DMC20

LESSONS AND OBSERVATIONS

ANDREW HOWARD

CHIEF JUDGE DMC19 & DMC20

CLOUD TEAM MANAGER, NCI

NCI AUSTRALIA

OVERVIEW….

- Introduction to DMC
- Network architecture
- The Challenge
- Issues
- Lessons
- Conclusion
- Questions

DATA MOVER CHALLENGE

- How do we operate and manage regional and global platforms which spans multiple administrative domains, supports high speed data transfers of large data sets and is able to effectively utilise 100G networks ?
- How do we encourage the development of innovative data transfer tools and techniques which are able to exploit contemporary hardware and networks efficiently ?
- Can we create a repository for containerised data transfer applications which can be utilised by the community ?

DATA MOVER CHALLENGE - META CHALLENGE

- How do we build the skills, knowledge and trust between the current and next generation NREN and HPC facility network engineers ?
- How do we exercise advanced network capabilities, understanding baseline performance, consequences of load, abnormalities etc.
- What information can we return to the community based on the lessons learned.

DMC SCENARIOS

- Surfnet -> simultaneous transfer to NSCC (via CAE-1-SingAREN) and AARNet (via CAE-1--Indigo)
    - (Participants can do staging or concurrent transfers)
- AARNet -> Surfnet (via Indigo-CAE-1-GEANT)
- NSCC -> StarLight (via SingAREN-I2-PacificWave-StarLight)
- NII -> NSCC (via SINET-US-NetherLight-GEANT-CAE-1-SingAREN)
- NICT -> KISTI (via NICT/JGN-SingAREN-Internet2-StarLight-Kreonet)

STRESS TESTING NEW NETWORK SERVICES

- CAE-1 (Europe to Singapore)
- Indigo (Singapore to Western Australia)

CHALLENGE CONSTRAINTS

- Resources contributed by a number of partners
  - No root access (node operators execute privilege commands)
- Tools must be containerised
- Non homogeneous hardware, storage, network cards
- Must use distribution release kernel

NETWORK PRESENTATION - THE DREAM

- Single extended VLAN
  - Isolate experimental from production
  - Provision over backup paths to reduce any impact on production traffic
  - “The way we’ve always done it”
  - “Reduced complexity”

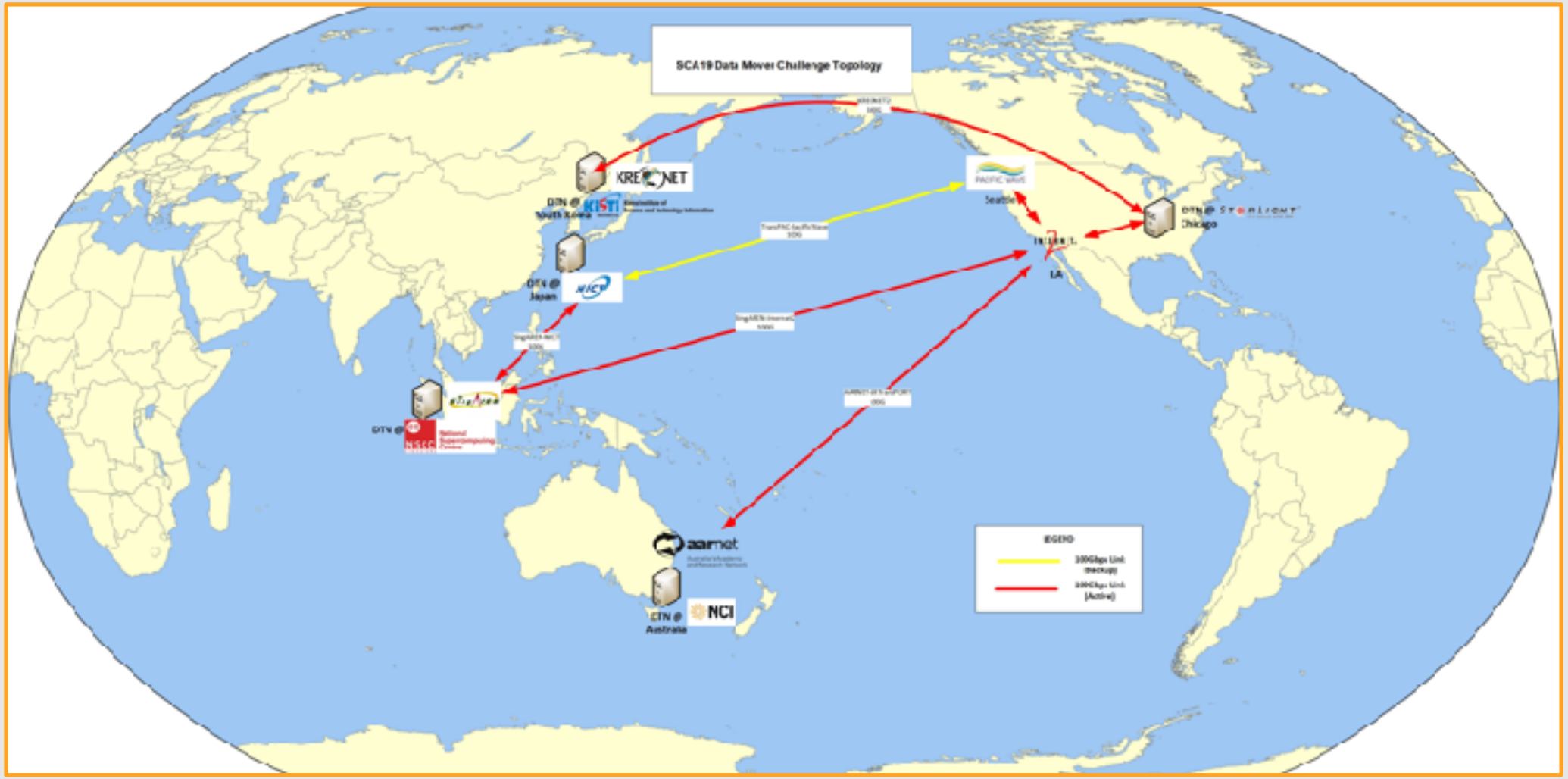
AND THE REALITY

- Single extended VLAN network
  - Significant stability issues
  - Difficult to effectively monitor or debug performance issues
    - Magic black box which had variable performance, latency and reachability
    - Requirement to know about looking glass, perfSonar and other measurement services to observe physical network load possibly correlating to VLAN performance
  - Significant complexity increase with number of networks, sites and systems
  - Requirement to tune for both outer (physical) and inner (VLAN) network environments
- * Shout out to Richard Hughes-Jones and the GEANT network team for accurately predicting many of the issues encountered

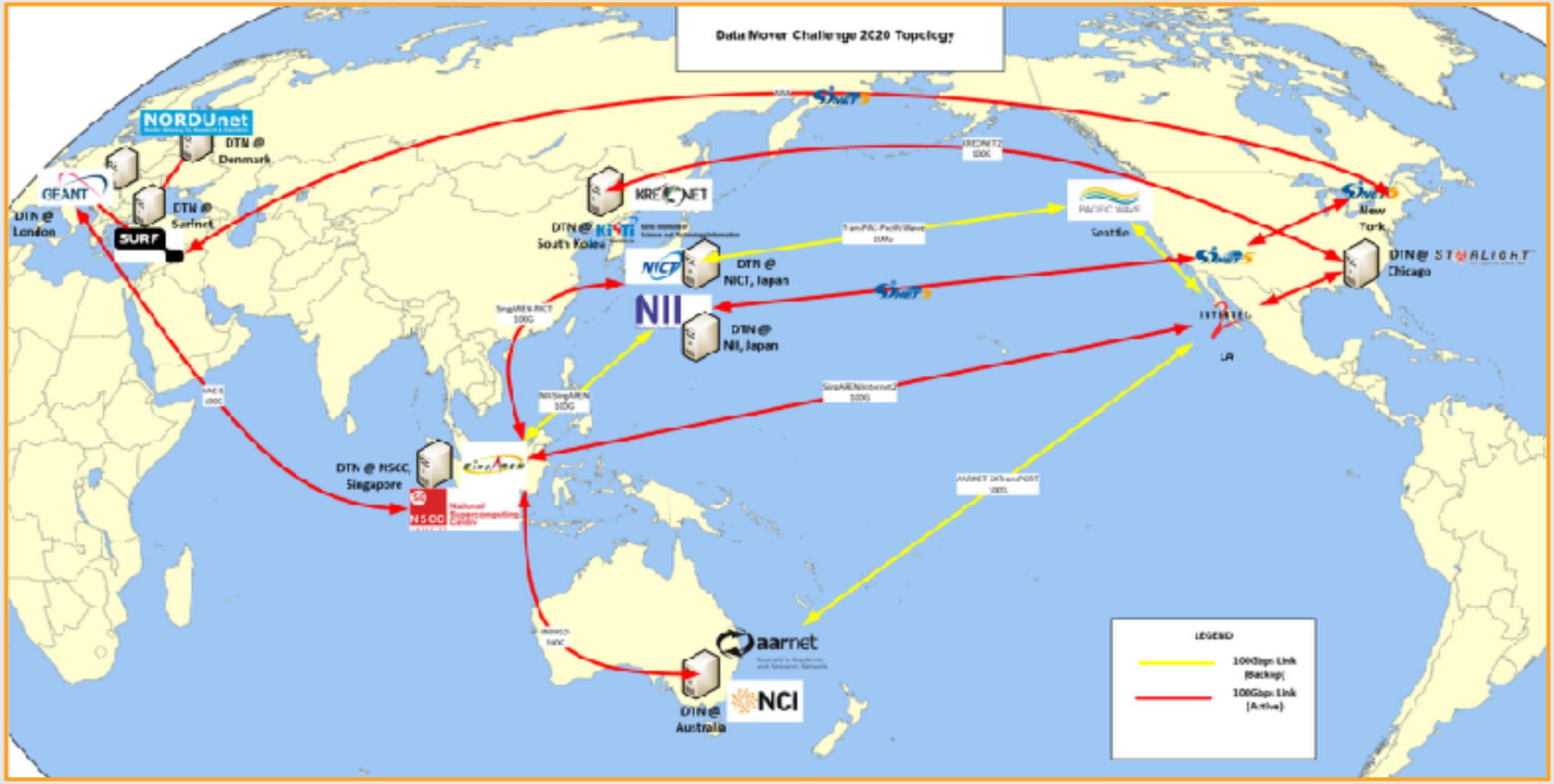
SOME OF THE ISSUES WE OBSERVED

- ARP table timeout
  - Solution: regular ping over VLAN
- “Routing on a stick (one arm routing) where 2 sites were connected to same switch but routing traffic over the same uplink to router. Having ip redirect enabled on router causes cpu to increase substantially and high retransmits on iperf3 tests”
- Previously performant network tuning caused window size issues

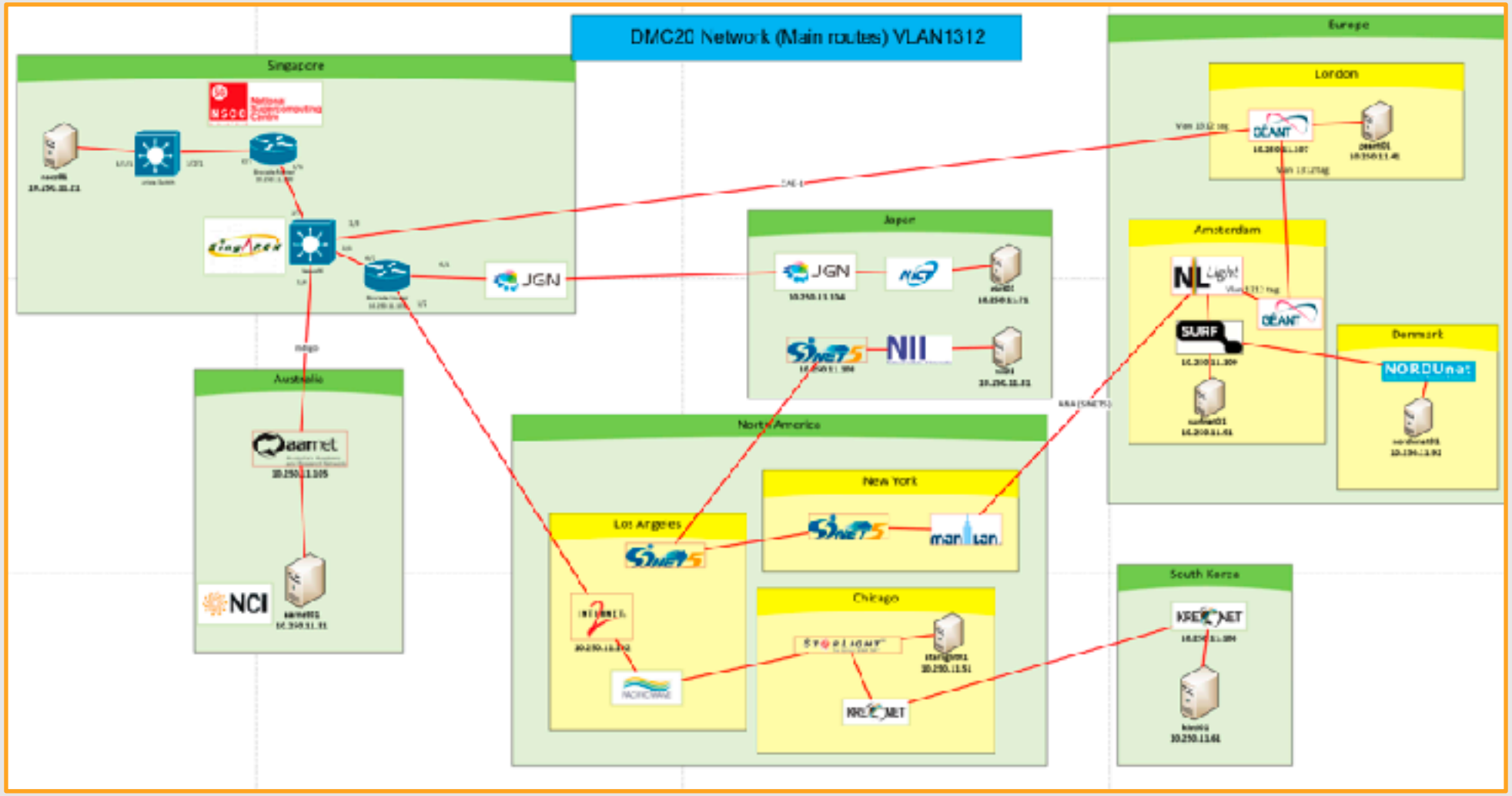
```
[ ID] Interval           Transfer     Bandwidth       Retr  Cwnd
[  4]   0.00-1.00   sec  86.0 MBytes   721 Mbits/sec    0   24.9 MBytes
[  4]   1.00-2.00   sec  1.49 GBytes  12.8 Gbits/sec    0   8.74 KBytes
[  4]   2.00-3.00   sec  0.00 Bytes    0.00 bits/sec    0   8.74 KBytes
```



DMC19 TOPOLOGY



DMC20 TOPOLOGY



DMC20 ROUTING PATHS

NETWORK TUNING

- TCP's max window size is 1,073,725,440 ((2^16-1)*(2^14))
- RX and TX Pause on interfaces
- Router buffer queue depth (Brocade)

```
qos queue-type 0 max-queue-size 65536
```

- Recommended reading:
  - https://fasterdata.es.net/assets/Papers-and-Publications/100G-Tuning-TechEx2016.tierney.pdf
    - However: the values Brian suggested result in unpredictable window behaviour

CENTOS7 SYSTEM TUNING DEFAULTS

- net.core.rmem_max=212992
- net.core.wmem_max=212992
- net.ipv4.tcp_mem=185259 247015 370518
- net.ipv4.tcp_rmem=4096 87380 6291456
- net.ipv4.tcp_wmem=4096 16384 4194304

FASTERDATA SYSTEM TUNING RECOMMENDATIONS

- # allow testing with buffers up to 128MB
- net.core.rmem_max=134217728
- net.core.wmem_max=134217728
- # increase Linux autotuning TCP buffer limit to 64MB
- net.ipv4.tcp_rmem="4096 87380 67108864"
- net.ipv4.tcp_wmem="4096 65536 67108864"
- # recommended default congestion control is htcp
- net.ipv4.tcp_congestion_control=htcp
- # recommended for hosts with jumbo frames enabled
- net.ipv4.tcp_mtu_probing=1
- # recommended for CentOS7+/Debian8+ hosts
- net.core.default_qdisc=fq

DMC SYSTEM TUNING RECOMMENDATIONS

- net.core.rmem_max=1073741824
- net.core.wmem_max=1073741824
- net.ipv4.tcp_rmem="4096 87380 536870912"
- net.ipv4.tcp_wmem="4096 87380 536870912”
- # recommended default congestion control is htcp
- net.ipv4.tcp_congestion_control=htcp
- # recommended for hosts with jumbo frames enabled
- net.ipv4.tcp_mtu_probing=1
- # recommended for CentOS7+/Debian8+ hosts
- net.core.default_qdisc=fq

RELIABLE SYSTEM TUNING RECOMMENDATIONS

- net.core.rmem_max=1073741824
- net.core.wmem_max=1073741824
- net.ipv4.tcp_rmem="4096 87380 268435456"
- net.ipv4.tcp_wmem="4096 87380 268435456”

- Further testing with > 2 nodes required to validate

SERVER TUNING

- System architectures
  - Memory
  - Dual Socket
  - NUMA affinity
- TCP Window Size
- 100G and 10G coexistence

STORAGE

- Variety of storage types used
  - NVMe
  - SSD
  - HDD
- Filesystems
  - Isolated to DTN local for security but not indicative of HPC centre parallel storage systems (Lustre, GPFS)

FUTURE DIRECTIONS

- Run the challenge over standard network offerings
  - Production equipment
  - Vendor support
  - Production stability
- IPV6
- Non Intel architectures

THANKS

- NRENS
- Site operators
- Network engineers and supporting staff
- Participating teams
- Judging Panel
- NSCC

Andrew Howard – *Chief Judge*	Cloud Systems Manager, National Computational Infrastructure (NCI)
Cees de Laat	Professor, Informatics Institute, Faculty of Science, University of Amsterdam
Eric Pouyoul	Network Engineer, Energy Science Network (ESnet)
Francis Lee Bu Sung	Associate Professor, School of Computer Science and Engineering, Nanyang Technological University (NTU)
Lawrence Wong	Emeritus Professor, Department of Electrical & Computer Engineering, National University of Singapore (NUS)
Tim Chown	Network Development Manager, Jisc


SCAsia 2020
Supercomputing
Gathering the Best of HPC in Asia
SG NSCC SINGAPORE
Data Mover Challenge 2020
Most Systemic Approach Award
Gauss Centre for Supercomputing (GCS) e.V.
SCAsia 2020
Supercomputing
Gathering the Best of HPC in Asia
SG NSCC SINGAPORE
Data Mover Challenge 2020
Best Speed & Science Integration Award
iCAIR / Starlight
SCAsia 2020
Supercomputing
Gathering the Best of HPC in Asia
SG NSCC SINGAPORE
Data Mover Challenge 2020
Experimental Excellence Award
Japan Aerospace Exploration Agency (JAXA) / National Institute of Information and Communications Technology (NICT)
SCAsia 2020
Supercomputing
Gathering the Best of HPC in Asia
SG NSCC SINGAPORE
Data Mover Challenge 2020
Most Innovative & Novelty Award
National Institute of Informatics (NII)

LESSONS

- More network and baseline system testing is required to ensure stability and fairness for contestants
- Extended VLANs are not sustainable or stable at this extent
- Contemporary troubleshooting skills exist at the physical network level but are stretched in a virtual context
- Logistics challenges
  - Timezones
  - Languages
  - Normal operational issues (circuit faults etc)

QUESTIONS ?