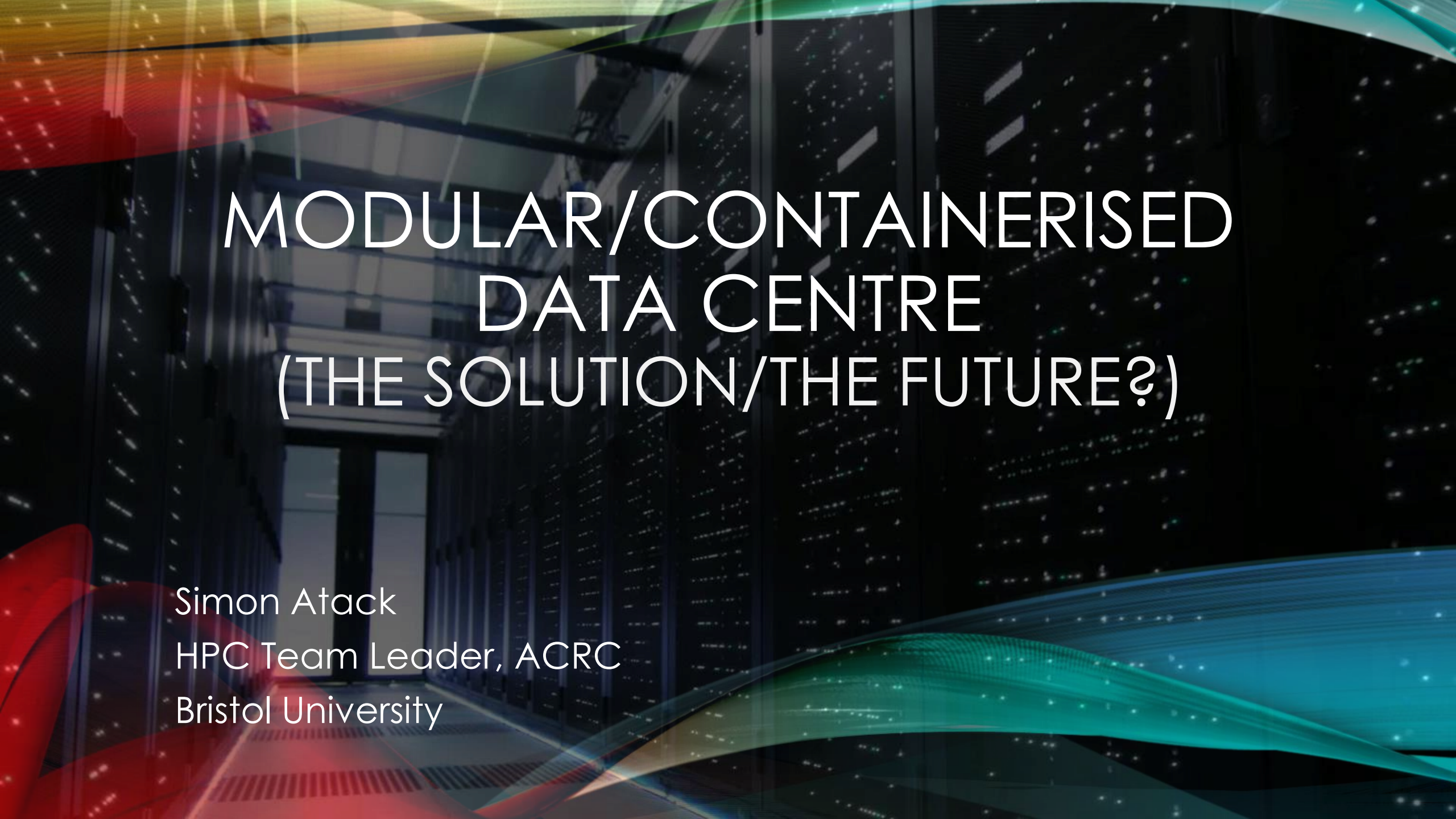


MODULAR/CONTAINERISED DATA CENTRE (THE SOLUTION/THE FUTURE?)

Simon Atack
HPC Team Leader, ACRC
Bristol University

EXISTING DATA CENTRES

- Aging DCs (Many DCs are now >10years old)
  - Tired Equipment
    - Time for a refresh
  - High PUE
  - Deeper Equipment >1000mm

- HPC has changed
  - Ever High Wattage Chips
    - >500W GPUs now, >1kW future?
  - Liquid Cooling may be only practical way to manage power/heat
    - Water has heat transfer coefficient 23.5 times that of air
    - Performance of air cooled may be significantly lower going forward
    - Density/Energy Efficiency –Fans use lots of Power to get enough airflow

BENEFITS

- Deployment time RAPID
  - Limited support needed from 'local facilities'
  - Can effectively be outsourced to specialists
    - Only Needs Power, Network & Pads
- Variable size SCALABLE
  - Can be tuned to what you need now
  - Future expansion is another unit
- Customised to your needs
- Modern
  - Low PUE – High Efficiency
- Site Placement opportunities
- Cost

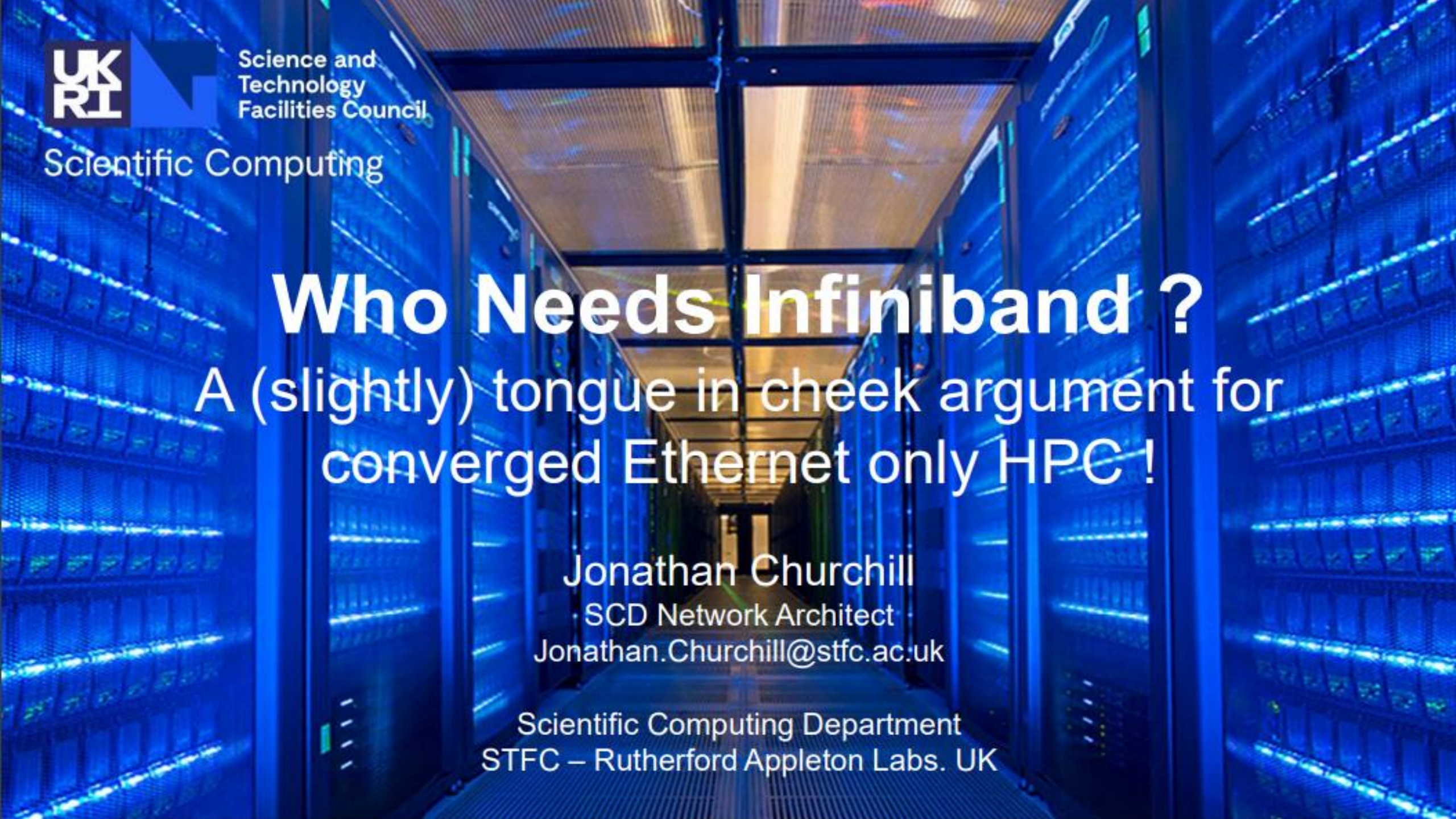
BRISTOL

- Our next DC will be modular/containerised solution
- Experimenting to see if this is how we will go for most future use
- Existing facility
  - Half got refreshed 5 years ago/Half >1years old
  - Design density of 20kW a rack
  - Pushed to 28kW non redundant at the minute
  - Very disruptive - Significant Shutdown would be required to upgrade
    - Currently Latest Cluster opporational in location – what to do with it
  - Equipment size -> Layout change likely needed
    - 20% deeper racks require 40% more volume
- Energy Efficiency
- Reuse of Heat/Energy

CONCLUSION

- It might now be time to evaluate what will be needed for the MID to LONG term
- A data centre is not a SHORT term investment.
  - It's a multimillion £ Investment
- Modular/containerised solutions could provide the scalable & flexible answer for the future

Simon Atack
HPC Team Leader
Simon.atack@bristol.ac.uk



Science and Technology Facilities Council

Scientific Computing

Who Needs Infiniband ?

A (slightly) tongue in cheek argument for converged Ethernet only HPC !

Jonathan Churchill

SCD Network Architect

Jonathan.Churchill@stfc.ac.uk

Scientific Computing Department

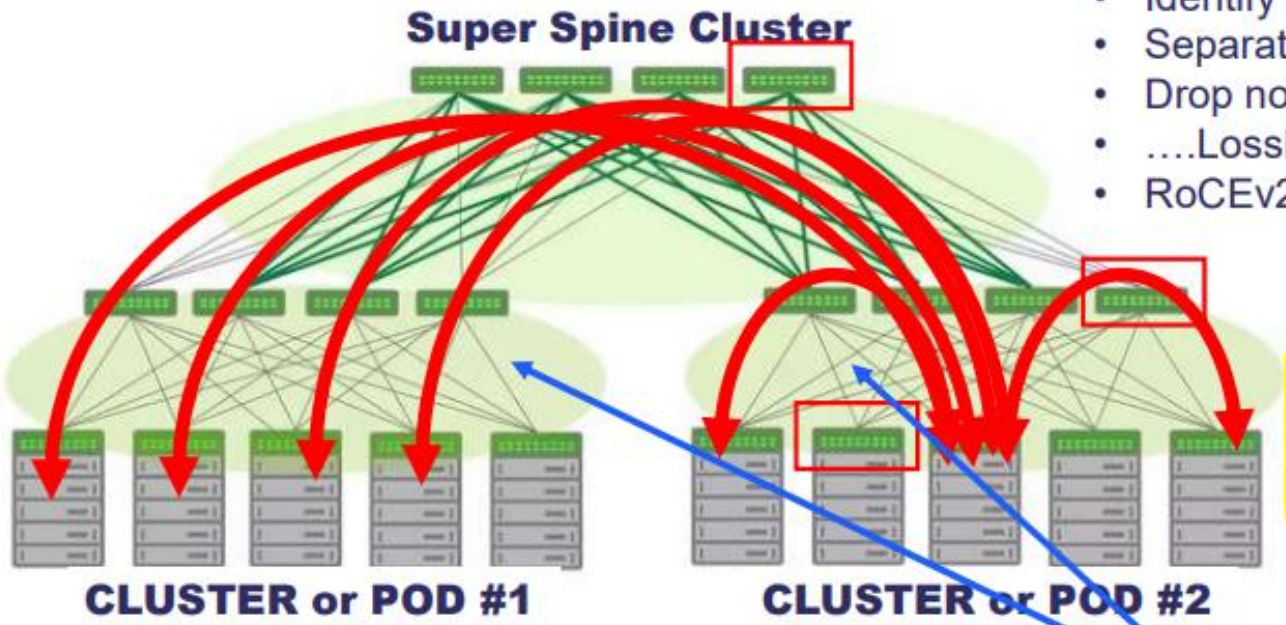
STFC – Rutherford Appleton Labs. UK

What is Converged Ethernet ?

Sample Topology

Congestion Control : RDMA & RoCE

- "RDMA over Converged Ethernet"
- RDMA Traffic mixed with general Ethernet traffic.
- Identify RoCE packets
- Separate PFC buffers for RoCE and non RoCE
- Drop non RoCE packets as buffers fill
- ….Lossless RDMA/RoCE traffic
- RoCEv2 = RoCE over L3 (aka L3 CLOS networks)



- w/o RoCEv2 JASMIN Ping-Pong 8-11uS
- w RoCE (estimated) Ping-Pong ~3uS
- c.f. std latency ethernet 100-200uS

This is ~ Infiniband PingPong Latencies
Plus Non-blocking Bandwidth

Science and Technology Facilities Council
Scientific Computing

https://enterprise-support.nvidia.com/s/article/understanding-rocev2-congestion-management
Onyx RoCE : https://docs.nvidia.com/networking/pages/viewpage.action?pageId=19807245
RoCEv2 : https://enterprise-support.nvidia.com/s/article/roce-v2-considerations

Pros and Cons

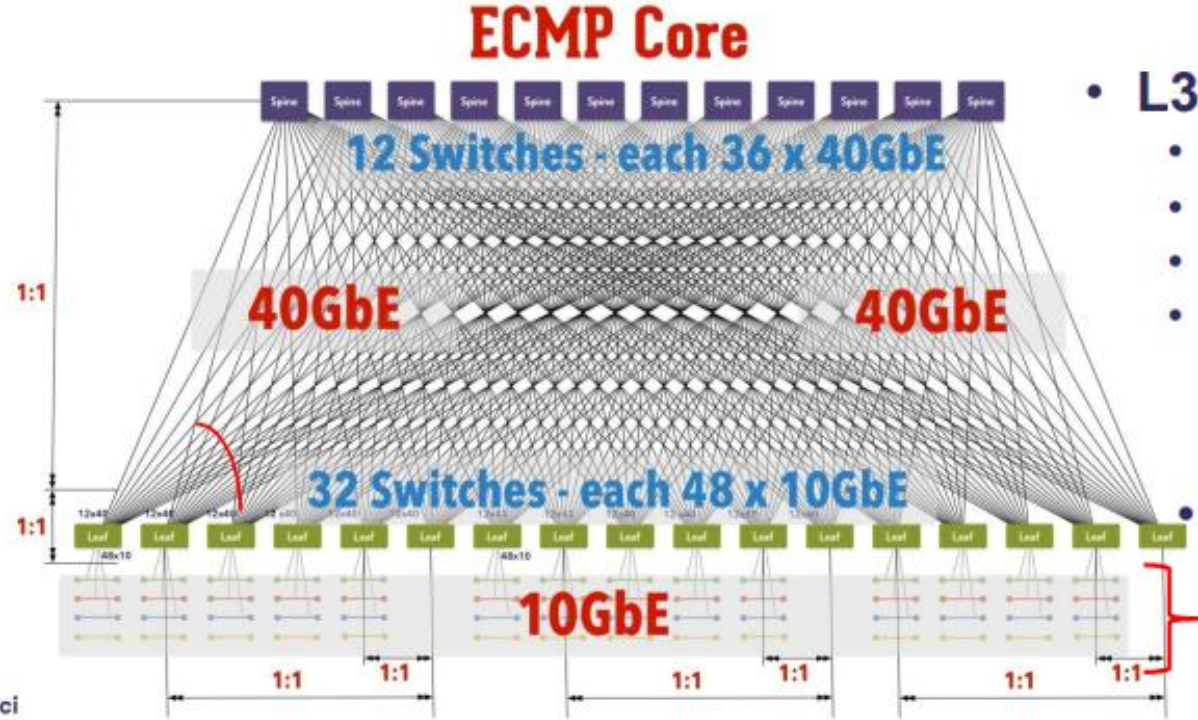
- ”Traditional” HPC cluster
  - General purpose network to access the nodes/launch jobs/ bulk storage
  - A “high performance”/”low latency” network for MPI/RDMA traffic / +/- storage
  - Maybe a block storage/iSCSI network
- Converged Ethernet cluster
  - Just One high performance , low latency, non-blocking Ethernet fabric
  - +/- RoCE config for Infiniband levels of RDMA/latency

- One network connection per node (lower cost nodes, less cabling)
- One network to config and monitor (fewer switches, lower cost, power)
- Limited IFB port counts ?? Vs 3 Tier CLOS Data centre fabric network.
- Lower training/skills requirement ?
- Latency spread/jitter higher than IFB ?

What applications do you run ?
How critical is MPI shared memory ?
Is message jitter important ?

Large Scale and Lower costs of L3 CLOS

- 36 leaf switches :1,728 Ports
- Non-Blocking. Zero Contention
- Low Latency (250nS L3 / per switch/router). 7-10uS MPI



- **L3 CLOS Leaf-Spine**
  - Leafs are L3 Routers
  - Spines reflect Leaf routes
  - OSPF Leafs to Spine
  - ECMP Traffic sharing

- **L2 Only on the Leafs**
  - Mostly /26. Max /24
  - Blast radius = 1 leaf
  - No ARP cache issues
  - No STP

Science and Technology Facilities Council
Scientific Computing

Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC

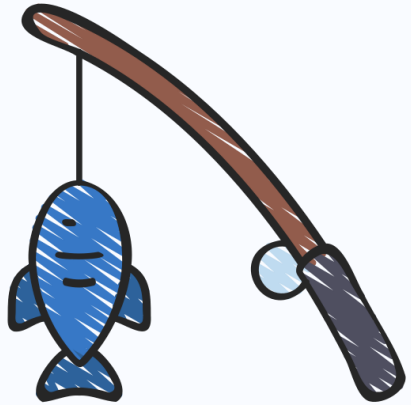
University of Southampton

Virtual Environments – Teaching to Fish

Taylor Haynes
Research Computing Systems Engineer
University of Southampton – HPC - iSolutions

Virtual Environments

- Package installation and dependency resolution eats up a lot of time
- Installing a package centrally and having it available in a module system often requires permissions a user doesn't have
- Is this a good use of time?
- A virtual environment makes a great fishing rod.



Conda

"Package, dependency and environment management for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, Fortran, and more"

- Conda is the virtual environment we have chosen to direct our users to.
- Community-driven channels are an invaluable resource.
- The majority of our Conda users have Python and R based workflows.
- Particularly useful for biomedical sciences and genomics researchers.

Management Challenges

- Upfront time cost in user education.
- In-place upgrades of the conda platform itself are very straightforward.
- Extra layer to troubleshoot in user support queries
- Inode management can be a headache with filesystem quotas

- Roughly 700,000 jobs were run on Iridis 5 in 2022
- 21.6% of these made use of a conda virtual environment