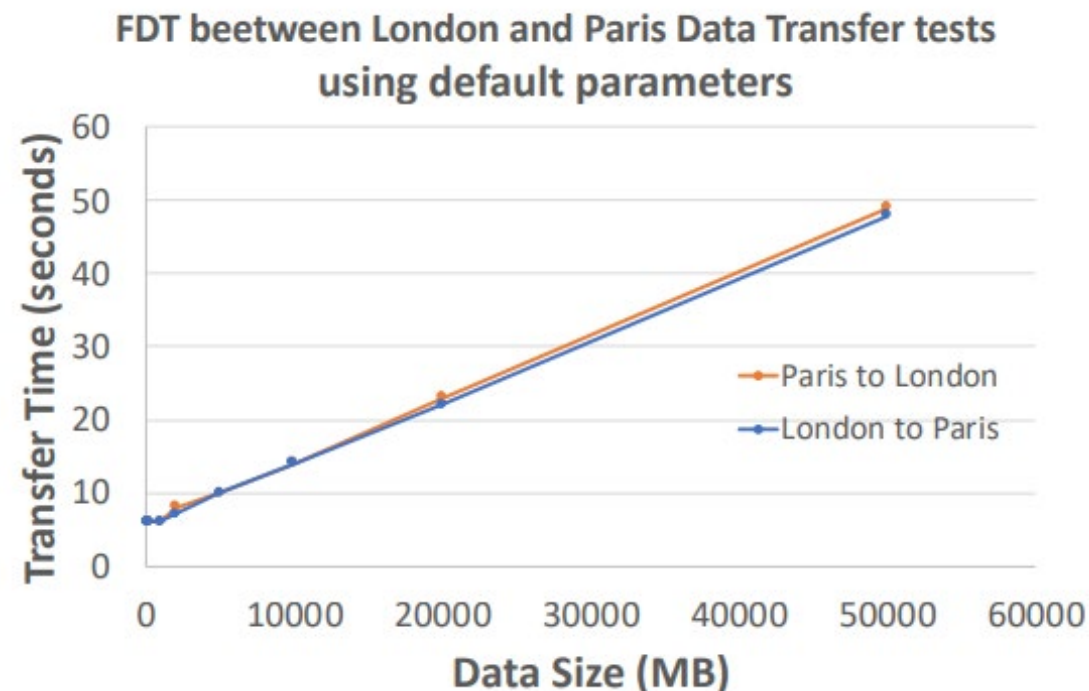
- Multiple streams achieved 75 Gbit/s
with 4 streams

- Long term stability testing showed constant
32.5 Gbit/s over 30 hours, no retransmits



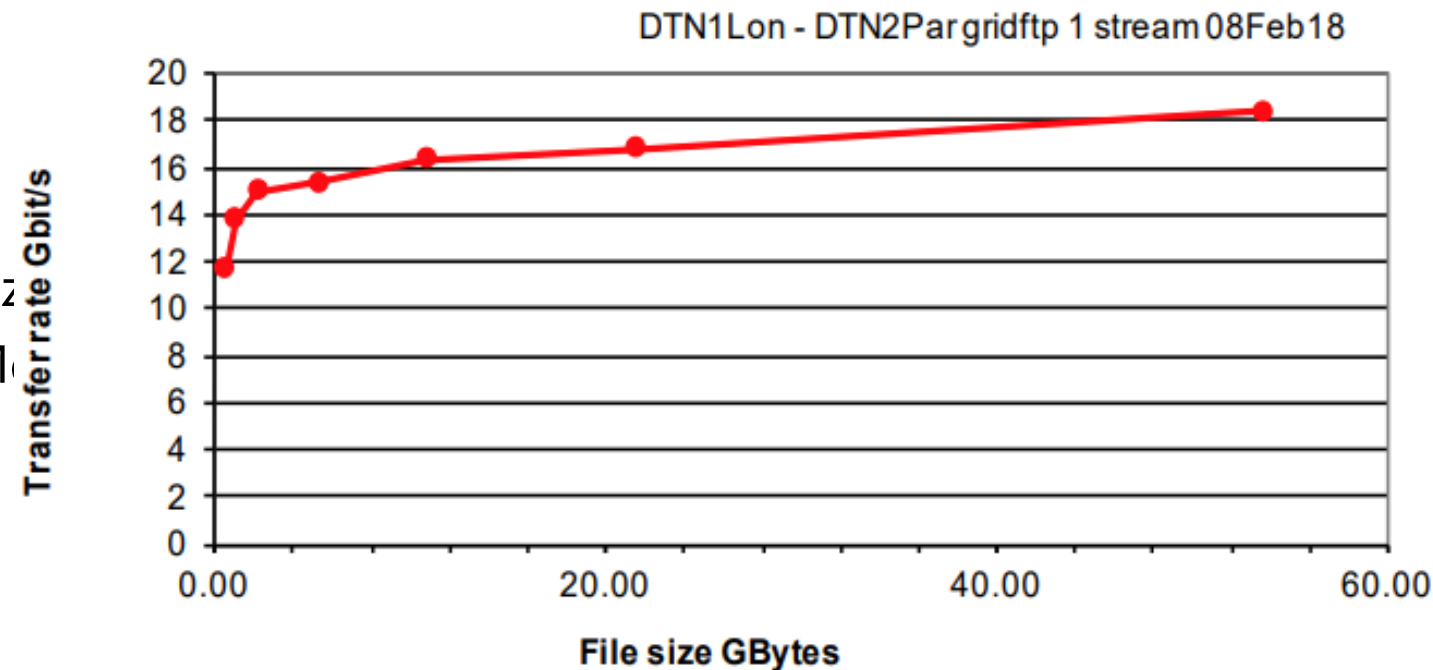
Deliverable 4.1: file transfers

- FDT 10 Gbit/s testing London <-> Paris
 - Files of sizes 100 MB -> 50 GB were created at each test node
 - Time for FDT to transfer each file size measured
 - 1 GB and below took 6s
 - overheads of starting an FDT transfer
 - Above 1 GB transfer time scaled linearly
 - 9.9 Gbit/s bandwidth achieved
 - Similar speeds in either direction
 - Useful tool for transferring larger (e.g. > 10 GB) files due to overhead of starting transfers



Deliverable 4.1: file transfers

- GridFTP 100 Gbit/s tests
 - 100 Gbit interfaces used between DTN nodes in London & Paris
 - Same file sizes as for FDT tests
 - Single TCP flow
 - Linear increase in transfer time vs file size above 5 GB file sizes
 - Maximum 19 Gbit/s @ 50 GB file size
 - Consistent with measured NVM₀ disk-memory transfer rate





Deliverable 4.1: recommendations & conclusions

- Shared hosts / virtual machines should be avoided for DTNs
 - High performance transfers esp. TCP-based can be severely affected if incoming packets are dropped due to the CPU being busy servicing another un-related process.
- Hyperthreading should be disabled
- CPU frequency scaling should be set to either 'performance' governor or alternatively set the CPU manually to the highest possible frequency
- On multi CPU systems, it is important that network card interrupts are handled by the CPU which provides the PCI express lane the NIC is connected to. IRQ balancing should be disabled.
- Increase NIC ring-buffers to largest supported value (different NICs have different max values)
- Expect SKA file sizes to be similar if not greater than LOFAR
- Automated replica system needed to cope with any failed transfers

Network testing at OSO

- Previously libVMA only accelerated UDP transfers
 - Offered performance increase 50 -> 90 Gbps
- Recently updated drivers enabled TCP acceleration on Mellanox ConnectX cards
 - No real benefit seen so far with TCP streams
 - Some cases a slight performance increase, others show a decrease
 - Possibly early drivers, better performance will come with more mature drivers?



NVMe testing at OSO (1)

- 3 different brands of NVMe drives tested (2 of each)
 - Adata XPG sx8200 480 GB
 - Samsung 970 Evo 500 GB
 - WD Black 500 GB
- Connected either directly to PCIe lanes of the CPU or to PCIe lanes provided by the chipset with 32 Gbit/s DMI bus to the CPU
- Read/write testing of individual drives connected by either method as well as several combinations of RAID-0



NVMe testing at OSO (2)

- Individual disk performance
 - WD & Samsung perform similarly
 - Both feature 'fast cache' – high performance section of the disk (ca 16 GB), write performance drops heavily if cache filled. Need to wait for cache to be flushed to main storage
 - Good for bursty writes, not good for streamed writes
 - 17 Gbit/s writing to 'fast cache'; drops to 6 Gbit/s
 - Read consistent 24 Gbit/s
 - Adata disks consistent 12 Gbit/s write, 24 Gbit/s read
- Software RAID0 testing
 - RAID 0 arrays were created using combinations of 2-4 Adata & Samsung drives
 - Max read speed achieved of ca 30 Gbps with 2 drives, no extra speed with 3 & 4 drives. Write speed max 20 Gbps with 2 Adata drives – write speed decreased when including Samsung drives.

NVMe tests at JBO

- mdadm software RAID 0, XFS filesystem, largeio mount option
- romulus: 6 x Intel DC P3700 SSD 800GB
- remus: 6 x 2 TB Samsung 960 Pro NVMe SSD Controller SM961/PM961
- 10 simultaneous dd writes (10^9 bytes):
 - 6.73 GBps (53.87 Gbps)
- 10 simultaneous dd read almost double write speed:
 - 12.39 GBps (99.12 Gbps)

Motherboard	AsusZ10PE-D8 WS (All)	7 x PCI-e sockets
Memory	DIMM DDR4 Synchronous 2400 MHz ECC	8 x 32 GB 256GB total
CPU	2 x Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz	10 physical cores per CPU
Storage	6 x 2 TB Samsung 960 Pro NVMe SSD Controller SM961/PM961	12 TB Storage
System Disk	Samsung SSD 850	256GB
Networking	2 x Gigabit Ethernet Motherboard	
	Mellanox ConnectX-4 100Gbit Ethernet	
OS	Fedora Core 27	

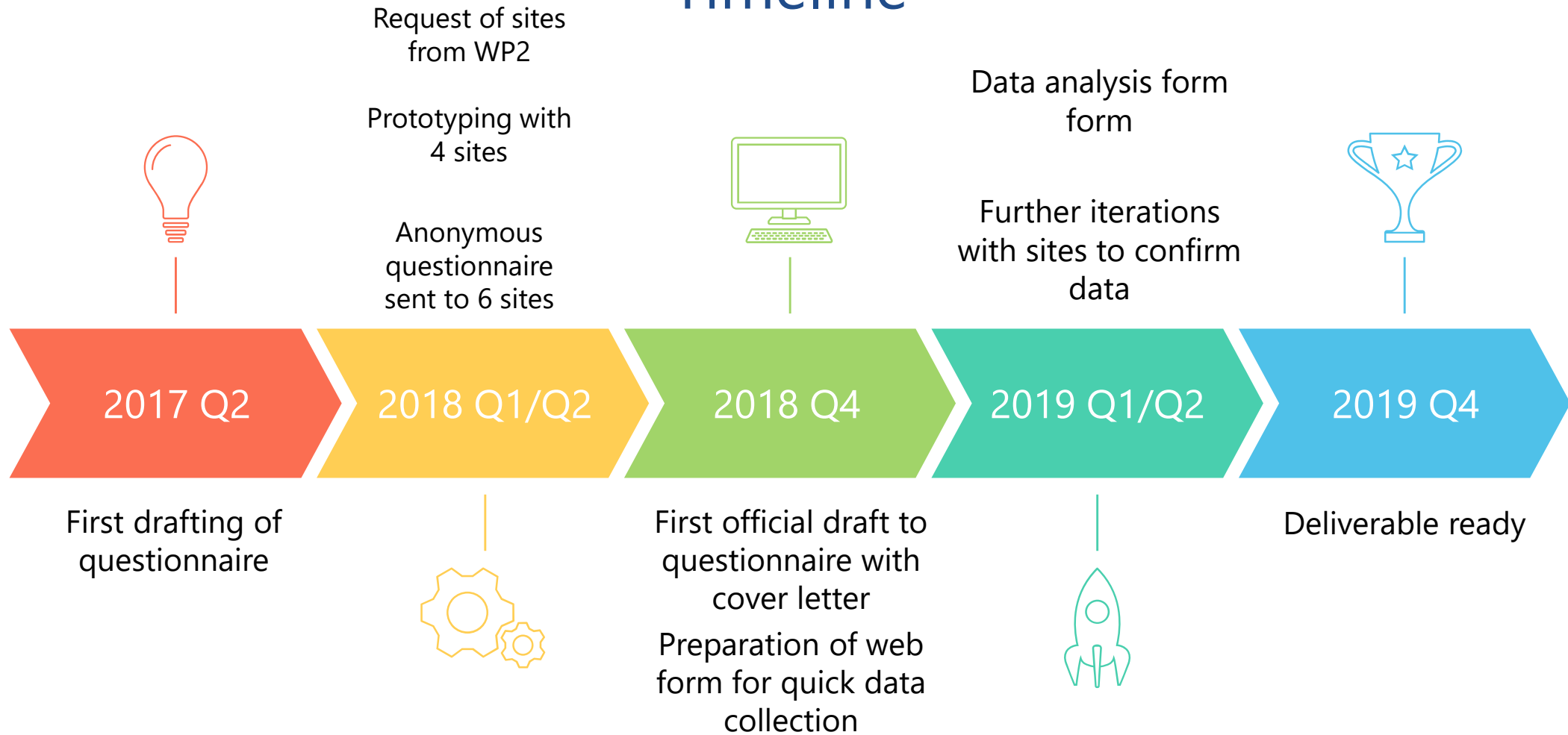


Task 4.2

D4.2 “Site Catalogue of the Storage and Network Capabilities”

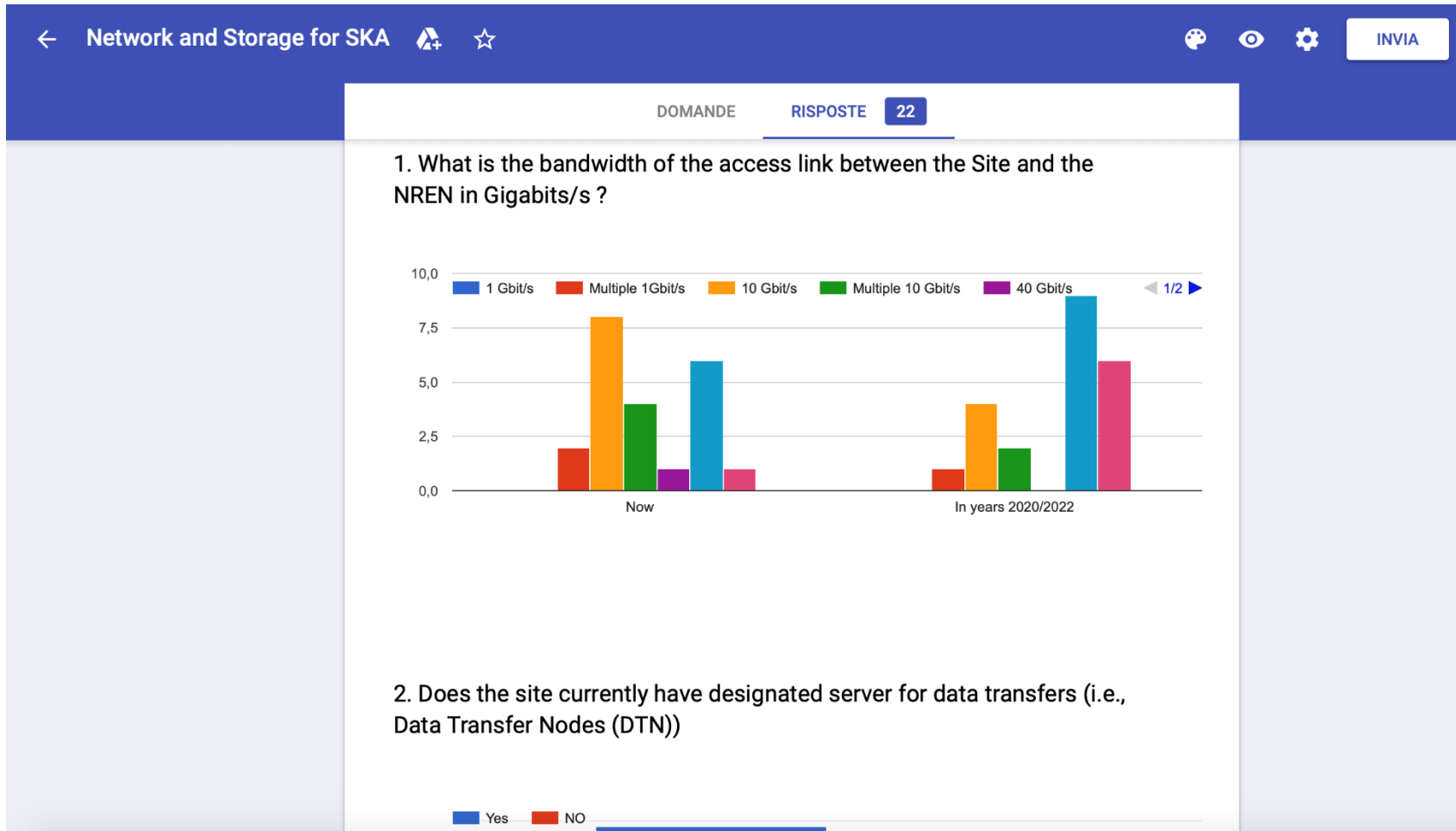
Matteo Stagni and Domenico Vicinanza

Timeline



Web Form and PDF

1



Questions on Networks and Storage for the Site Catalogue of Potential Infrastructure Providers

Work Package 4: Questionnaire for Deliverable D4.2





1st official contact - 19 Dec 2018

- Inventory of the storage and network capabilities of potential sites to form the SKA European Science Data Centre
- Pin-pointed sites



2nd official contact - 5 feb 2019

- Aeneas Survey on Network and Storage Capabilities of Sites
- PDF questionnaire with cover letter
- Link to web form to provide answers



Deadline for Answers – 15 feb 2019

- Latest answer was provided a month later
- Ongoing discussion in WP4 meetings how to validate the data...



Confirmation e-mail 10 Jun 2019

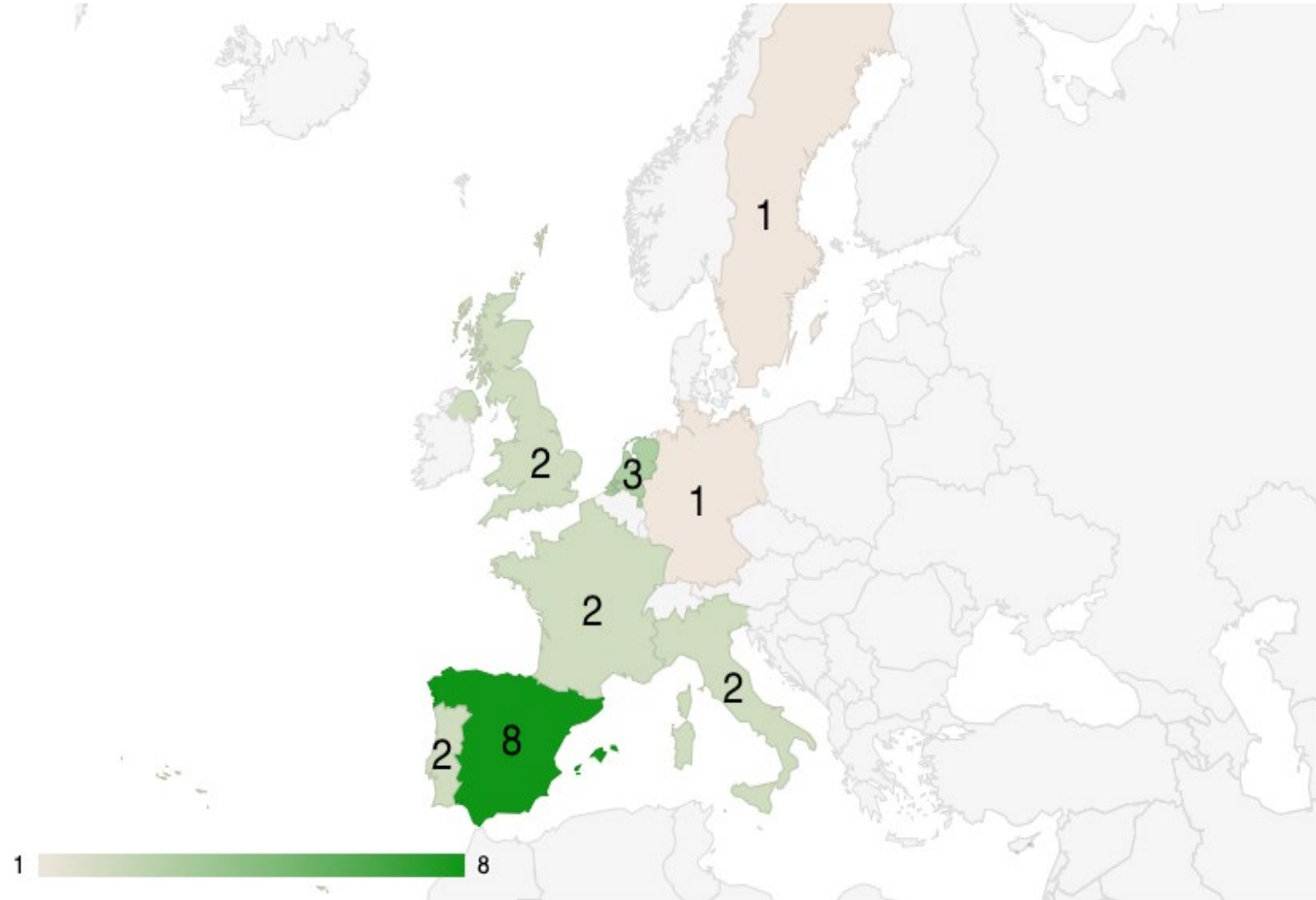
Dear ...,

we are kindly asking you to review the answer you provided for the AENEAS questionnaire 'Network and Storage for SKA' - there is no need for a reply in case you find the information to be truly representing your facility. Should you find any incongruences in your answers please reply to this email detailing the issue **NO LATER THAN JUNE 17th 2019**.

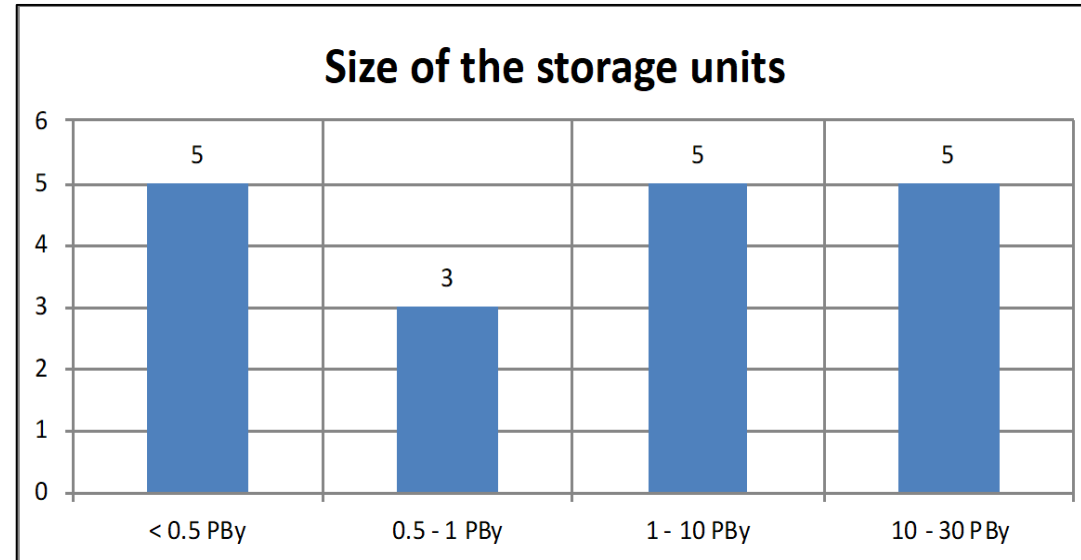
- Containing site data as reminder



Responses per Country

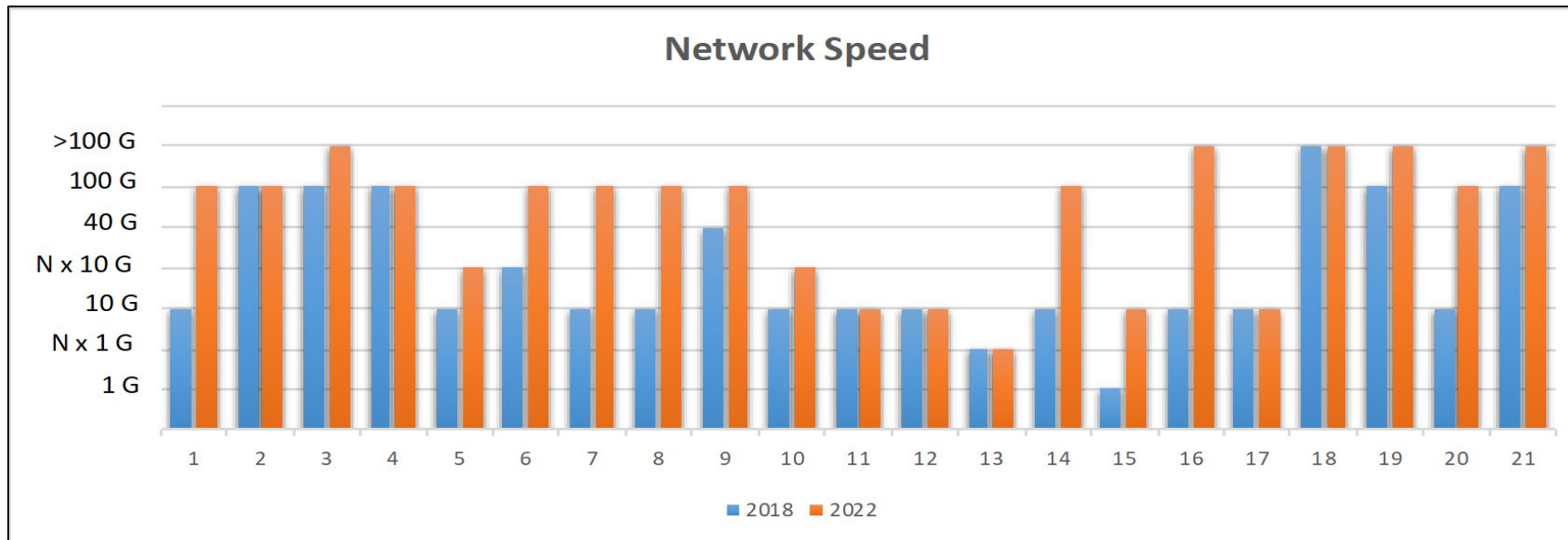


Storage – Size and writing speed



- Disk space ranges from 0.2PB to 29.5PB across sites
- Solid-state disks (SSDs) seems to be common
- Six sites use RAID-6 and three use RAID-5 (others: networked FS)
- Two thirds of the sites can handle (overall) writing speeds > 10Gbit/s.
- The local file system most commonly used is XFS.

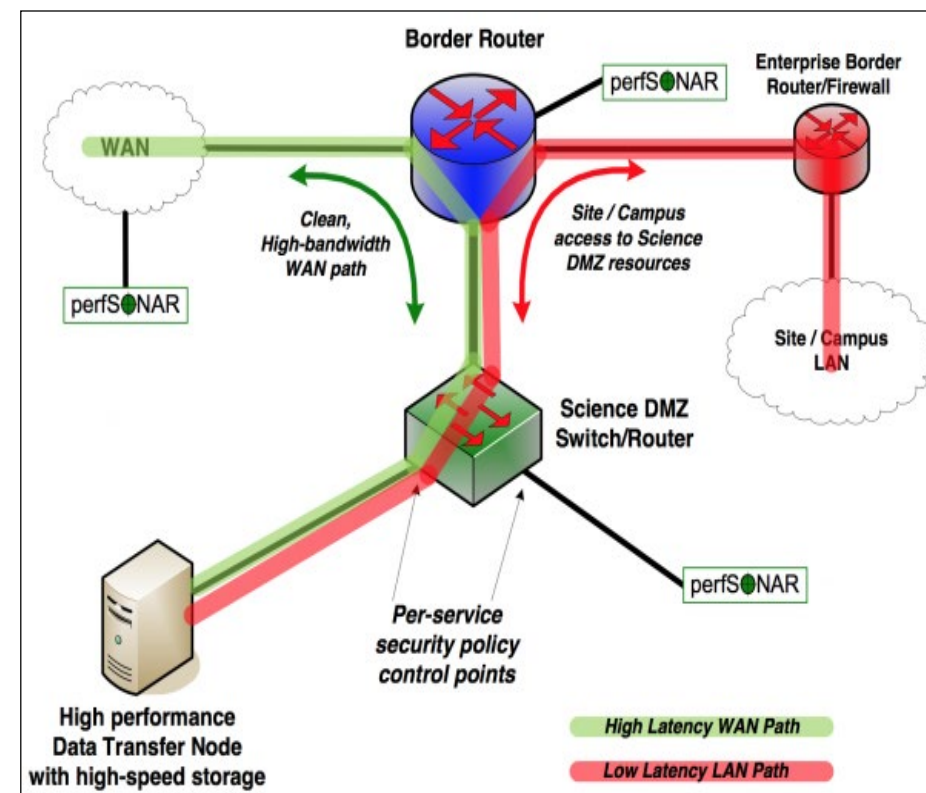
Network speed



- Most of the sites have at least 10Gbit/s connections
- 30% already reach 100Gbit/s
- 17 sites (66% of the total) see no issues in increasing their bandwidth to 100Gbit/s and beyond.

DTN, DMZ and Firewalls

- 60+% of the sites have or will have DTNs
- 50% of the sites have DMZ
- All sites have firewalls
 - Most of them report the firewalls are can deal with full available bandwidth.





Storage Network Protocol

- Infiniband-40-56 Gbit/s is the most common technology
 - to interconnect storage systems
 - to build parallel systems
- Some sites are evaluating the migration to Ethernet 100 Gbit/s:
 - higher latency but
 - could be more effective in transferring large amount of data.



Tape Libraries

- Still an essential tool, not just for backup copies. They will be used:
 - To store the SKA required quantities of data at an affordable price and low maintenance.
 - In data centres using a Hierarchical Storage Management (HSM)
 - To dynamically extend the capacity of a 'data lake'.
- Cheap to buy and operate:
 - Low costs: 12-14 EUR/TB (up to 150 EUR per TB with disk.)
 - Lower power: Tapes consume energy only when there is an IO request.
- However, they are slower: reading 1 TB takes 1 hour
- Tape libraries that can support more than 10000 cartridges and 32 reading devices.
- Current cartridges storage size: 6 TB
 - Forecast (10 years time): 48 TB/cartridge, 3 times faster (1TB in 20 min)



Conclusions

- Data centres for SKA present challenges that require addressing in the next years
- Data centre **networks** are sufficiently adequate
 - or they could become so by 2022
- No issues with the local NRENs providing the required bandwidth.
- Storage will have to be provided by SKA and funded by the SKA community.
- Data centres need to grow to twice the present size.
- Sites are willing to provide Science DMZ and Data Transfer Nodes
- The existing sites have personnel with considerable experience
- The implementers of the ESRC must discuss requirements with the centres.



Task 4.4

D4.5 “Report on Data Transport Tests
and Recommendations”

Update on AARNet & Indigo

Update on SANReN Fibre Rollout

Shaun Amy and Siju Mammen

The Network Paths between GÉANT (Lon, Par) – AARNet (Canberra, MRO)

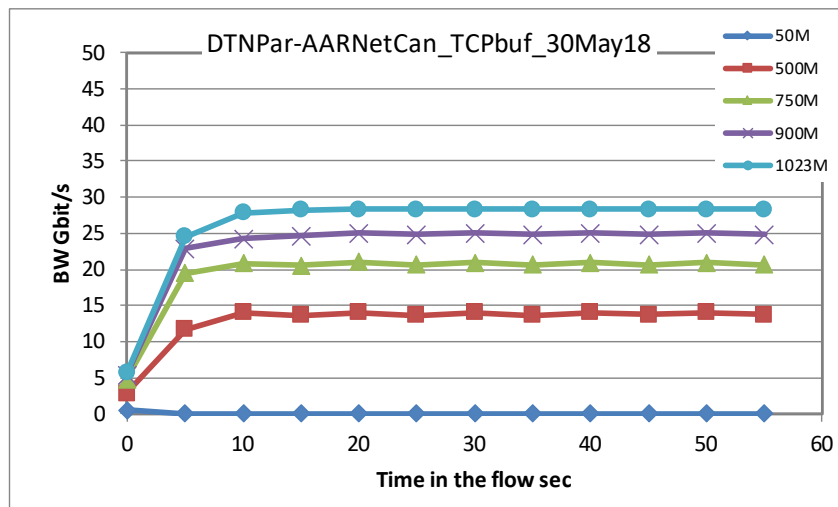
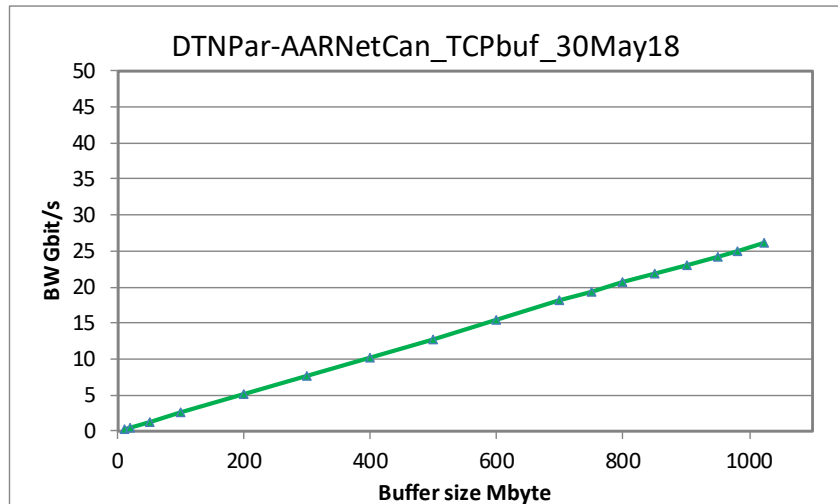
RTT 303 ms.



RTT 259 ms.

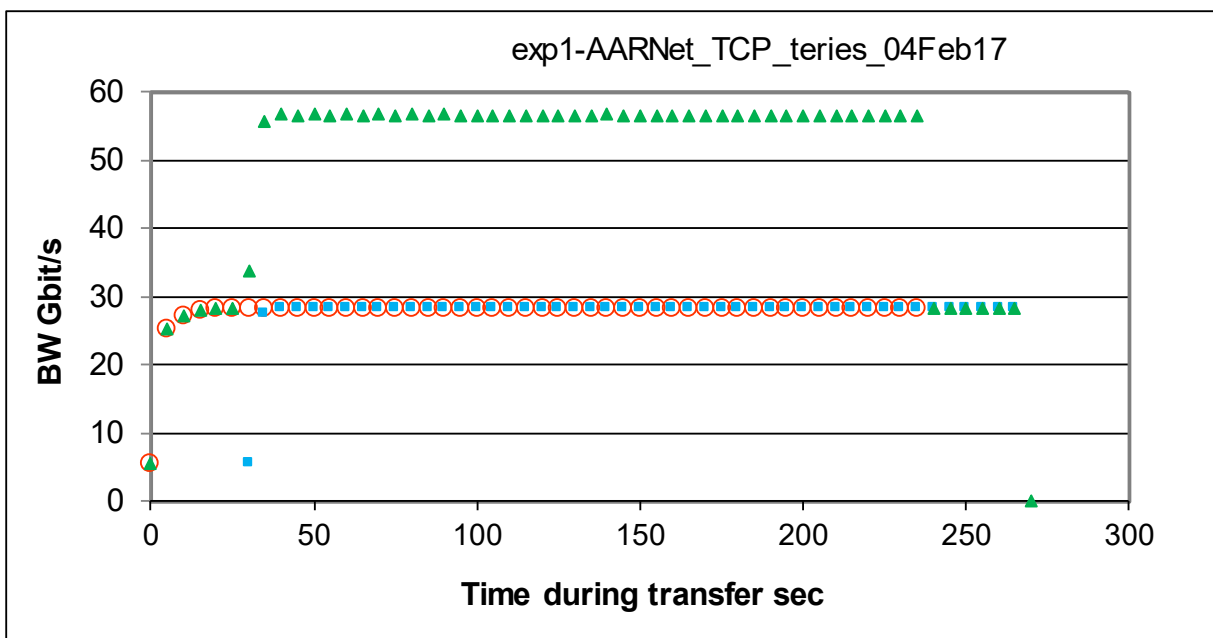


100 Gigabit between GÉANT Paris and AARNet Canberra



- Route GÉANT, ANA300, Internet2, & AARNet: Paris-New York-Seattle-LosAngeles-Sydney-Canberra
- TCP offload on, TCP cubic stack
- RTT 303 ms.
- Delay Bandwidth Product 3.78 GB for 100 Gigabit
- One TCP flow rises smoothly to **26.1 Gbit/s** at 1023 MBytes including slowstart.
- No TCP re-transmitted segments
- **Rate after slowstart 28.3 Gbit/s**
 - Plateau after ~15s
- Reach the limit of TCP protocol
Max TCP window is 1 Gbyte
- Rate for RTT 303 ms and TCP window 1023 MB
28.32 Gbit/s
- CPU core only 75-80 % in kernel mode

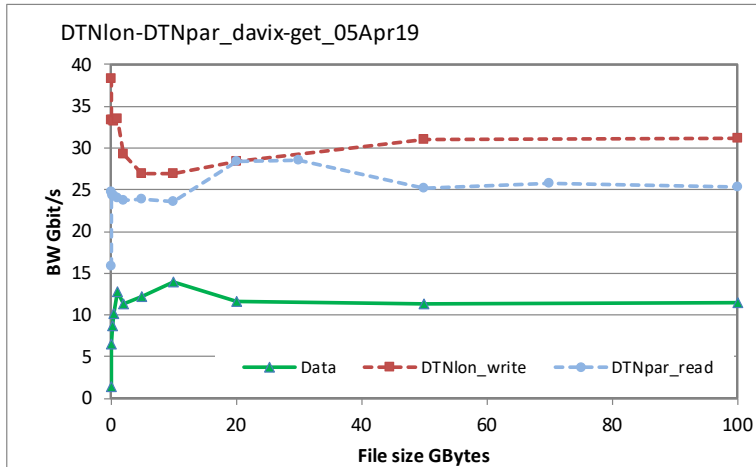
100 Gigabit: Multiple flows between GÉANT and AARNet



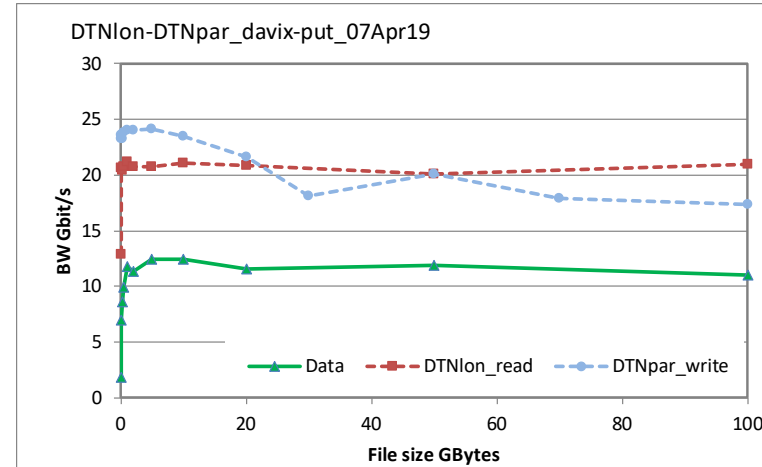
- Route GÉANT, ANA300, Internet2, & AARNet: Paris-New York-Seattle-LosAngeles-Sydney-Canberra
- RTT 303 ms.
- TCP window 1023 MB.
- Two 4 minute TCP flows
- Second flow started 30s after the first
- Each flow stable at 28.3 Gbit/s
- Total transfer rate 56.6 Gbit/s
- 1.55 TBytes data sent in 4.5mins.
- No TCP segments re-transmitted.
- Demonstrates the stability of the academic network

Disk to Disk Throughput vs File Size Scan davix:http

davix-get DTNlon-DTNpar



davix-put DTNlon-DTNpar



- 1 TCP flow
- Concern re low transfer speeds half disk speeds.
- Disk - memory
Lon zfs write 30 Gbit/s
Par xfs RAID0 read 25 Gbit/s

Instrument davix and xrdcp

- Time measurement better than 1 sec
- Set fixed size chunk reads from the network
- Measure times for "getChunk" (socket read) & "putChunk" (disk write)
- Note TCP performance parameters using tcp_info struct
- Record as time series in memory
- Option not to write a file (davix-get)

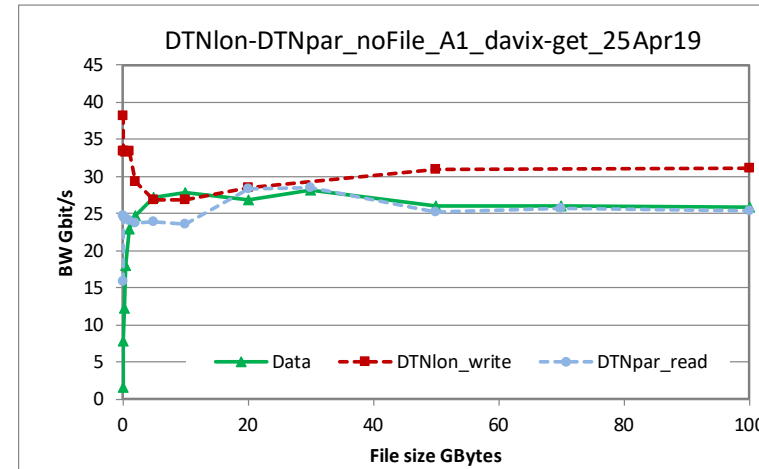
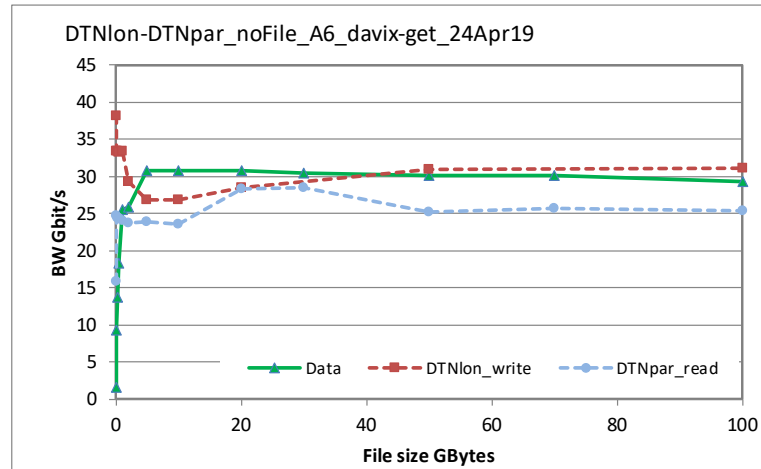
- Help from Andy Hanushevsky CHEP: lots of improvements in Xrootd v4.11.0
- But Tests done with v 4.8.4 !
 - Using 1 TCP flow
 - Did set the chunk size
- So data should be OK

davix-get: Effect of CPU core. File Size Scan, (No File Write)

DTNlon-DTNpar A6 good core

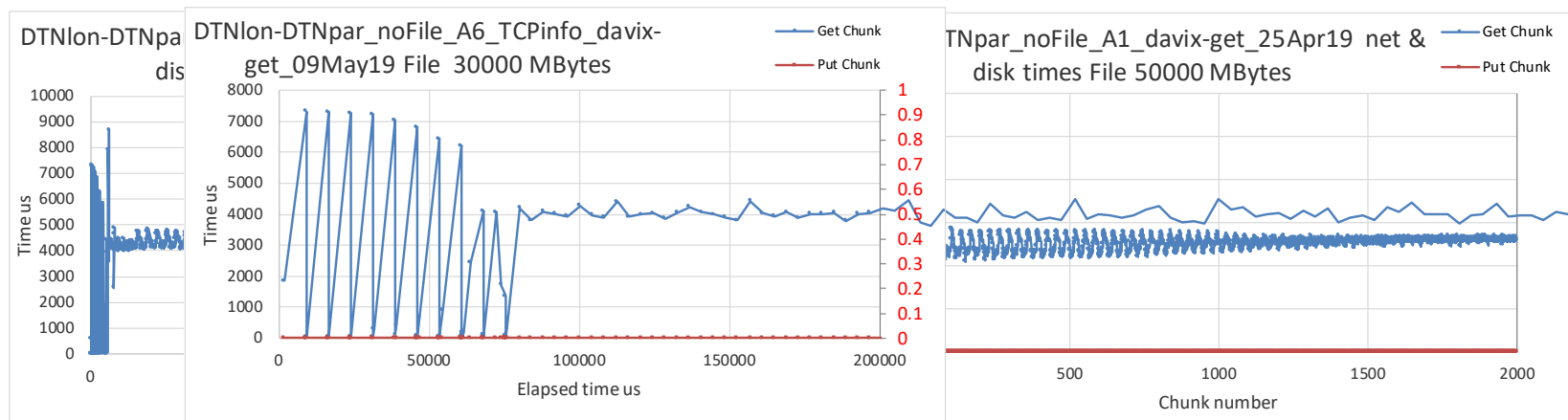
DTNlon-DTNpar A1 bad core

Throughput



- 1 TCP flow
- Disk - memory
Lon zfs write 30 Gbit/s
Par xfs RAID0 read 25 Gbit/s
- Use of bad core ~5 Gbit/s less similar to iperf3
- Constant read size

getChunk, putChunk time series



- Network getChunk times more stable locked to a CPU
- Note variation at start of the transfer
- Linked to TCP Slowstart

- Shows good performance of Xrootd server.

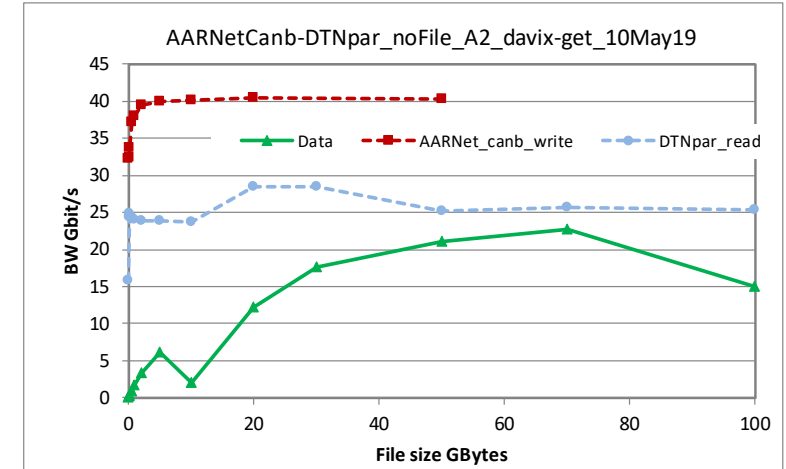
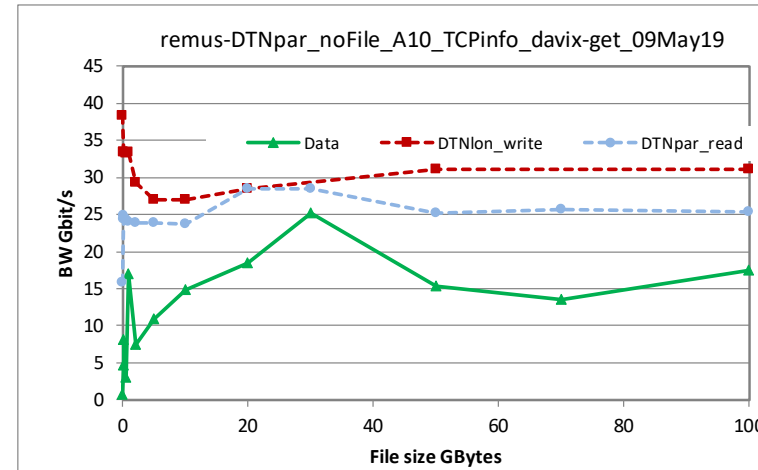
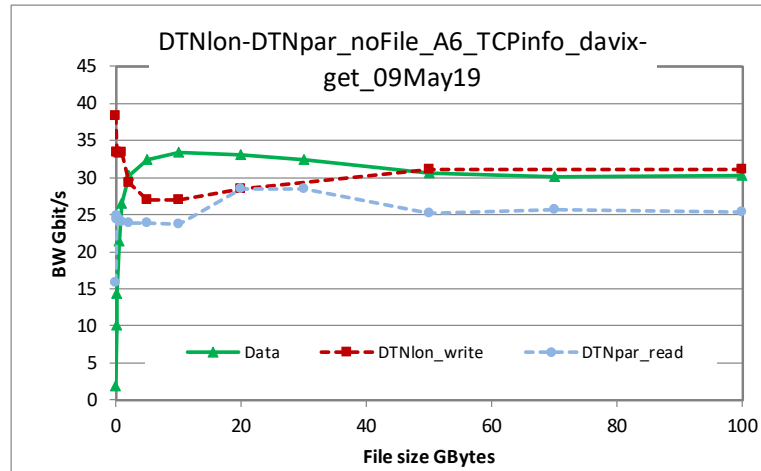
davix-get: Different RTT (No File Write)

DTNlon-DTNpar RTT 7.4 ms

remus-DTNpar RTT 14.1 ms

AARnetCanb-DTNpar RTT 302ms

Throughput

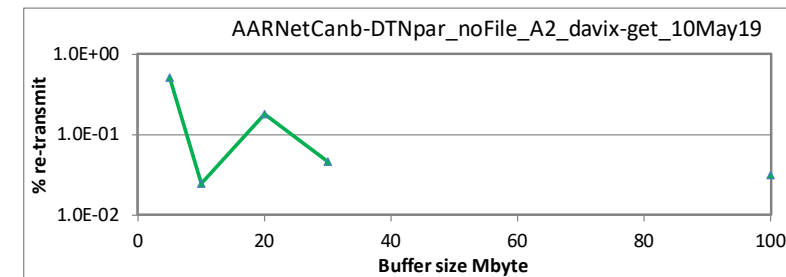
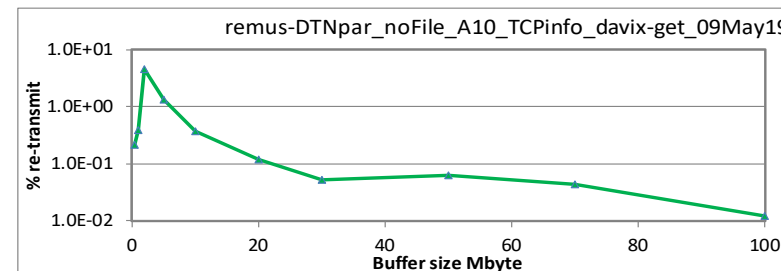


- >30 Gbit/s for files > 2 Gbytes
- Remote disk → local memory at disk speed
- Shows good performance of Xrootd server

- Packet loss – need TCP re-transmits
- Impact on performance

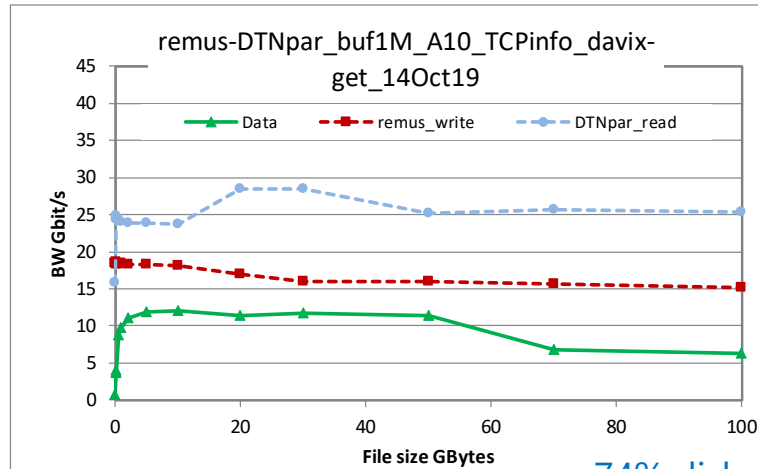
TCP re-transmits

No packet loss

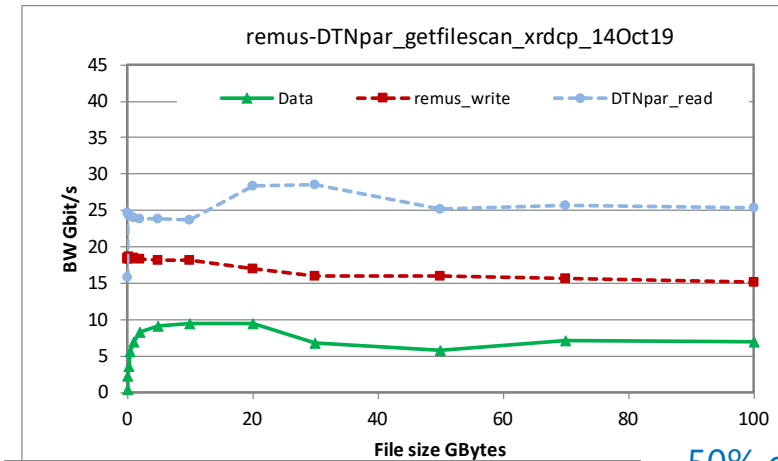


Protocol Comparison JBOremus ← DTNpar RTT 14.1 ms

davix-get //http: 1MB buffer

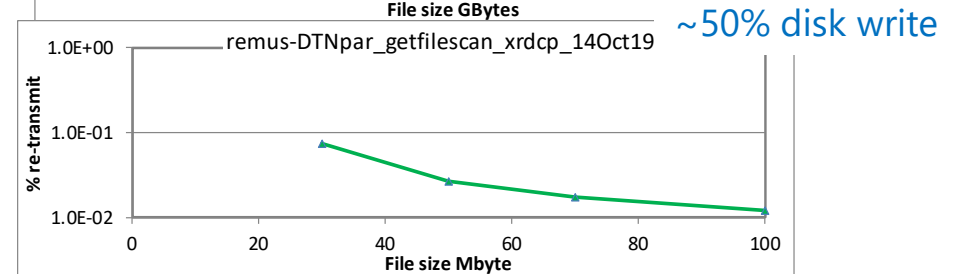
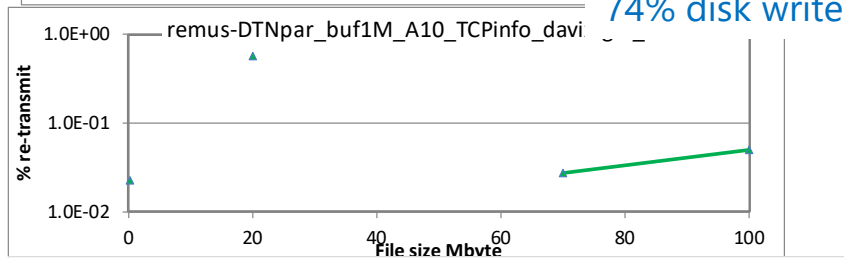


xrdcp //root: 8 MB buffer

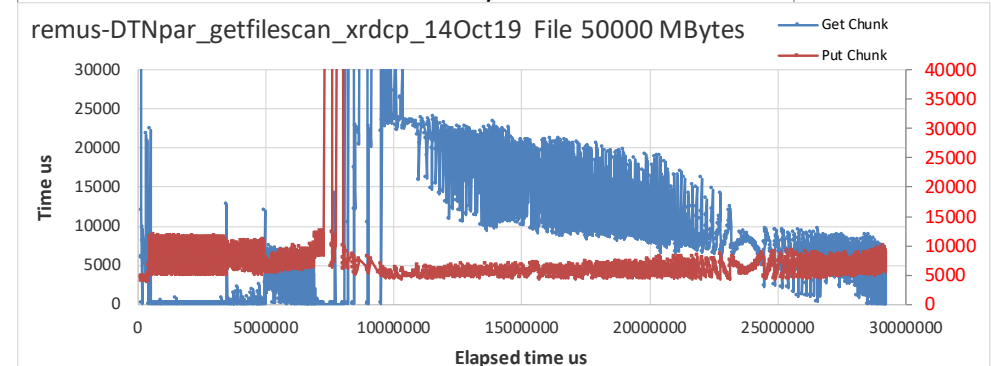
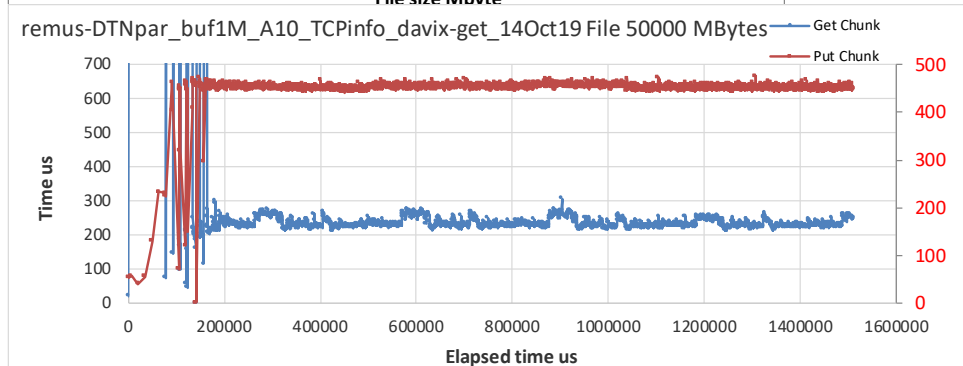


**Throughput
Disk-to-Disk**

TCP re-transmits



**Chunk time series
Net read
Disk write**

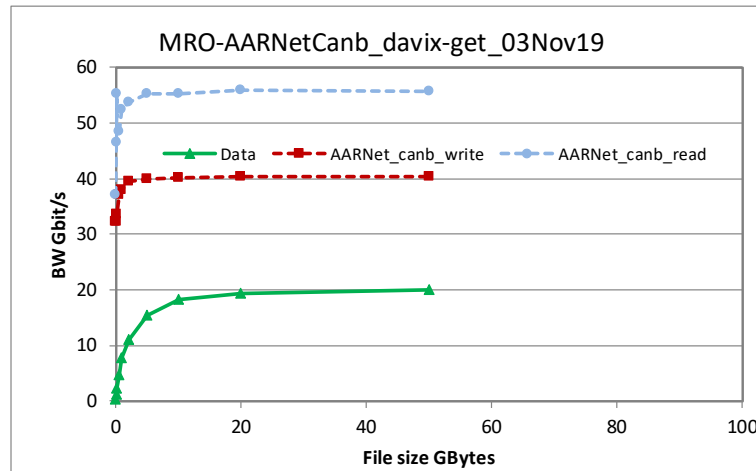


Protocol Comparison MRO ← AARNetCanb RTT 56.5 ms

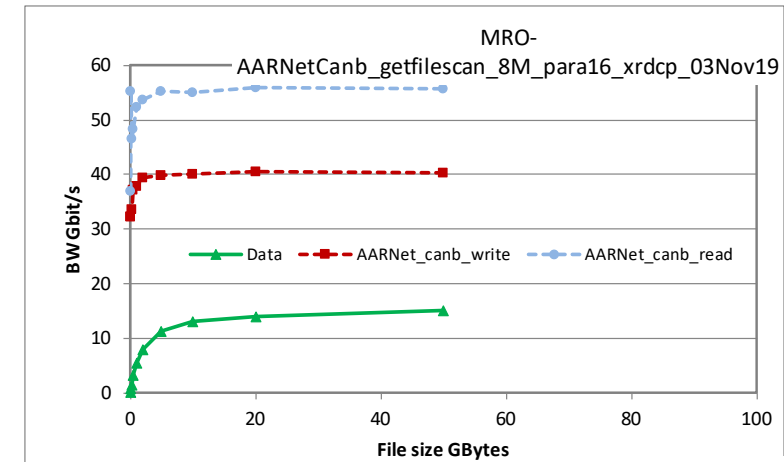
davix-get //http: 1MB buffer

xrdcp //root: 8 MB buffer XRD // CHUNKS=16

**Throughput
Disk-to-Disk**

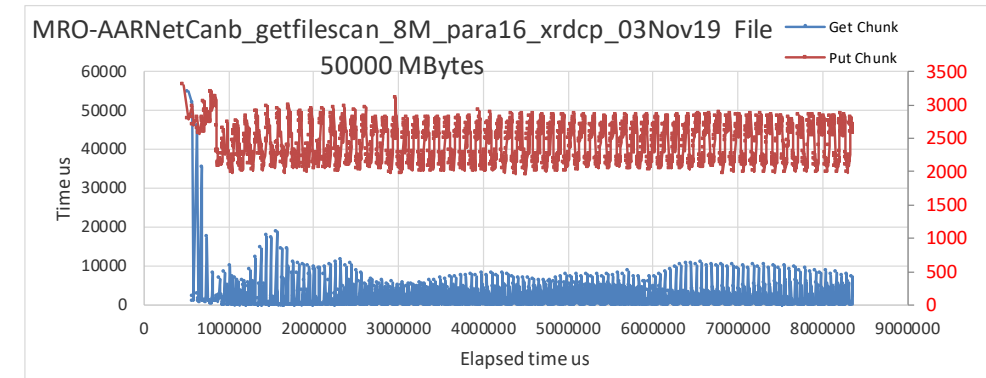
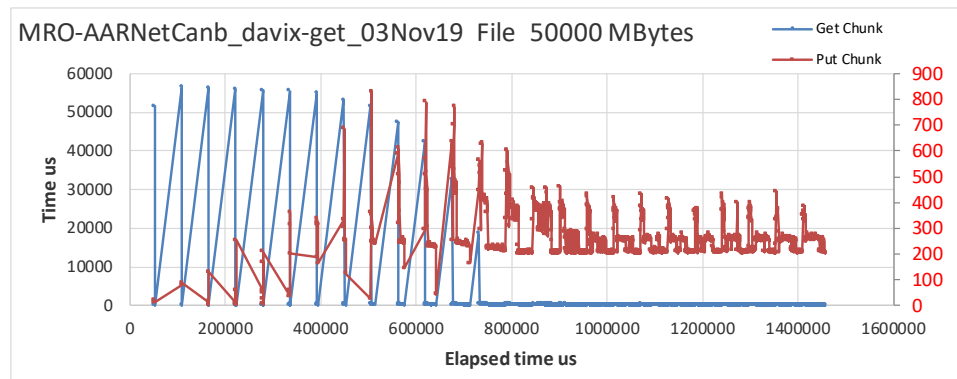


50% disk write



~37% disk write

**Chunk time series
Net read
Disk write**



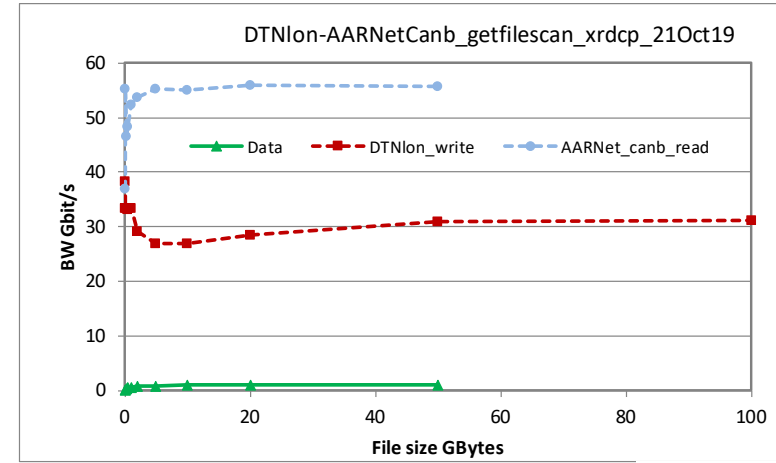
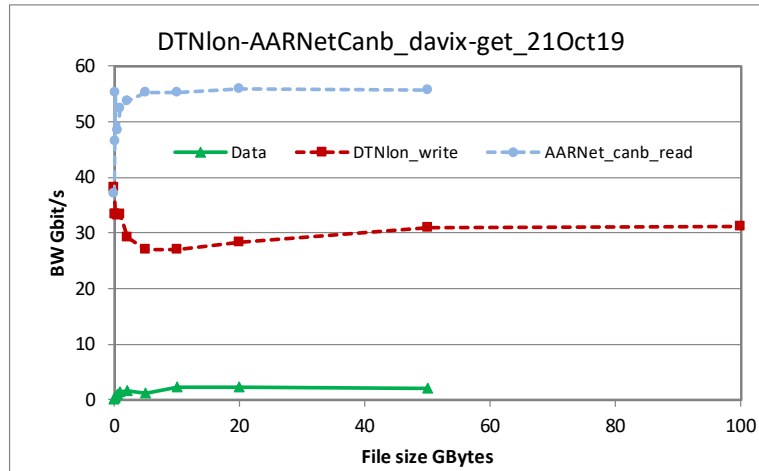
Stable after TCP slowstart.

Protocol Comparison DTNlon ← AARNetCanb RTT 259 ms

davix-get //http: 1MB buffer

xrdcp //root: 8 MB buffer XRD // CHUNKS=16

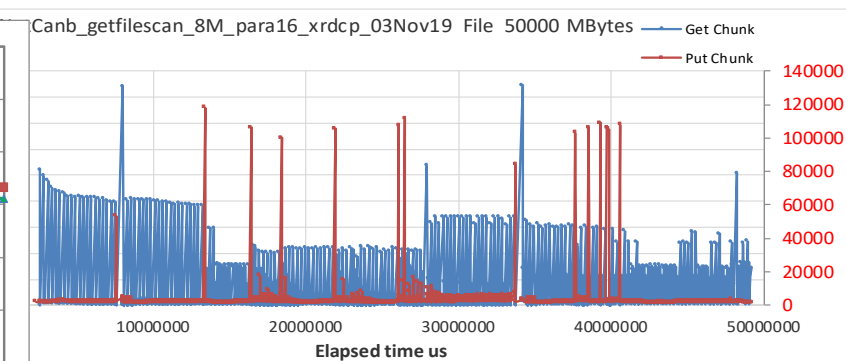
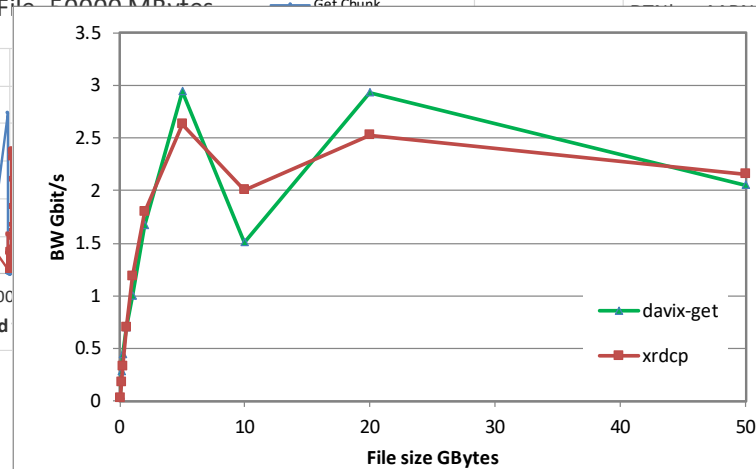
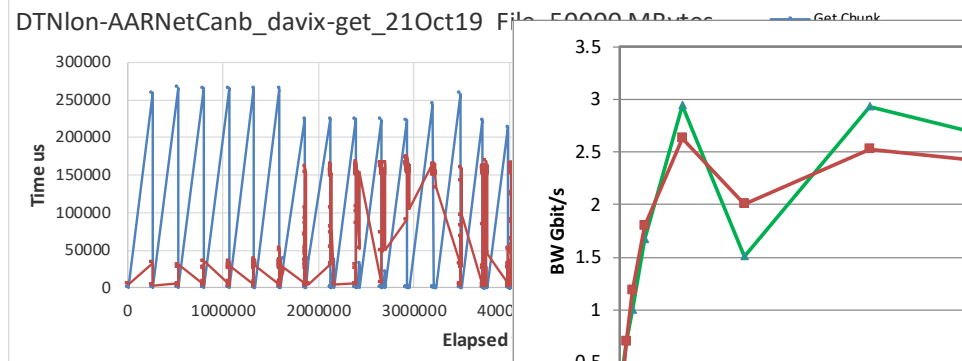
**Throughput
Disk-to-Disk**



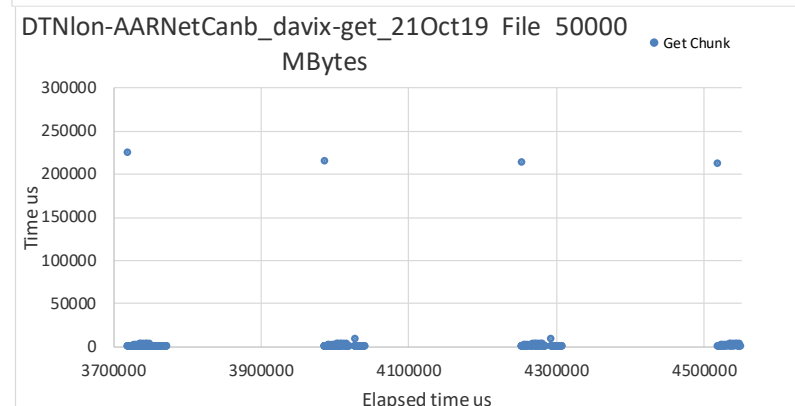
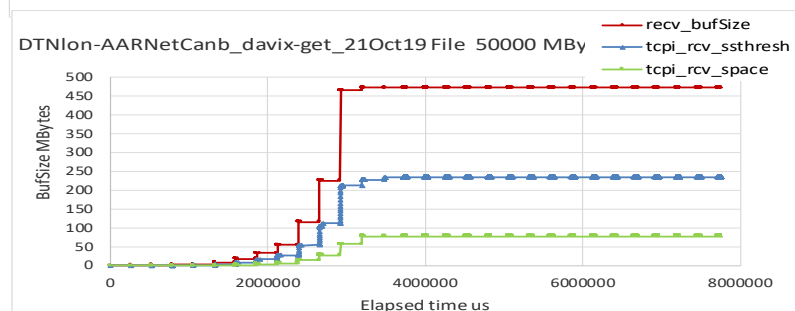
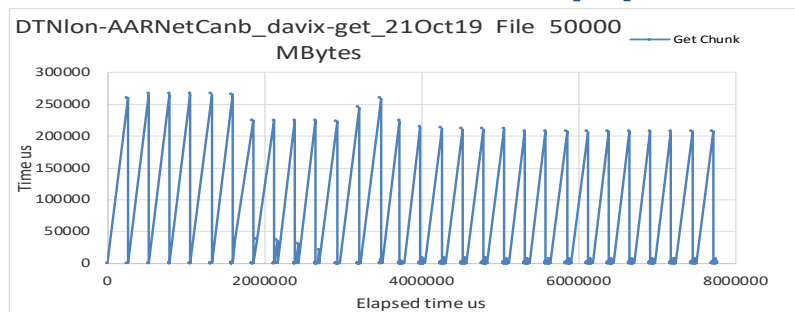
~7.4% disk write

~3% disk write

**Chunk
time series
Net read
Disk write**



What is happening between DTNlon and AARNetCanb?



- Look at the “Get Chunk” Network read times
 - Large value = delay in reading from the network
- Would expect stability after TCP slowstart
- Slowstart ends at approx $10 * RTT$ 259 ms. $\approx 2.6s$
- Confirmed by the receive buffer size plateau.
- Delay Bandwidth Product for 3 Gbit/s ~ 97 MBytes
- Every RTT
Read about 100 MBytes then wait 210ms
- Why?
 - App not providing data to the TCP socket
 - TCP auto tuning is working but slower than expected over these RTT – iperf is OK
- Investigations continue

What have we learnt?

- WebDAV/http(s) and xrd protocols both work well for moving bulk data.
- A simple client loop of “Get Chunk” - “Put Chunk” clearly reduces disk-to-disk throughput.
- Application disk-to-disk transfers work well up to RTT 56 ms.
- Use of zero-copy e.g. sendfile() on the server gives a big improvement.
- Use of multiple parallel disk accesses is a help.
- TCP will send what the app gives it – keep the socket full.
- TCP auto-tuning works well at medium RTT but may be slower at large RTT.



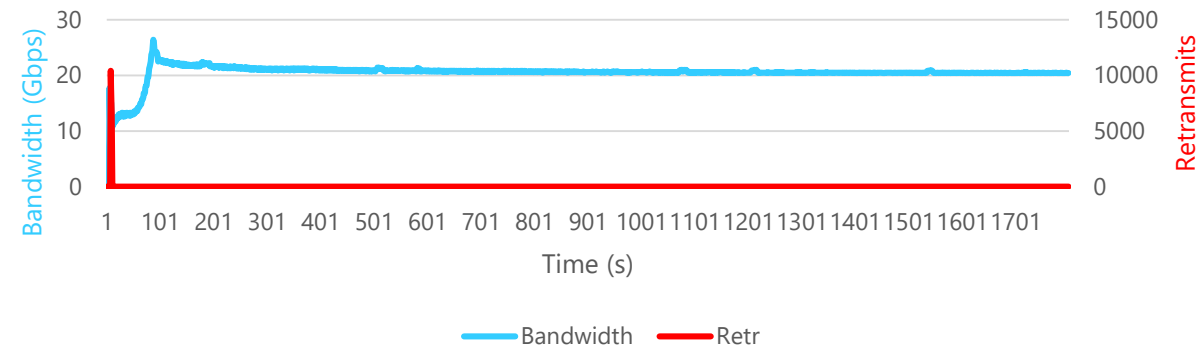
Next steps

- Check out xrootd v 4.11.0 server and client
- Look at multiple TCP flows

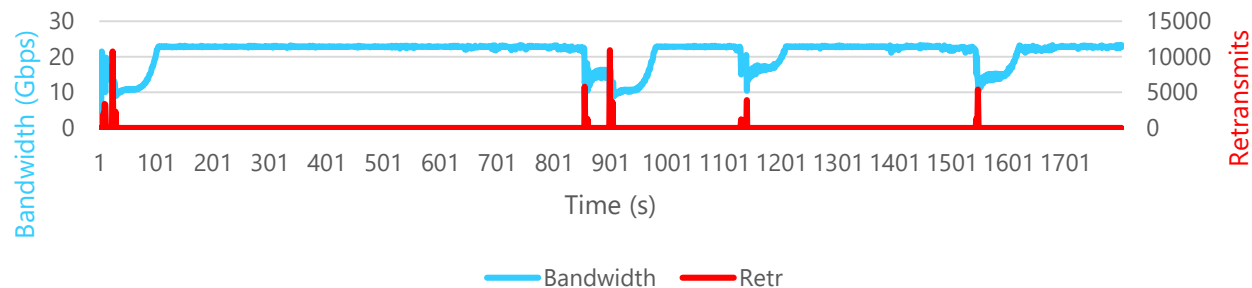
Long haul iperf3 tests

- Canberra to JBO via Singapore
- 30 minute test
- Default TCP auto-tune transfers (let TCP find best parameters)
- Pin processes on sender and/or receive

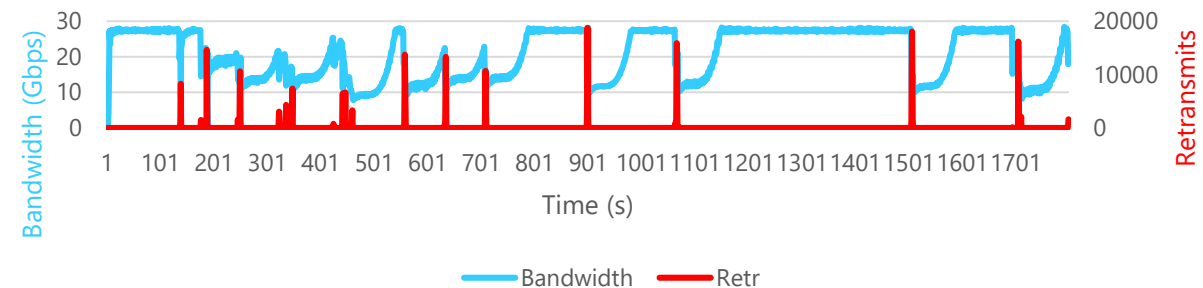
30 minute TCP iperf3 Canberra to JBO
20.5 Gbps average



30 minute TCP iperf3 Canberra to JBO
client processor affinity
21.0 Gbps average



30 minute TCP iperf3 Canberra to JBO
client and server processor affinity
20.5 Gbps average



XRootD transfers romulus to remus at JBO

- Transfers of files of various sizes – 1 GB, 10 GB & 100 GB
- Modes of transfers:
 - Consecutive
 - Simultaneous (multiple (300, 30 or 3) xrdcp processes running at the same time)
 - Recursive from a directory
 - Recursive with multiple TCP streams

File size and number	Consecutive (Gbps)	Simultaneous (Gbps)	Recursive (Gbps)	Recursive with 15 TCP streams (Gbps)
300 x 1 GB	12.16	38.94	12.92	13.70
30 x 10 GB	14.52	43.07	13.33	14.56
3 x 100 GB	11.28	24.79	10.51	10.42

XRootD transfers JBO to Paris

File size and number	Consecutive	Simultaneous	Recursive	Recursive with multiple TCP streams
300 x 1 GB	5.49	28.29	8.77	8.51
30 x 10 GB	12.06	34.64	10.16	9.57
3 x 100 GB	9.14	16.09	10.09	9.49

XRootD transfers Canberra to JBO

Transfers of files of various sizes – 1GB, 10GB & 80GB
Simulate SKA-low to European SRC transfer

File size and number	Consecutive	Simultaneous	Recursive	Recursive with multiple TCP streams
80 x 1 GB	0.88	33.95	1.65	1.60
8 x 10 GB	1.36	12.46	1.27	1.19
	1 TCP Stream	2 Streams	4 Streams	15 streams
1 x 80 GB	1.48	1.33	1.22	1.39

Pinning processes to a particular core (taskset)

Best performance for a single 80 Gbps file 1.16 Gbps (single or multiple TCP streams)



Update on AARNet & Indigo

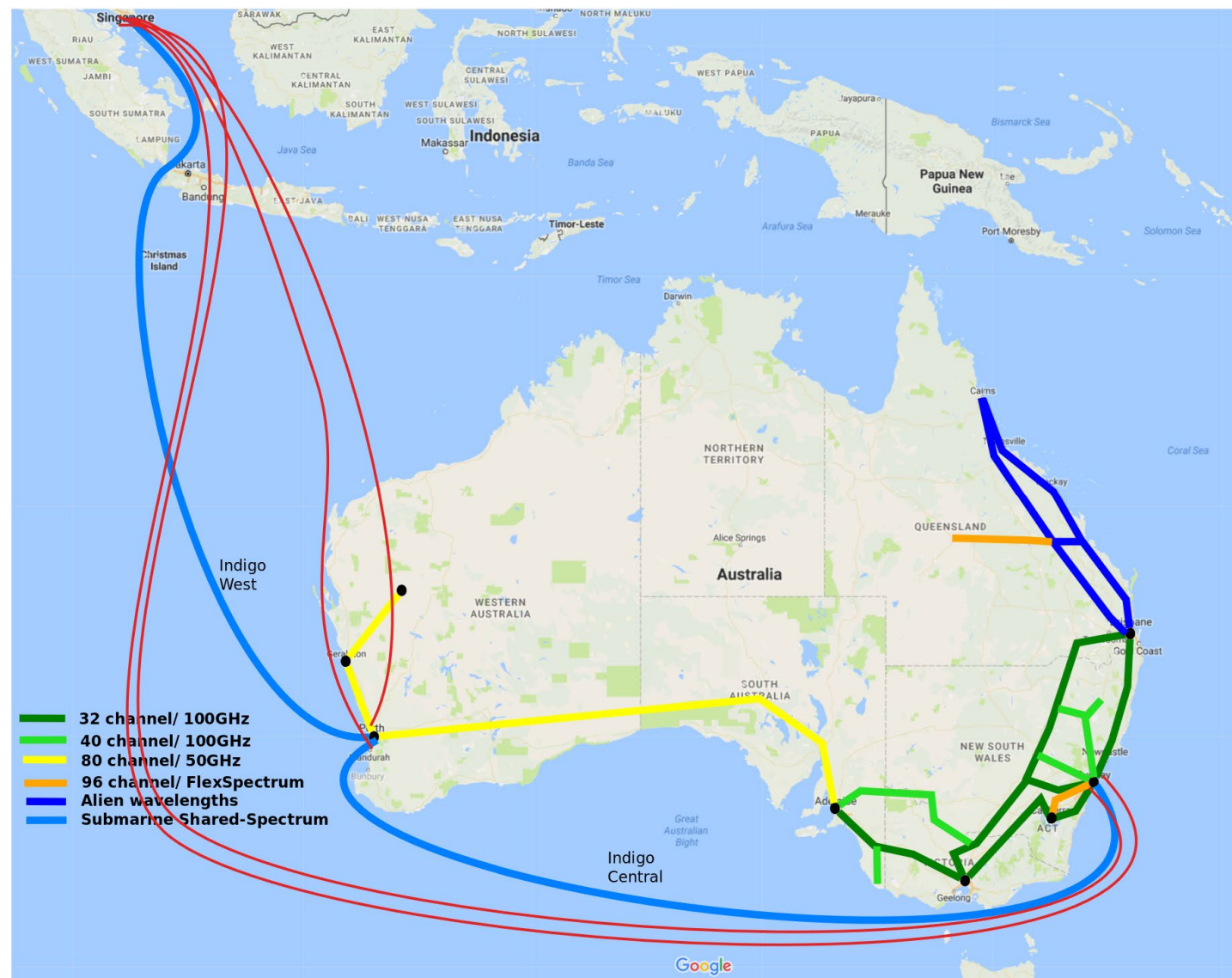
Shaun Amy & Tim Rayner

Coming Ashore



The AARNet Optical Network

- Began ~2004 now > 12,000km of DWDM network
- Using Cisco NCS2006/ONS15454 equipment (0 – 15 years old)
- Queensland Alien wavelength project: 2015 to 2017
- CSIRO were early drivers for high bandwidth paths for eVLBI.





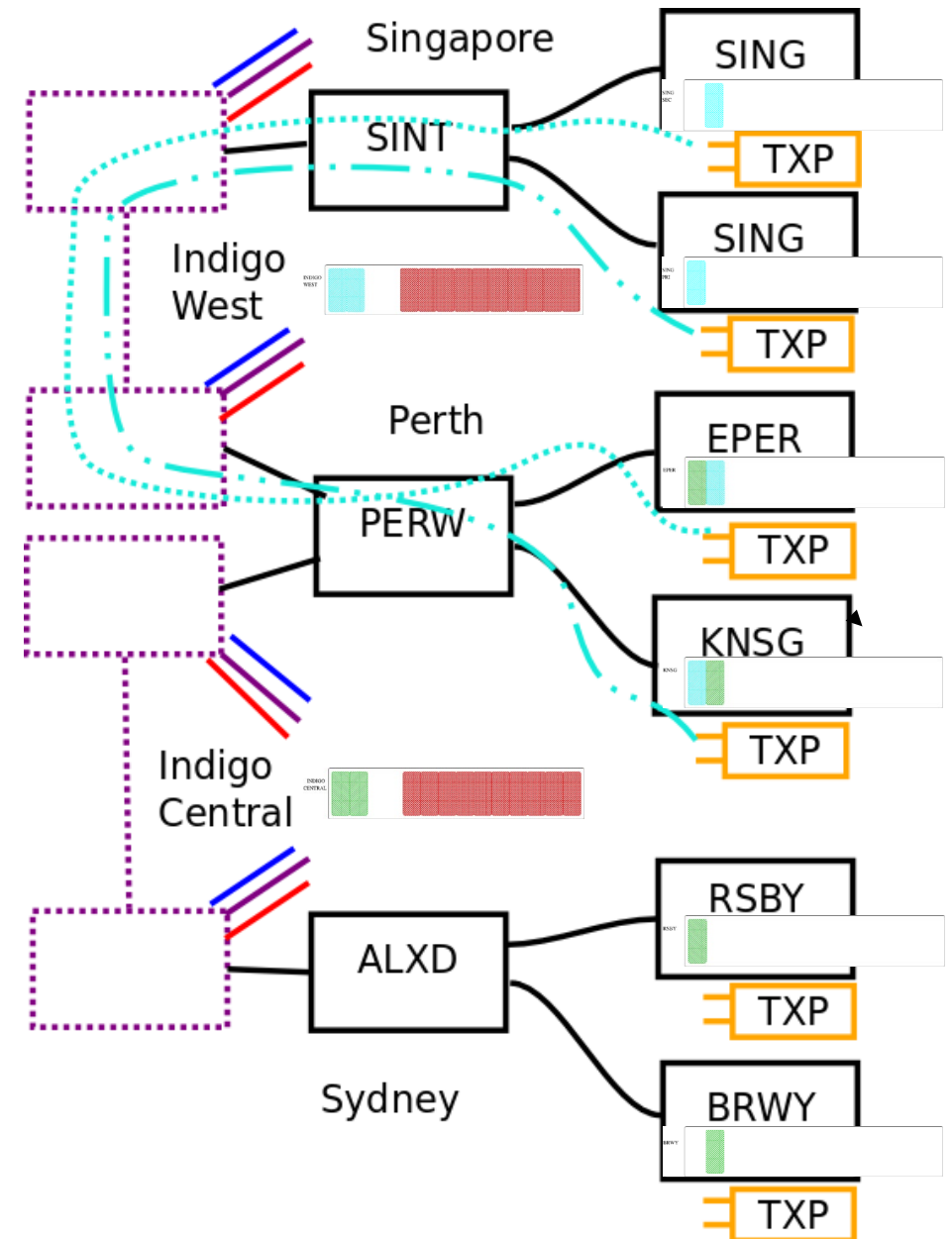
INDIGIO Details

- No in-built dispersion compensation
- Indigo West
 - 2 Fibre pairs Perth to Singapore
 - Additional 2 pairs Singapore to Jakarta – not relevant for us
- Indigo Central
 - 2 Fibre pairs Sydney to Perth
- Client Capacity
 - Capacity is assigned in 1/4 spectrum "blocks"
 - AARNet has capacity on both Indigo West and Indigo Central – using the same wavelengths.
 - This allows wavelengths to be routed directly from Central to West without re-gen.



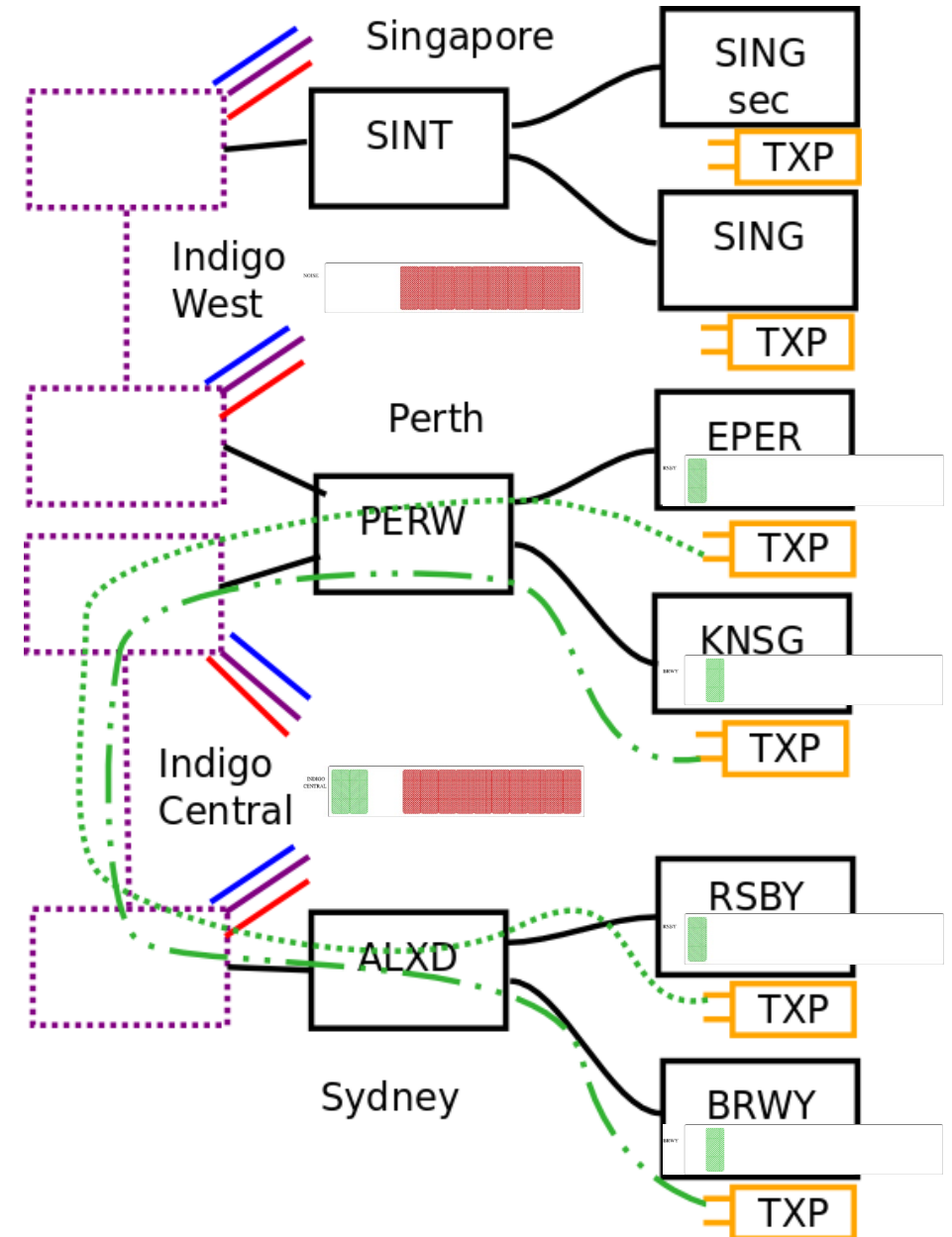
AARNet INDIGIO Interfacing

- Transponders away from the landing stations
- Maximise Diversity
- Diverse tails to regular PoPs
- ROADM nodes @ PoPs, multi-degree ROADMs at landing stations
- Tail length: 2 to 50km (+360km to Canberra also tested)
- Own redundant noise generation @ single site with remote loopback
- Aim to avoid Indigo Noise Insertion ie. maintain a transparent spectrum from client to client.
- 3 dB drop when half channels are removed.



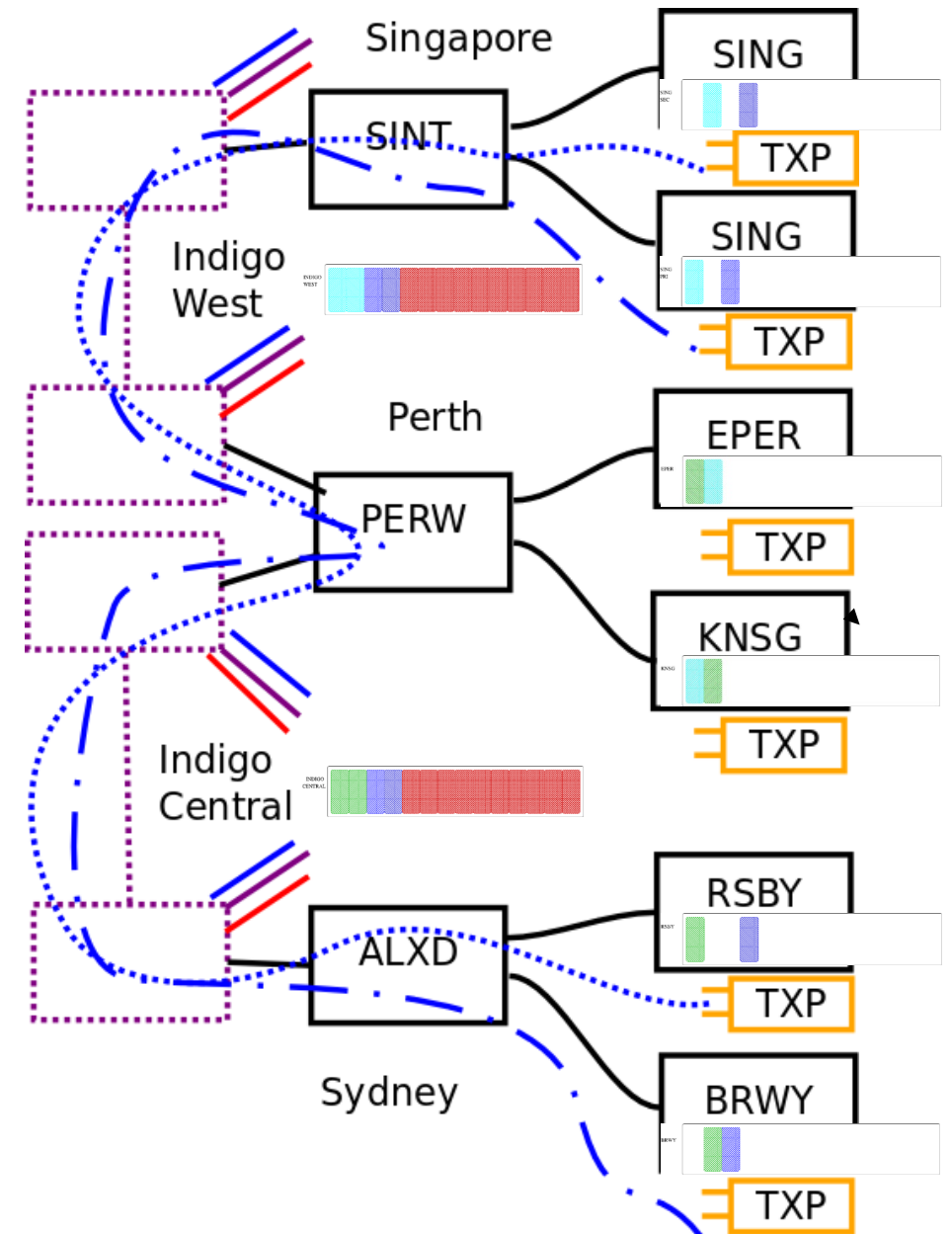
AARNet INDIGIO Interfacing

- Transponders away from the landing stations
- Maximise Diversity
- Diverse tails to regular PoPs
- ROADM nodes @ PoPs, multi-degree ROADMs at landing stations
- Tail length: 2 to 50km (+360km to Canberra also tested)
- Own redundant noise generation @ single site with remote loopback
- Aim to avoid Indigo Noise Insertion ie. maintain a transparent spectrum from client to client.
- 3 dB drop when half channels are removed.



AARNet INDIGIO Interfacing

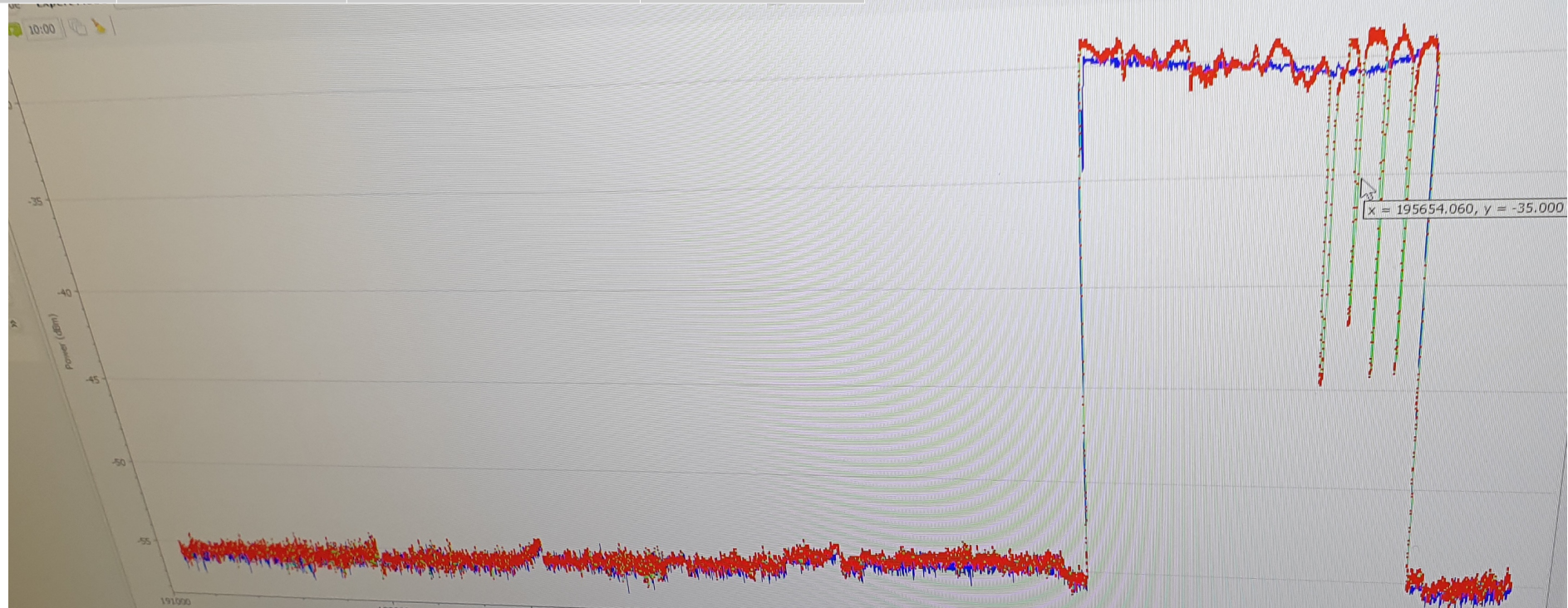
- Transponders away from the landing stations
- Maximise Diversity
- Diverse tails to regular PoPs
- ROADM nodes @ PoPs, multi-degree ROADMs at landing stations
- Tail length: 2 to 50km (+360km to Canberra also tested)
- Own redundant noise generation @ single site with remote loopback
- Aim to avoid Indigo Noise Insertion ie. maintain a transparent spectrum from client to client.
- 3 dB drop when half channels are removed.



INDIGIO Performance

	Distance	Best λ	Channel Width	Production λ	Channel Width
SYD-PER	4600km	400G	87.5 GHz	300G	75 GHz
PER-SIN	4600km	400G	87.5 GHz	300G	75 GHz
SYD-SIN	9200km	300G	87.5 GHz	200G	75 GHz

- We will have 1Tbps of production wavelengths running over Indigo – both Central & West.
- 2 x 300G + 2 x 200G
- Next Upgrade ? - We have plenty of spare spectrum – Higher Baud rates ?





CAE-1

- Consortium of 6 NRENs:
AARNet, GÉANT, NORDUnet,
singAREN, SURF, TEIN*CC
- 100 Gbit/s Singapore - London
- Indigo and CAE-1 Significantly change
how we globally connect the world.





Update on SANReN Fibre Rollout

Shaun Amy and Siju Mammen

Brief South Africa NREN Overview

- The South African National Research and Education Network (SA NREN) consists of an infrastructure and services commons, established through a long-term collaboration between the Tertiary Research and Education Network of SA (TENET) and the South African National Research Network (SANReN)
- SA NREN is regarded as one of the 16 world-leading NRENs - Member of Global REN CEO's Forum
- Service more than 1.2 million users (students, researchers, lecturers, admin staff, etc) daily
- All 26 public universities connected, most research councils, larger research initiatives (e.g. SKA, SALT)
- Multiple 10 Gbps links make up the national backbone network, upgrade to 100Gbps under way
- Several major metro networks at 10 Gbps each (Johannesburg, Tshwane, Ethekewini, Cape Town, Bloemfontein, East London), several new metros under way (e.g. Pietermaritzburg), upgrades to commence to 100 Gbps for various
- SANReN owns 7.3% of total available capacity on the West Africa Cable System (WACS) undersea cable

Current SANReN Network

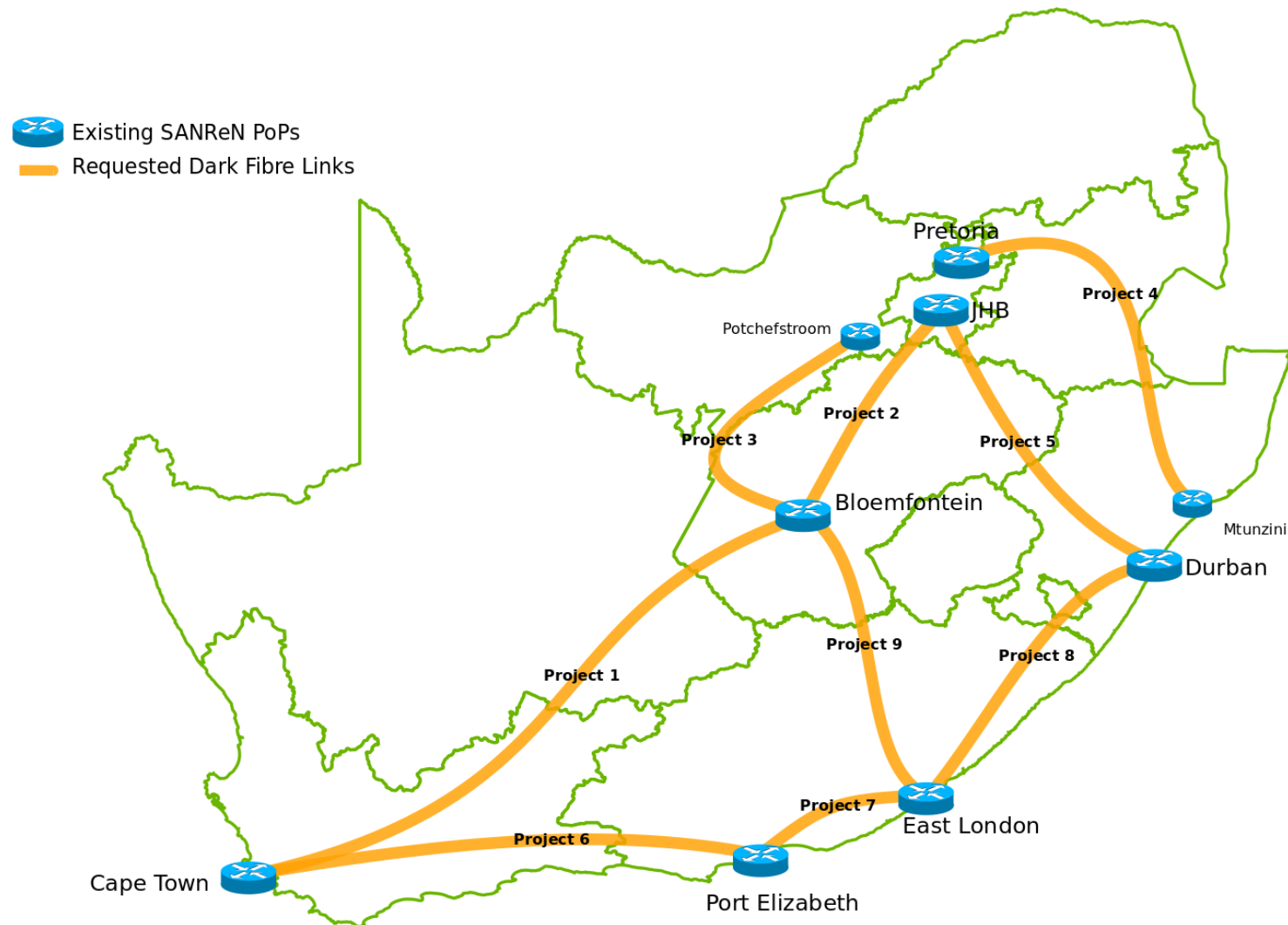




SANReN National Dark Fibre project



- SANReN embarked on a project to procure National Dark Fibre in 2015.
- The goal was to replace portions of the national backbone with dark fibre and run a DWDM system on top of the fibre
- Rationale:
 - Existing Backbone links were becoming extremely constrained.
 - Large projects (e.g. SKA) required access to Dark Fibre to transfer the enormous data that they produce.
 - Purchasing managed services from Telcos are very expensive.
 - Dark Fibre gives SANReN the flexibility to activate new capacity as required.
- The project will cost SANReN approximately €10 million for the Fibre and €3 million for the DWDM equipment when completed.

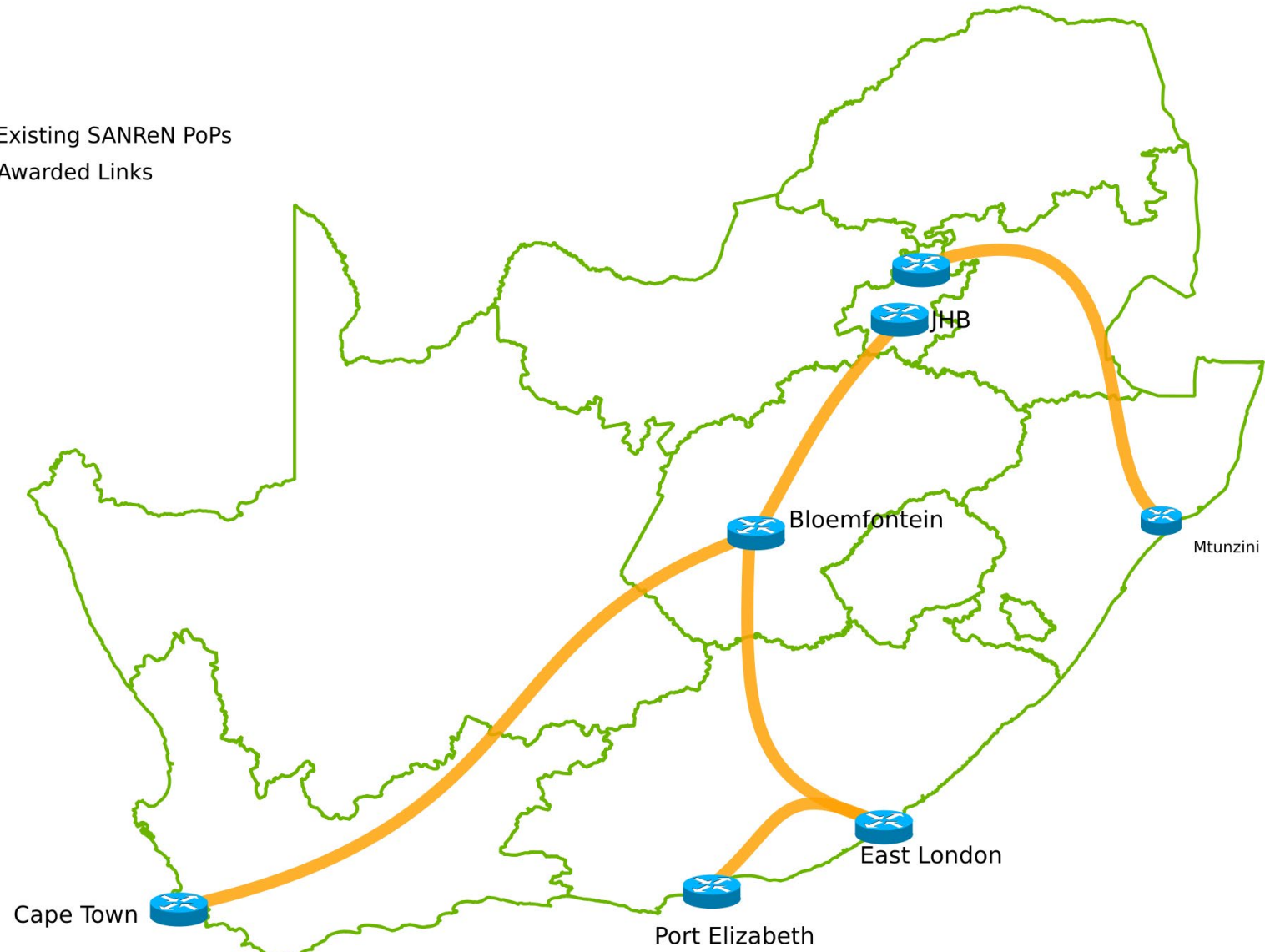
Requested Fibre links





Awarded Fibre Links

-  Existing SANReN PoPs
-  Awarded Links









Current status of SANReN Fibre Rollout

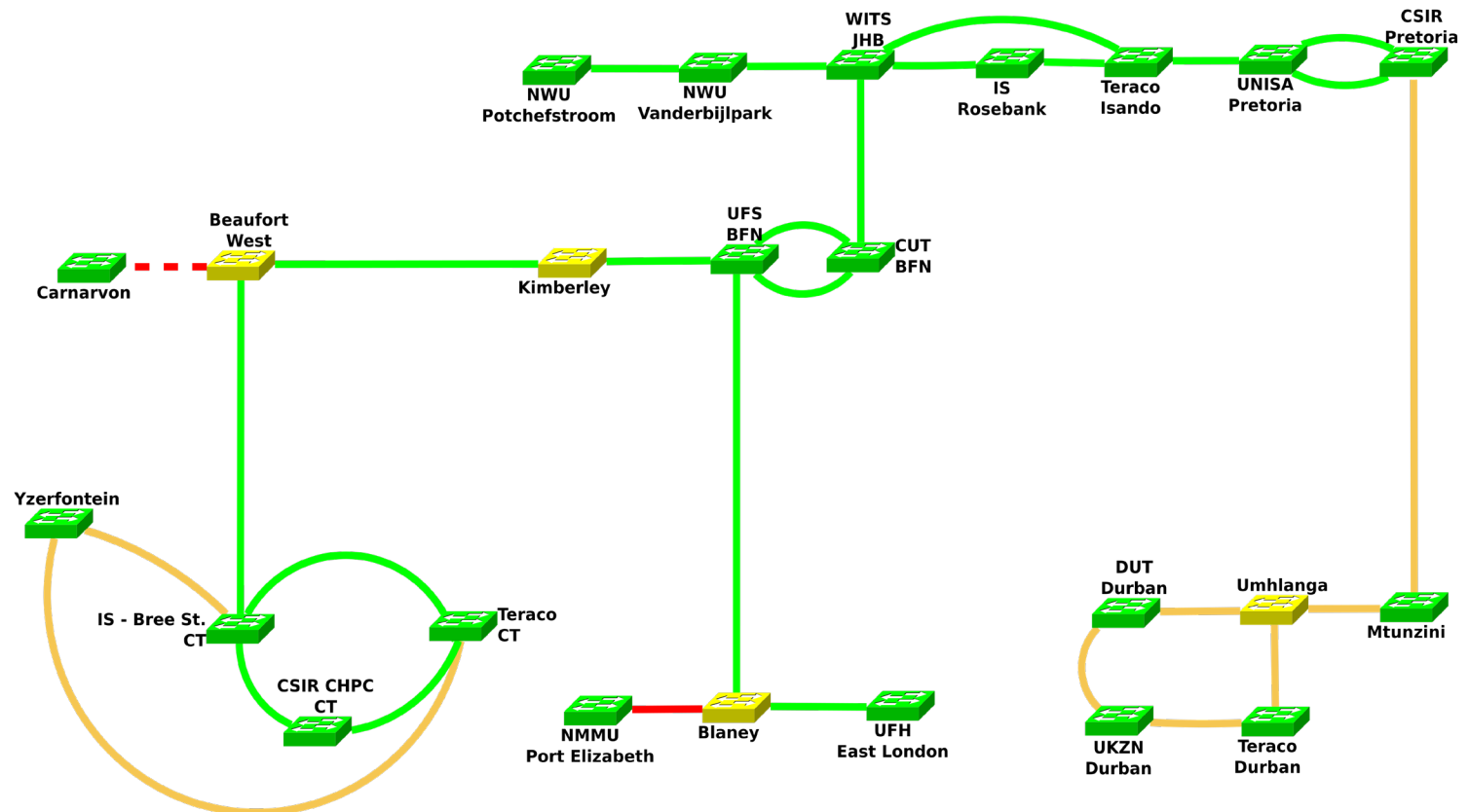
SA NREN 100Gbps DWDM Network

List of Acronyms

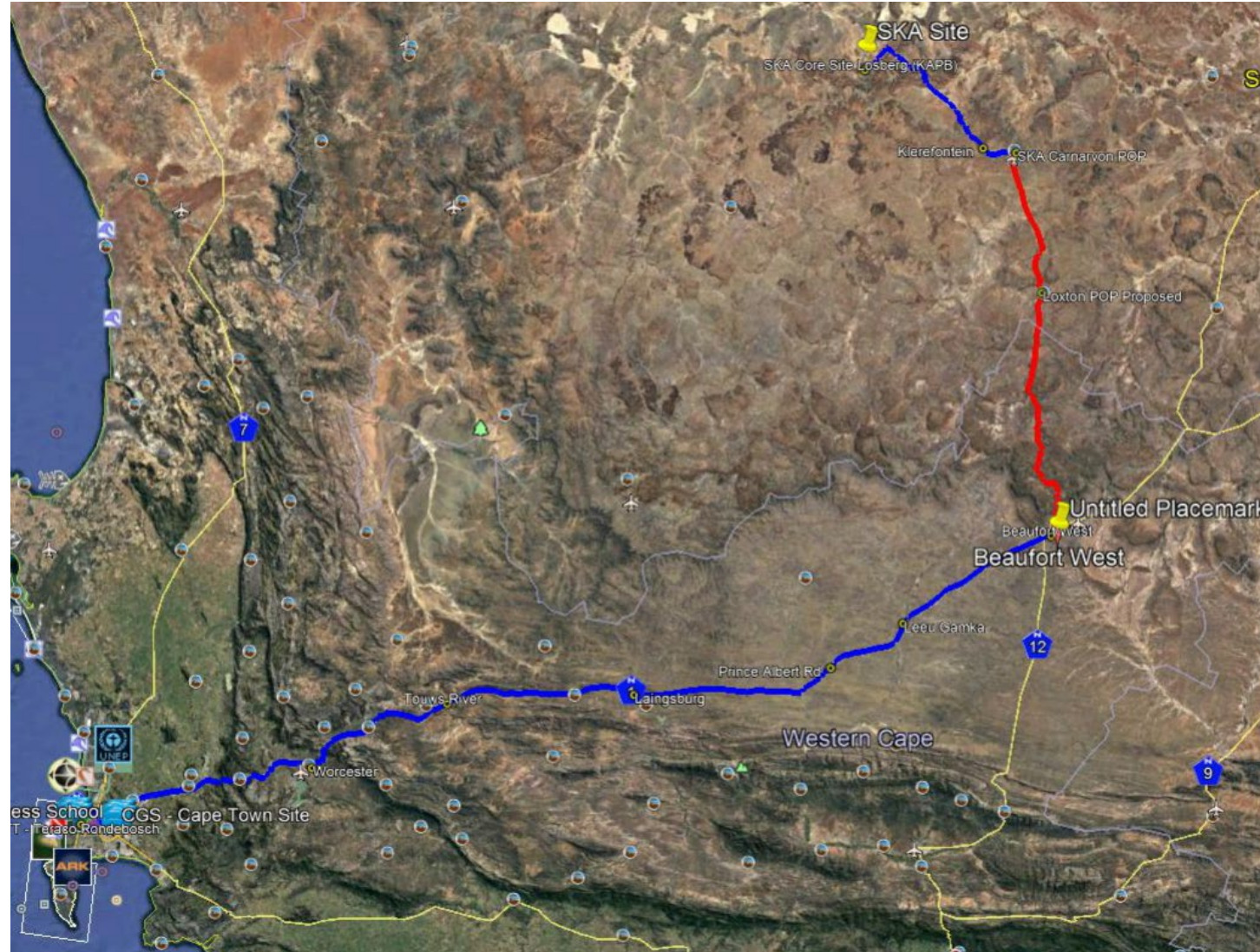
BFN - Bloemfontein
CHPC - Centre for High Performance Computing
CSIR - Council for Scientific and Industrial Research
CT - Cape Town
CUT - Central University of Technology, Free State
DUT - Durban University of Technology
IS - Internet Solutions
JHB - Johannesburg
NMMU - Nelson Mandela Metropolitan University
NWU - North West University
UFH - University of Fort Hare
UFS - University of the Free State
UKZN - University of Kwazulu Natal
UNISA - University of South Africa
WITS - University of the Witwatersrand

LEGEND

-  SANReN POP
-  Add Drop at Regen Site
-  DWDM links operational
-  DWDM links imminently operational (<1 month)
-  DWDM links to become operational in 2 - 3 months
-  DWDM links to become operational in 18 months



SANReN Fibre Network and the SKA








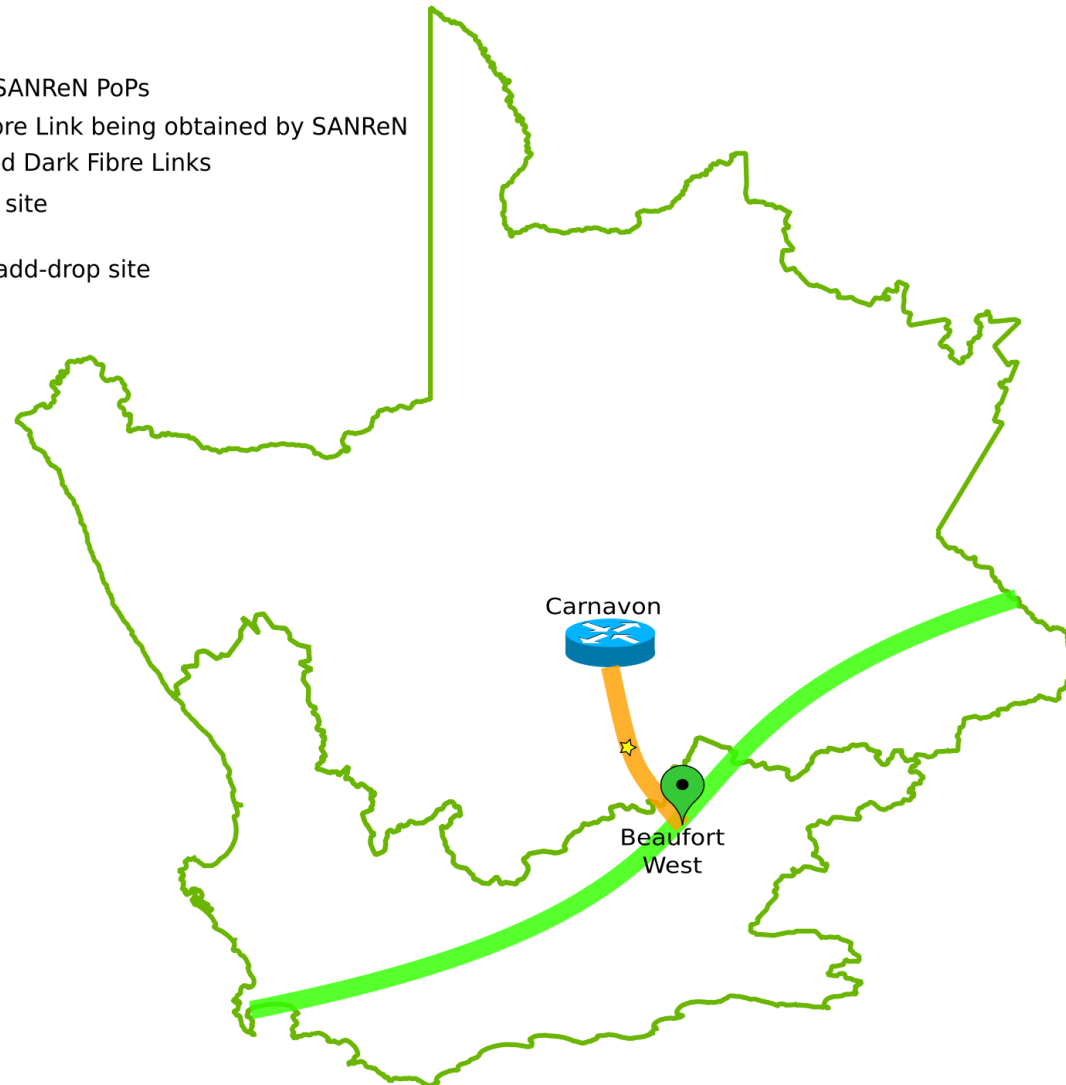


Project to build the required fibre infrastructure

- SARAO and SANReN embarked on a project to build dark fibre between Beaufort West and Carnarvon.
- SARAO is funding the project
- SANReN is managing the project on behalf of SARAO
- SANReN estimates that this project will be completed in 18 months.
- The DWDM System will be extended to the SKA Core Site and will immediately supply 100Gbps for the Meerkat project.
- The same DWDM system will cater for the connectivity requirements for SKA1.
- SKA1 has budgeted for fibre between Beaufort West and the SDP in Cape Town. SANReN's fibre network can act as a redundant route from Beaufort West to Cape Town.

SKA Dark Fibre Project

-  Existing SANReN PoPs
-  Future fibre Link being obtained by SANReN
-  Requested Dark Fibre Links
-  Repeater site
-  SANReN add-drop site

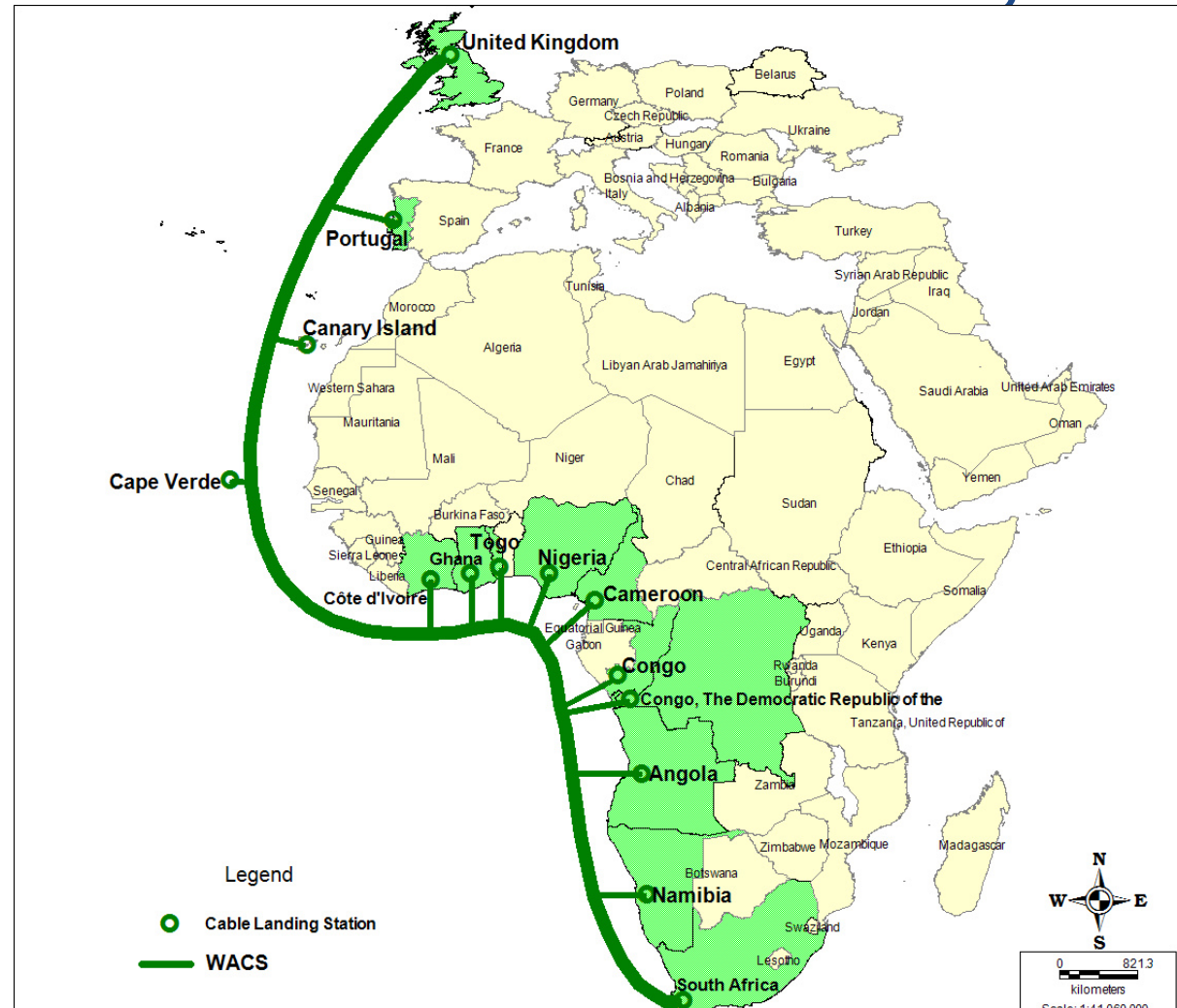




SANREN, SKA and WACS

- As mentioned previously, SANReN owns 7.3% of total available capacity WACS undersea cable.
- Currently, this translates to roughly 890 Gbps between Cape Town and London
- SANReN has participated in all of the WACS upgrades to date.
- The most recent upgrade has allowed SANReN access to 100Gbps wavelengths.
- The SA NREN has leveraged its WACS capacity to swap capacity with other undersea cable providers.
- Around 220 Gbps is currently activated between Cape Town and London
- 100 Gbps capacity has been set aside for the exclusive use of the SKA1 project.

MAP of the WACS Cable System

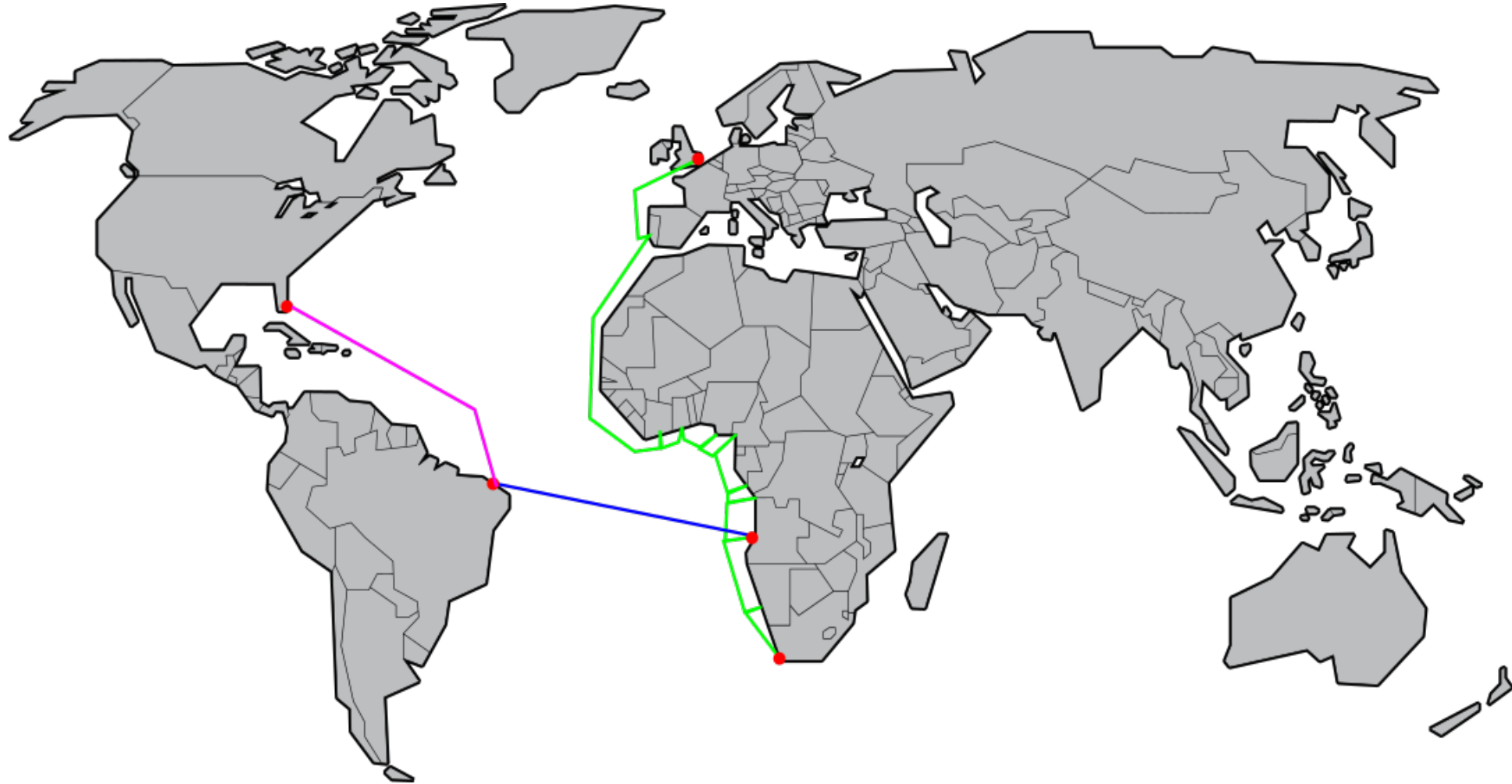




Leveraging WACS to support the global R&E Community writ large

- The AmLight Project has partnered with SANReN to build a 100Gbps link that connects:
 - Miami, USA
 - Fortaleza, Brazil and
 - Cape Town, South Africa
- SANReN leveraged its WACS capacity to provision a link from Sangano, Angola to Cape Town, South Africa
- The AmLight project build a link on the SACS cable to Angola.
- This potentially offers an alternate path to the SRC in Canada.
- Similar projects can be undertaken with other cable systems in the Indian ocean.

Amlight, SACS, WACS Project





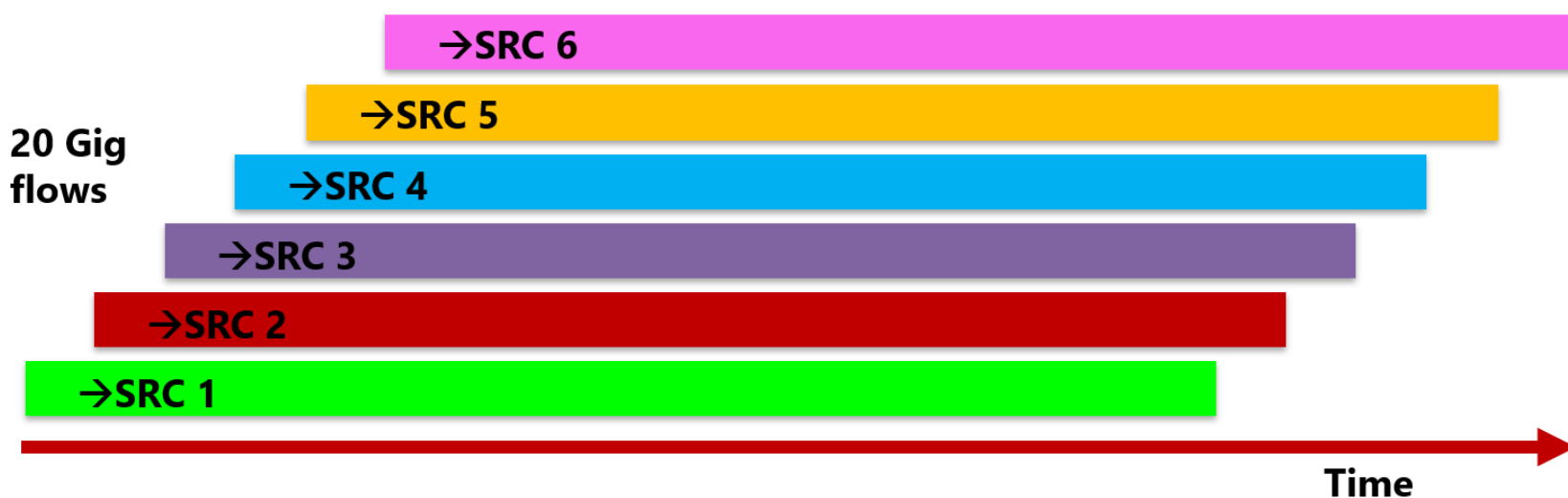
Task 4.3

D4.3 "Architecture and cost model for the European ESRC network"

D4.4 "Architecture and cost model for a World-wide network for SKA"

Richard Hughes-Jones

Models of Data Flows from the Telescopes to a SRC



- Data sent to each SRC in turn.
- Requires clean 100Gbit/s paths to each SRC.
- Worst case the SRC would require a 200 Gbits/s link.

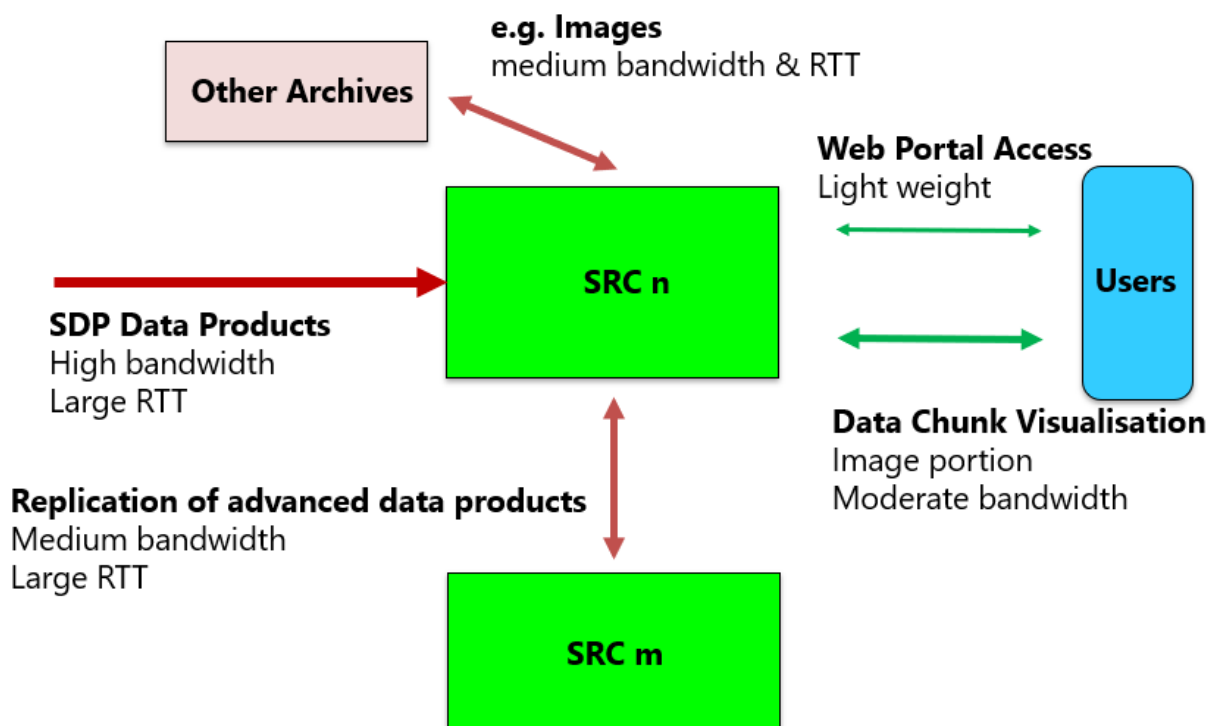
- Several flows to different SRC taking place in parallel.
- Operationally more realistic.
- Makes efficient use of the network to each SRC.
- WP4 demonstrated stable 28 Gbit/s flows – TCP limit.

The SDP push model gives the advantage of scheduling the use of the bandwidth on the telescope access link .



The European ESRC network

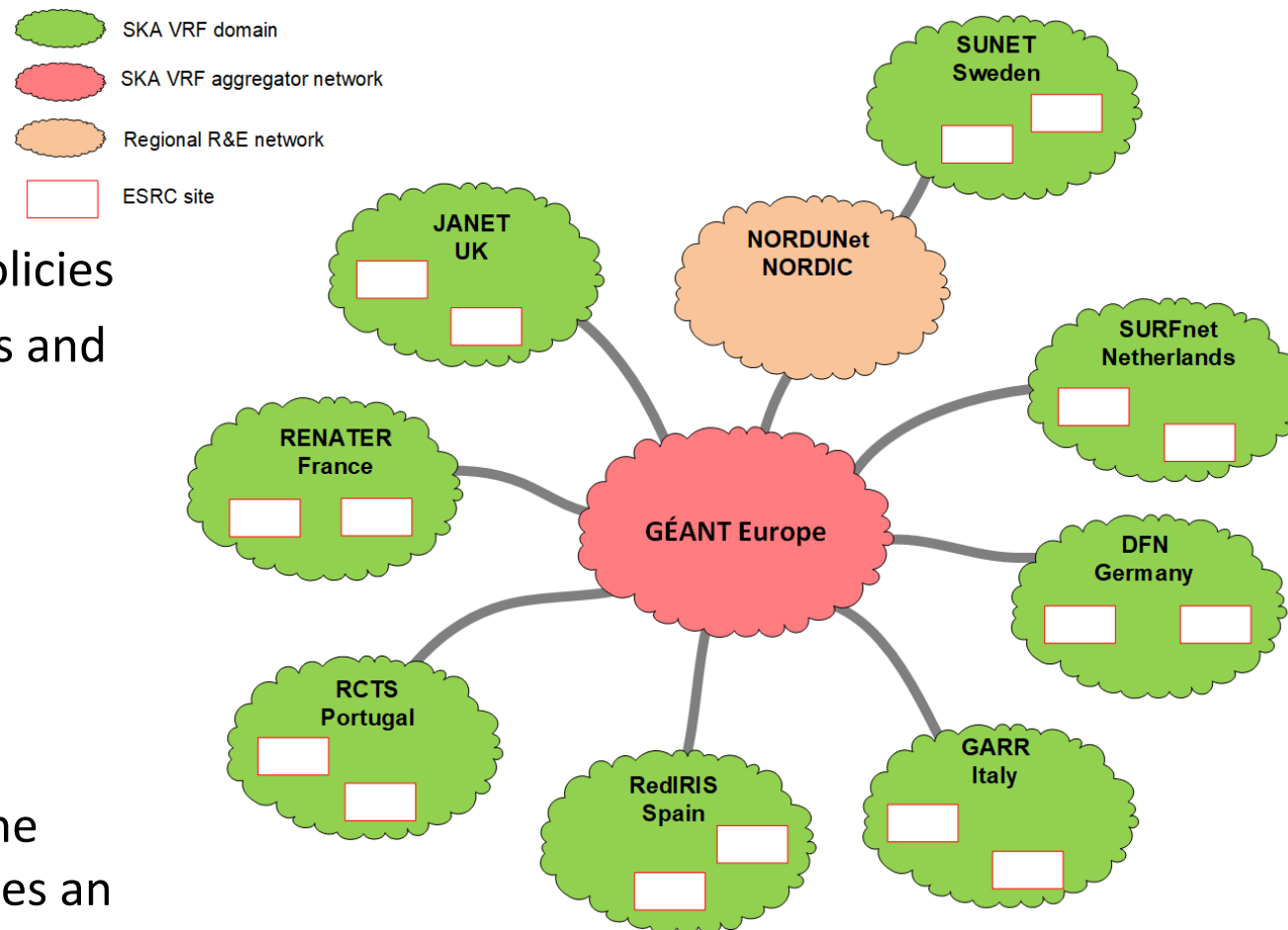
Network Traffic into & out of a SKA Regional Centre



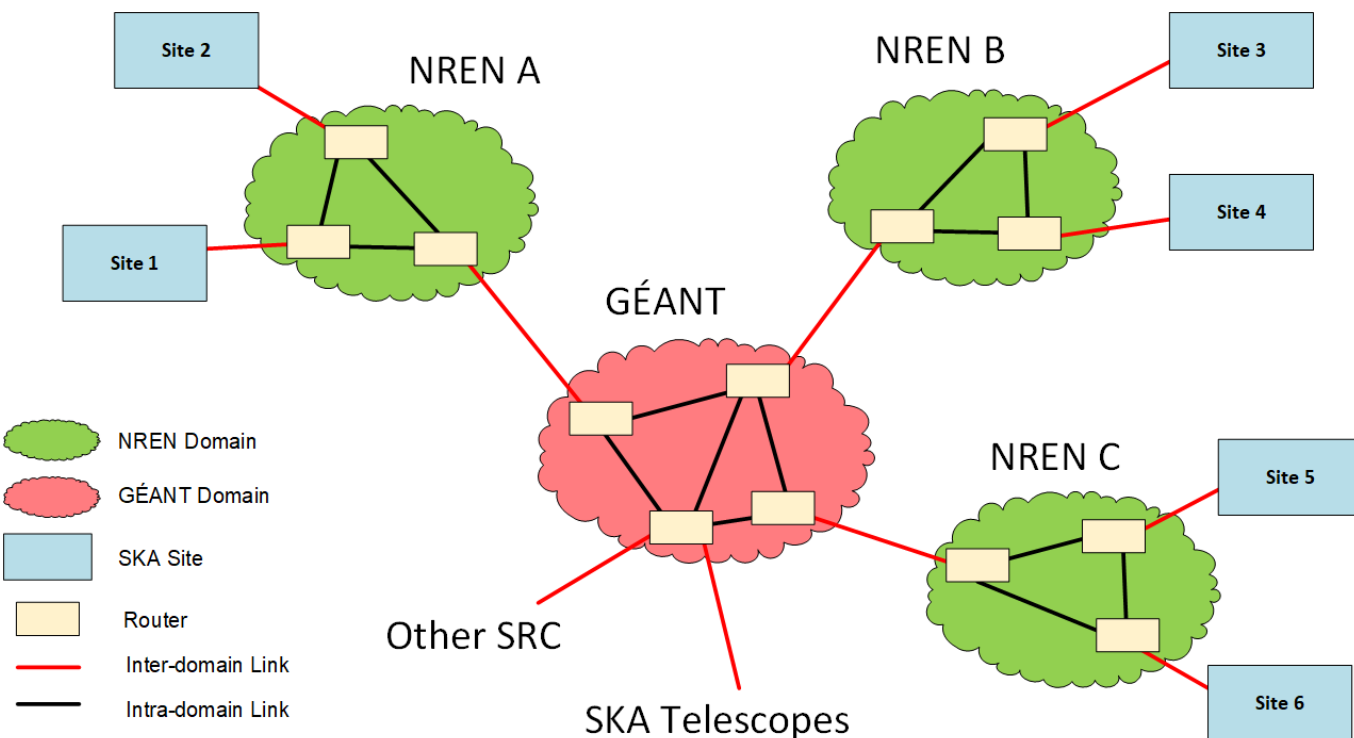
- Assumptions
- 20 Gbits/s at a time from each Telescope peak 40 Gbit/s to a site
- At a given time all of a Data Product goes to one site
- 10 Gigabits/s to other ESRC sites
- 10 Gigabits/s to non SKA data Archives
- 10 Gigabits/s to other RDC
- 70 Gigabit/s (peak) to each site.
- Flexible but secure ACL and high performance DMZ
- Say 100 Gigabit/s for SKA to each site in 2024/25. This is affordable.
(c.f. the UK WLCG sites which have 10-40 Gbit/s dedicated links now)

Network Architecture for the ESRC

- VRF based overlay on the academic networks
- Isolation of SKA traffic from other users
- Easier for NRENs to implement the routing and policies
- SKA traffic can be engineered to use specific paths and routes to provide the high bandwidth
- Layer 3 routing provides isolation of any network configuration issues and strictly limits broadcast storms
- Layer 3 will re-route traffic as long as there is an alternative network path
- Configuration actions have to be undertaken by the NREN and a Site to join the SKA VRF, which provides an extra layer of security



Forming the VRF - Connecting ESRC Sites to the NREN



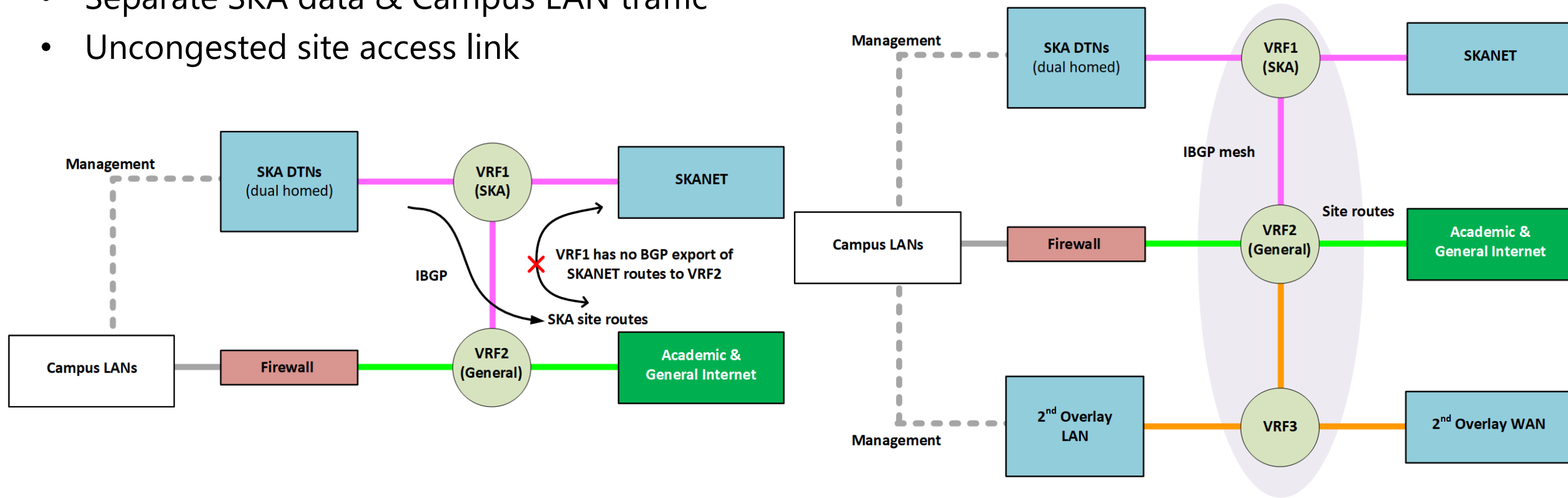
- Site routers connect to VRF in the NREN routers
- NREN backbone forms the VRF in that country.

Operational aspects:

- Use NREN access link to peer with routers on the GÉANT backbone
- Sites will need to define and implement local Site Policy and Filtering Requirements.
- Project-wide Acceptable Usage Policy to be defined by the whole SKA community.
- NRENs and GÉANT will need to implement an access policy based on the set of accepted ESRC site prefixes.

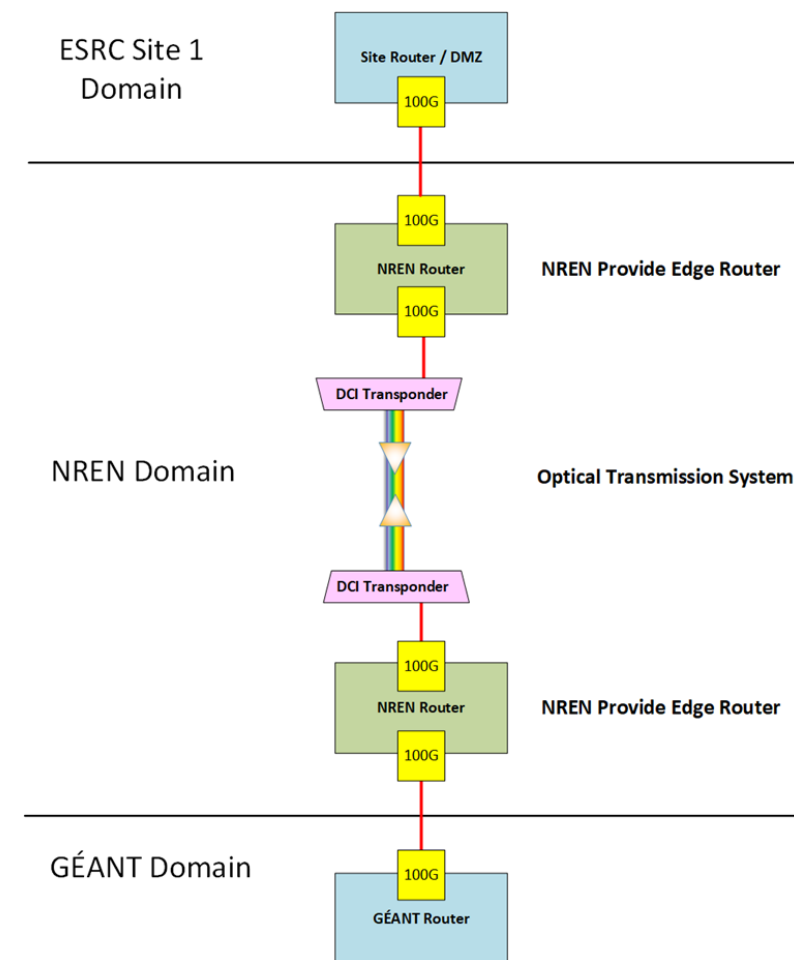
Network Considerations for a ESRC Site

- Need for high performance Data Transport Node hardware
 - Tuned for RTT ~300 ms
 - Network – disk transfer rate ~20 Gbit/s
- Flexible but secure ACLs and high performance DMZ connected to the VRF
- Separate SKA data & Campus LAN traffic
- Uncongested site access link



Network Cost Model for the ESRC

- The correct way: establish the sites and GÉANT would ask the NRENs for indicative costs & give help with contracts.
- For AENEAS GÉANT Product Management team used the equipment and technology used in the GÉANT upgrade (Summer 2019) and the current prices that have been procured for a 100Gbit/s slice.
- Assumed 16 sites – calculate for 3 scenarios:
- **Option A Layer 3 VPN Service Over the Shared IP links**
CAPex about €0.8M and the annual OPex about €0.25M
- **Option B Single Dedicated 100 Gbit/s link**
CAPex would be over €1.5M and the annual OPex about €0.5M
- **Option C Fully Resilient with two Dedicated 100 Gbit/s links**
CAPex would be over €4.5M and the annual OPex over €1.5M

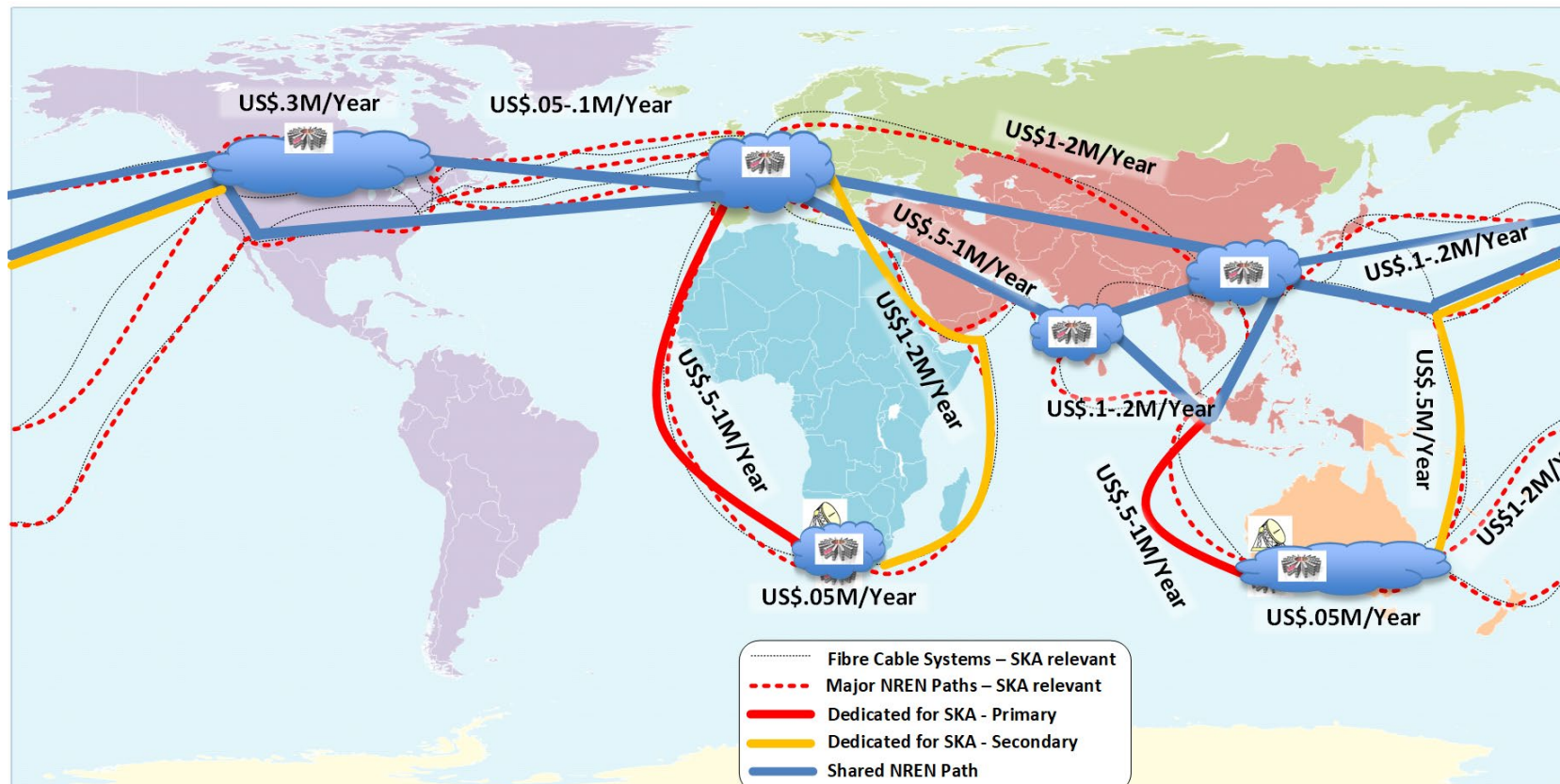




World-wide Network for SKA

Global Network & Paths of Interest to SKA

- Dedicated Primary (red lines) & Backup links (yellow lines) from both telescopes
- Use of the shared academic network (blue lines).
- 1 PetaByte/day pushed by SDP from each Telescope → 100 Gigabit/s
- Costs based on 10 to 15 year IRU per 100 Gbit circuit projected to 2020 prices



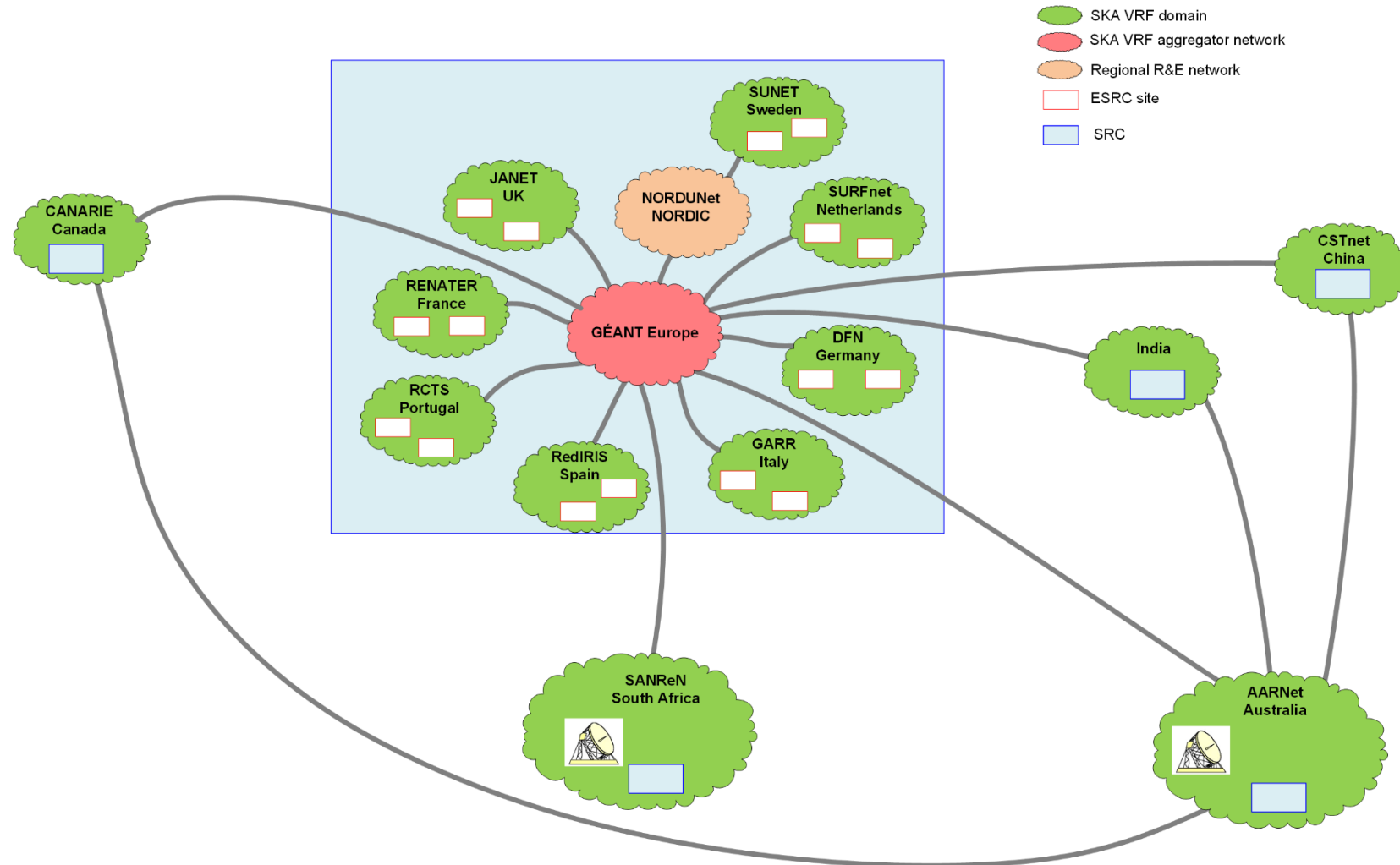
Budgetary OPex Costs

- Primary links USD 2M per year
- Backup links USD 4M per year

Now a new affordable path
Singapore – Europe direct

Global Network Architecture for SKA

- Global VRF based overlay with peering linked over the shared academic network

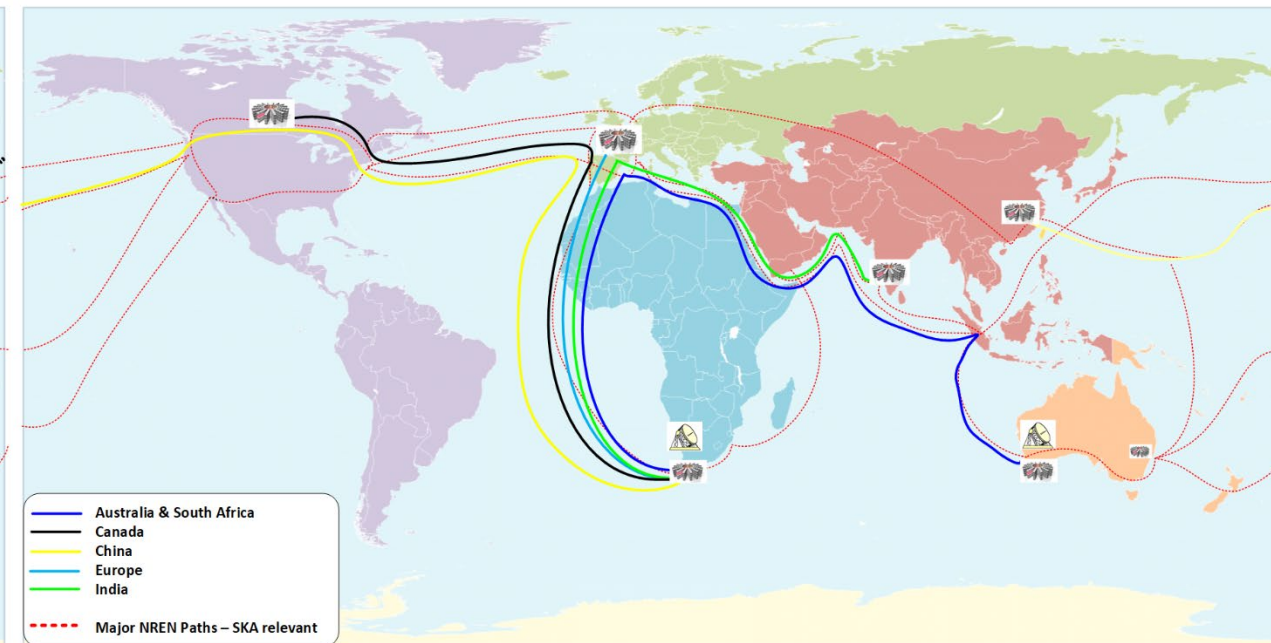
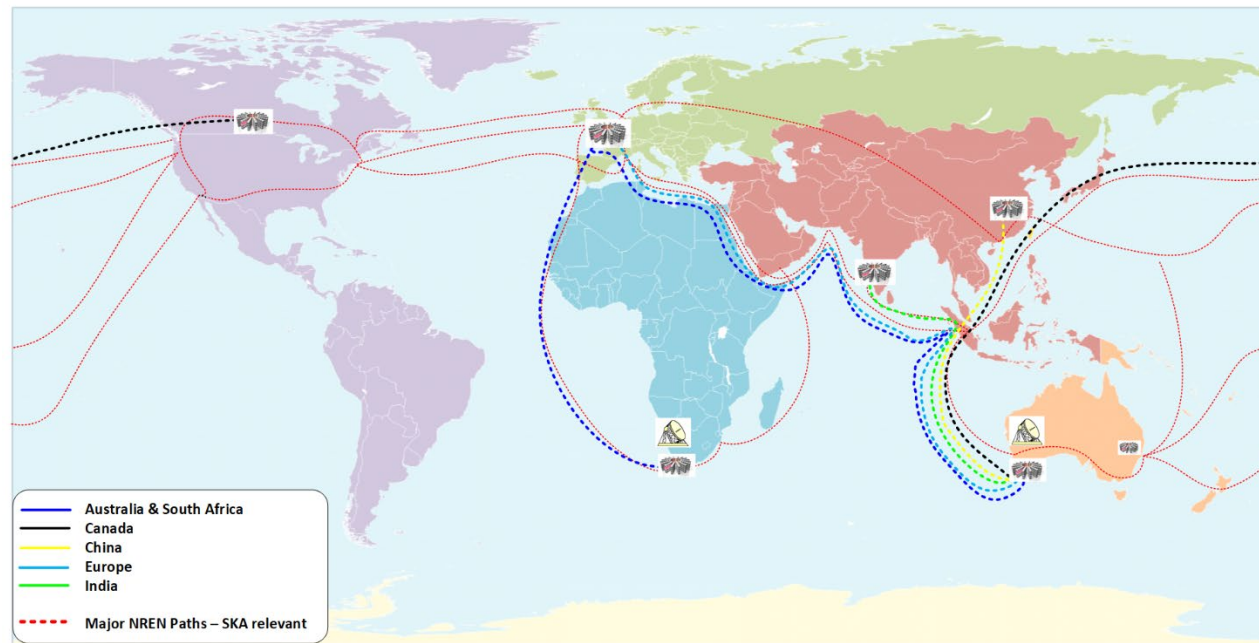


Global Paths of the Data Flows Pushed to the SRC – 1 Replica

- Five flows on the submarine cable from Perth to Singapore .
- Then join the general purpose routed IP academic network.
- Single flows on the routes to Canada, China and India, Australia is local, and two 20 Gbit/s flows would be carried to London to reach SRCs in Europe and South Africa.
- Five flows on the submarine cable from Cape Town to London.
- Then join the general purpose routed IP academic network.
- Different submarine cables used to reach India and Australia, Europe is local, and two 20 Gbit/s flows cross the Atlantic to SRC in Canada and China.

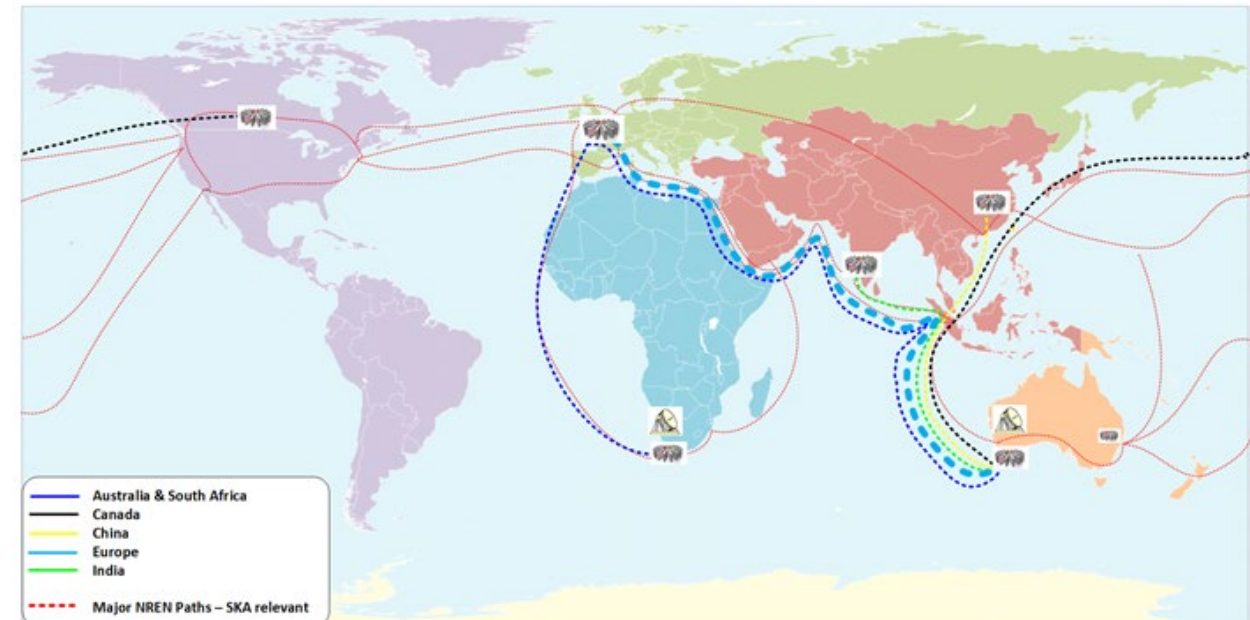
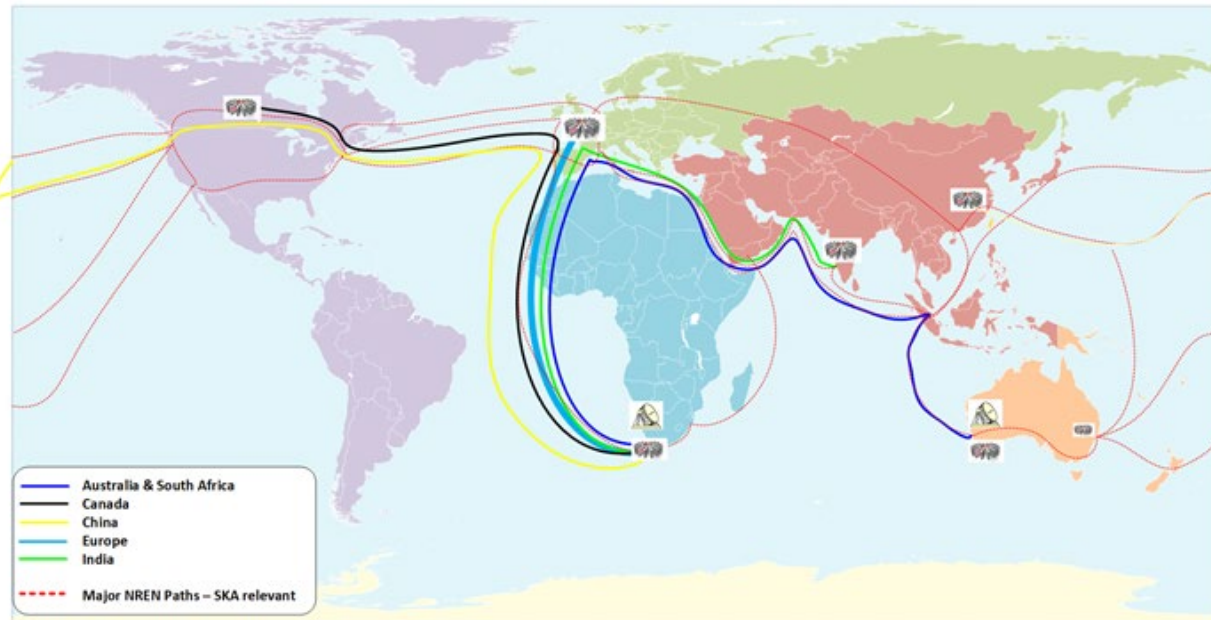
SKA1-LOW Australia

SKA1-MID South Africa



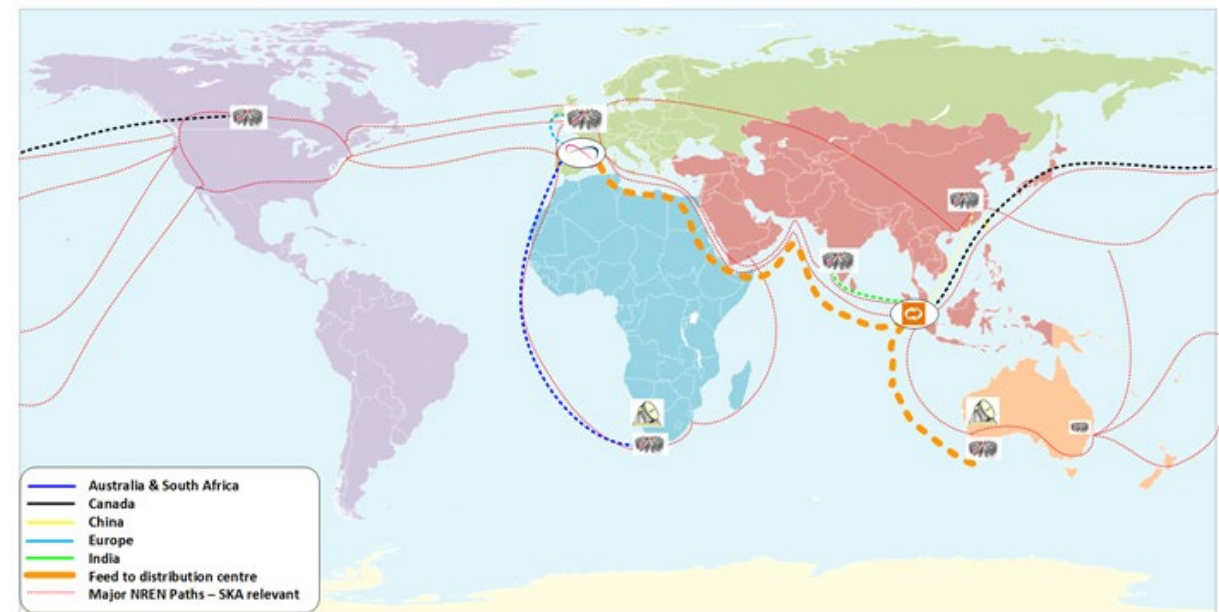
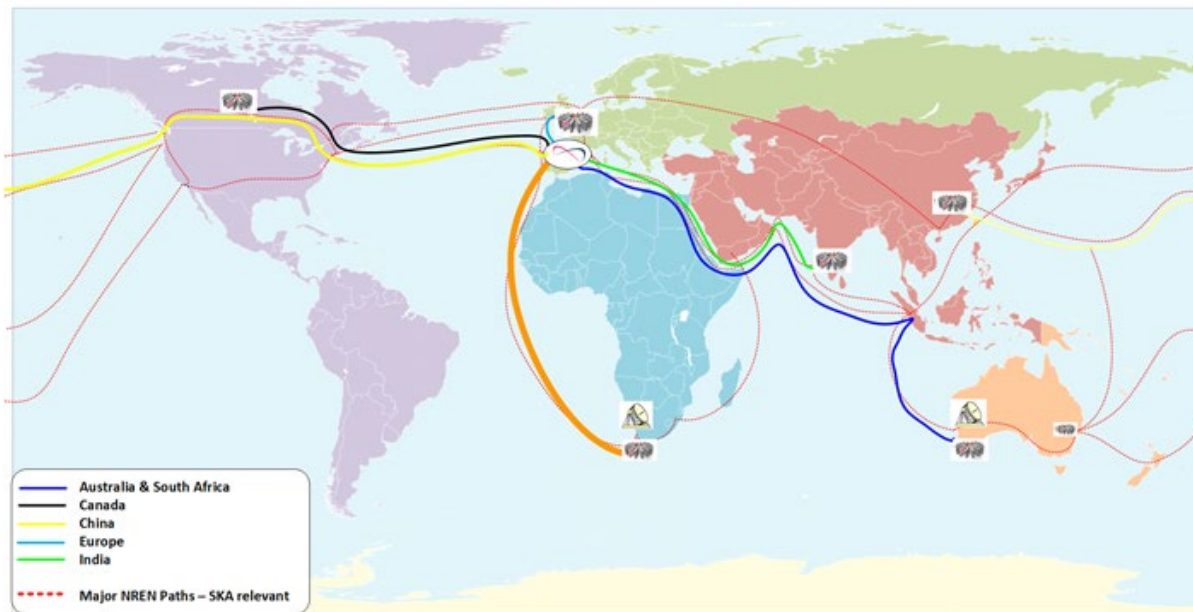
Global Data Flows if ESRC has a Full Copy of SKA Data

- One lambda carries the ~ 20 Gbit/s flows to the global SRCs like the 1 replica scenario
- Another 100 Gbit/s supplies the other Data Products to ESRC
- Europe funds extra intercontinental OPEX of USD 1.7M/year.
- Extra load on SDP buffer systems.
- Consideration to be given to the cost of provision of diverse backup paths.



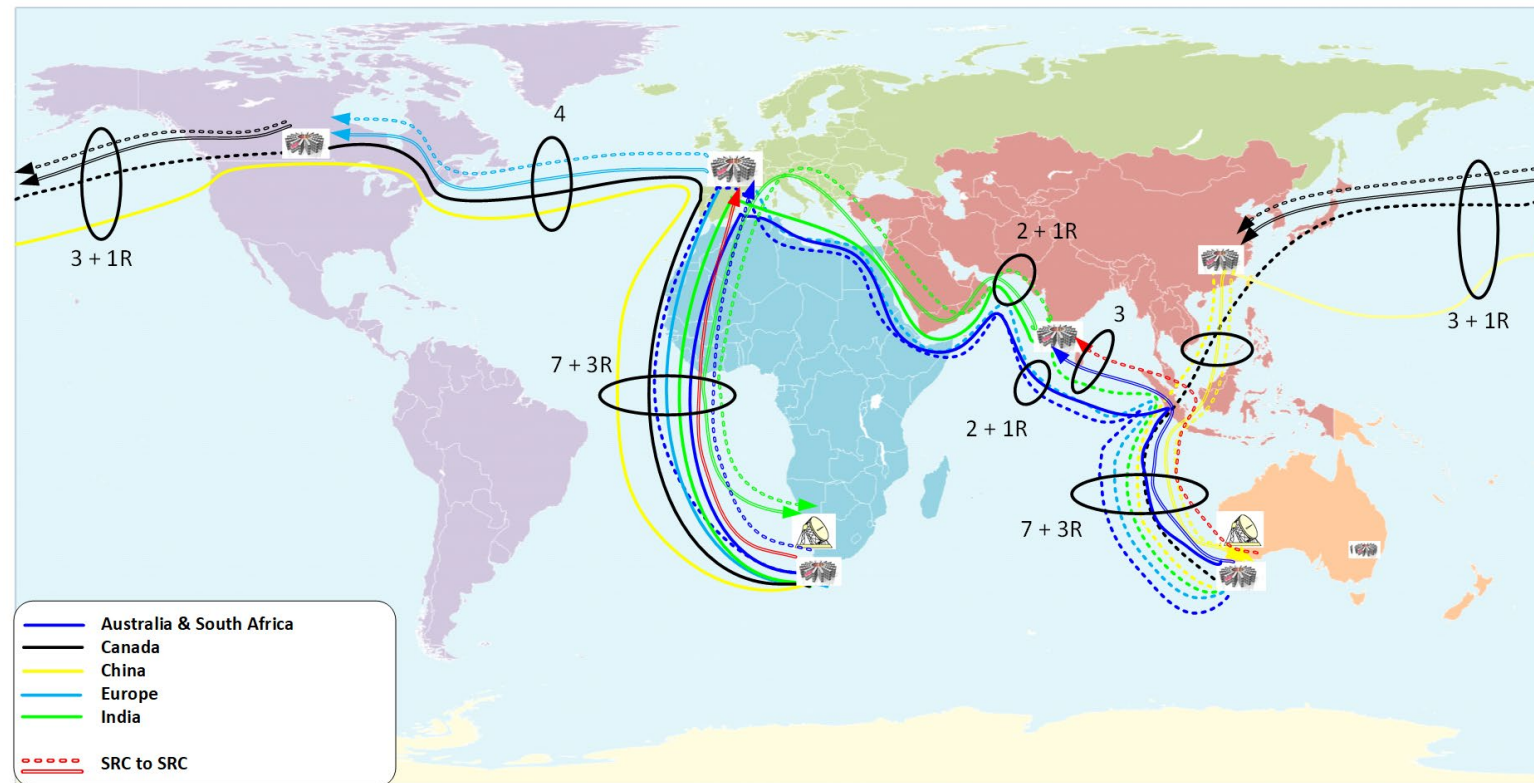
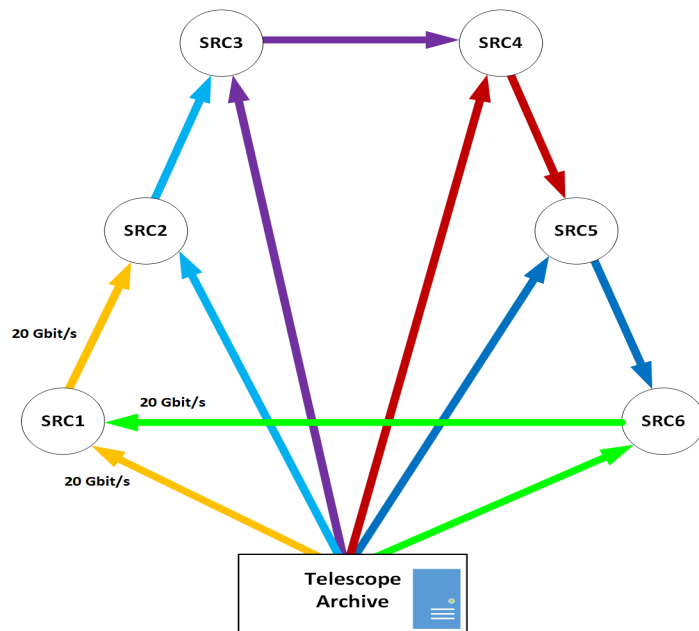
Distribution Centres Send Data to the SRCs

- Observatory Data Products sent to SKA Distribution Centres located at the remote ends of the expensive Submarine Cable links.
- These centres forward the data to the SRCs as required.
- London good location for SKA1-mid Singapore for SKA1-low
- Two days data from a telescope is ~ 2 PBytes of storage
- Need a medium size data centre, similar to a WLCG Tier 2 site
 - expensive 24/7 cost with power, staffing, etc. estimate ~ 0.3 M Euro/year/centre



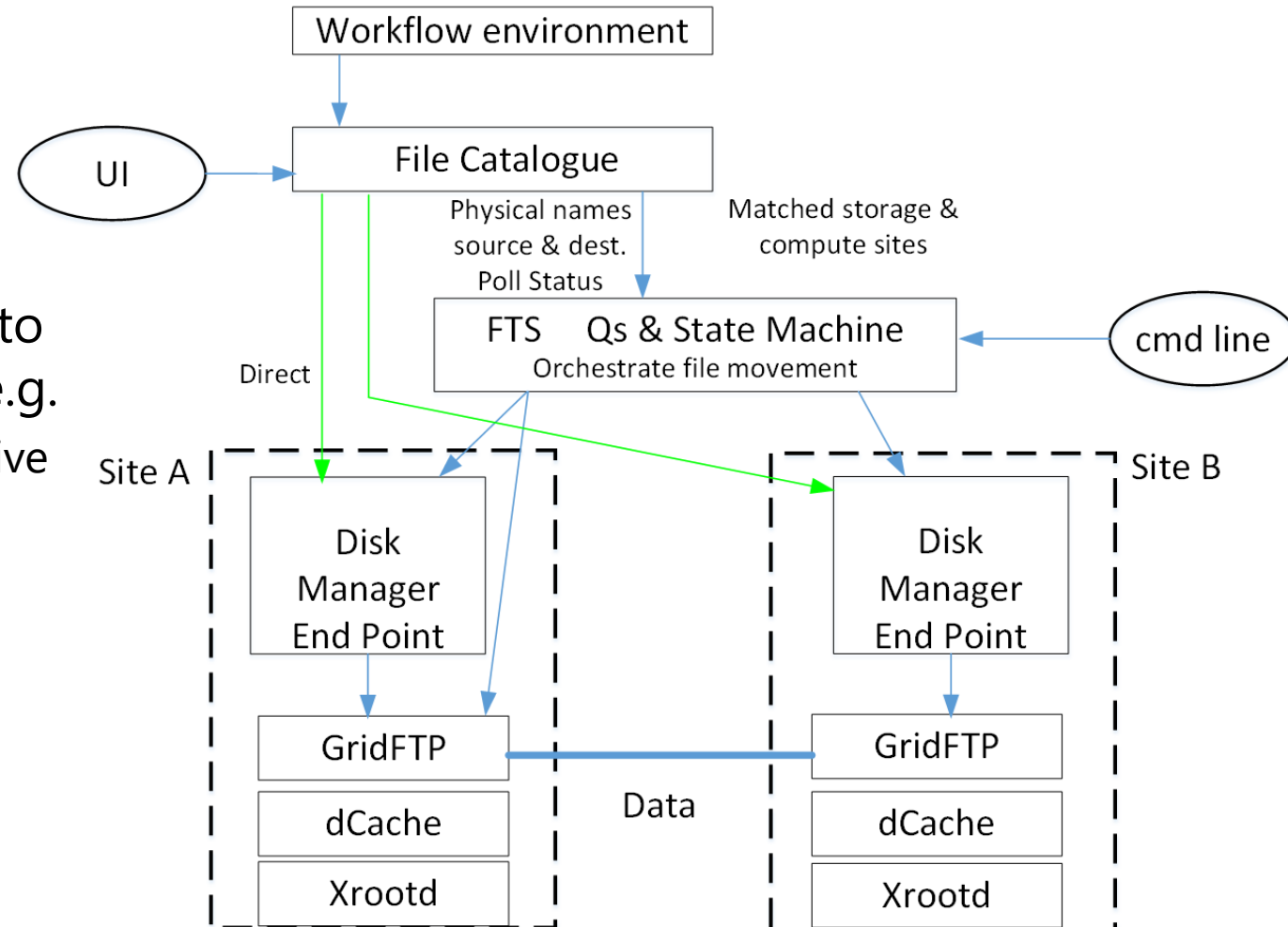
Global Data Flows if the SRC Re-distribute data – 2 Replicas

- Each SRC accepts its fraction of the Observatory Data Products and re-distributes to another SRC.
- SRC has 20 Gbit/s flow from the telescope & a second continuous 20 Gbit/s flow from another SRC.
- Each SRC sends out a 20 Gbit/s flow.
- Makes substantial use of the shared academic network which would imply charges to the SKA community.
- Cost to SKA community Very approx. ~ 0.8 M USD/year not allowing for the extra BW from the telescopes.



Integrating Global Distributed Computing Sites (AENEAS) & SDP

- Need for state of the art middleware for:
 - Replica management.
 - Orchestration of the data file streaming.
 - Workflow management
 - E.g. Rucio & FTS
- The SDP design needs to be integrated into the distributed computing environment, e.g.
 - Read-write loads on hot/cold buffers or archive
 - Linking SDP & global file catalogues
 - Defining the policy interface
- Realistic flow demonstrations are an excellent infrastructure candidate for SKA Data Challenges





Summary

WP4 Conclusions, Costs and Recommendations

Richard Hughes-Jones

Recommendations: The Global SKA Network

- The global SKA network be constructed as an overlay network formed from a set of VRFs.
- These VRFs are connected with suitable inter-continental paths to enable the peering.
- That SKAO funds the use of dedicated 100 Gbit/s paths out of each telescope on the submarine cable routes to the Northern hemisphere.
The budgetary OPex cost of these two primary paths is USD 2M per year projected to 2024 prices.
- These primary dedicated 100 Gbit/s paths out of each telescope be commissioned and in service to assist with:
 - The hardware and science commissioning of the telescopes.
 - The establishment of the SRCs and the SKA data challenges.
- The purchase of dedicated secondary 100 Gbit/s backup paths from each telescope be considered only when SKA is fully operational – probably about 2028.
- The SKA community gives careful consideration to the use of one replica of the Observatory Data Products external to the SDP archives.



Recommendations: ESRC Connectivity

- The SKA network in Europe be constructed as an overlay network formed from a set of VRFs each one covering a country and aggregated by a VRF on the GÉANT backbone.
- In discussion with the NRENs sites provision a suitable access link for SKA data
- ESRC sites and the Science Data Centres establish DMZs for SKA data transfers:
 - Ensuring a loss free end site network with no bottlenecks
 - Suitable access control policies for SKA data flows
 - Support for both high volume data transfers and public access to certain datasets
- The European SKA community make provision for funding the cost of the SKA VRF network estimated at
CAPex of about €50k and an annual OPex of about €15k for each 100 Gbit/s path over the network

Recommendations: Data Transfer

- ESRC sites provide a small number, e.g. ~4, of high performance Data Transfer Nodes for SKA located in the SKA VRF.
- The DTNs are tuned for high bandwidth long-haul data flows with RTT up to 300ms.
- Suitable open source data transfer applications are available at sites
 - Have on-site operational experience & support & widely used by the science community
 - Use open protocols e.g. Webdav/http(s) - Allow Third Party Copy
 - Integrated with global AAI e.g. use of tokens
 - Current example is Xrootd
- The SDP is integrated with the Global Distributed Computing Sites (AENEAS & others)
 - Replica management & Orchestration of data file streaming.
 - Linking SDP and global catalogues
- Start with Rucio & FTS
- Realistic flow demonstrations are an excellent infrastructure candidate for SKA Data Challenges

Recommendations: General Implementation & Operation

- That GÉANT works with the SKA community during the implementation phase of the ESRC
 - To assist in determining the sites, the bandwidth and other network requirements.
 - Coordinate requests for information and costs to the NRENs
- That GÉANT & global NRENs work with the SKAO / SRCSC to form a SKA-NREN forum.
 - The forum should cover technical, governance and operational aspects
 - Include European and global astronomy networking & infrastructure members.
 - The forum should enable the exchanges of roadmaps, requirements, and facilitate operational aspects.
 - Worked very well for VLBI and WLCG



Future Work

- Continue investigating the performance of long-haul applications & protocols.
- Extend the tests to include transfers to Data Centres and storage sub-systems e.g ceph & erasure coding systems.
- Support the work in ESCAPE (e.g. WP2 e-infrastructure).
 - Testing sustained data transfers
 - Comparison of low level performance with that from the Rucio level
- Collaborate and assist with the work of the SRCSC.
 - Involvement in technical working groups

