U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ESnet

ESnet Perspective: DC26 preparation, Clouds

Eli Dart

Network Engineer, Science Engagement

dart@es.net

LHCONE #53, Virtual

IHEP, Beijing, China

9 October, 2024

Upcoming Data Challenges

- DC26 and DC28
- LHC schedule recently changed
  - Run 3 extended into 2026
  - LS3 now runs through 2029
  - Run 4 (HL-LHC era) starting in 2030
- Not yet clear what this means for data challenges
  - DC26 → DC27?
  - DC28 → DC29?
  - Plans will emerge soon, I expect
- Regardless of schedule, we have work to do

Mini-Challenges

- Mini-challenges were valuable before DC24
  - Identified issues in time to fix them
  - Enabled sites to determine readiness
- Mini-challenges are likely to continue
  - Shawn, others in WLCG talking about this
  - ESnet is interested in participating as resources allow
- Key element: exchange point capacity
  - We cannot risk saturating exchange point interfaces
  - If a site wants to, that's fine - ESnet will not allow this
  - Ensure we collaborate on mini-challenges, data paths, etc.

Networking Progress → Host Focus

- Kate and Dale just presented ESnet update
  - Significant transatlantic capacity
  - Also transpacific, in collaboration with partners
- Significant improvements by other NRENs also
- Remember: this is a system
  - Computing, Networking, Storage
  - Distributed software stack (Rucio, FTS, etc.)
  - Physics code (production, user analysis, etc.)
- Initial analysis of DC24 results: host focus is good
  - Significant variation in host performance
  - This could be a good focus area for the next DC

Performance Differences

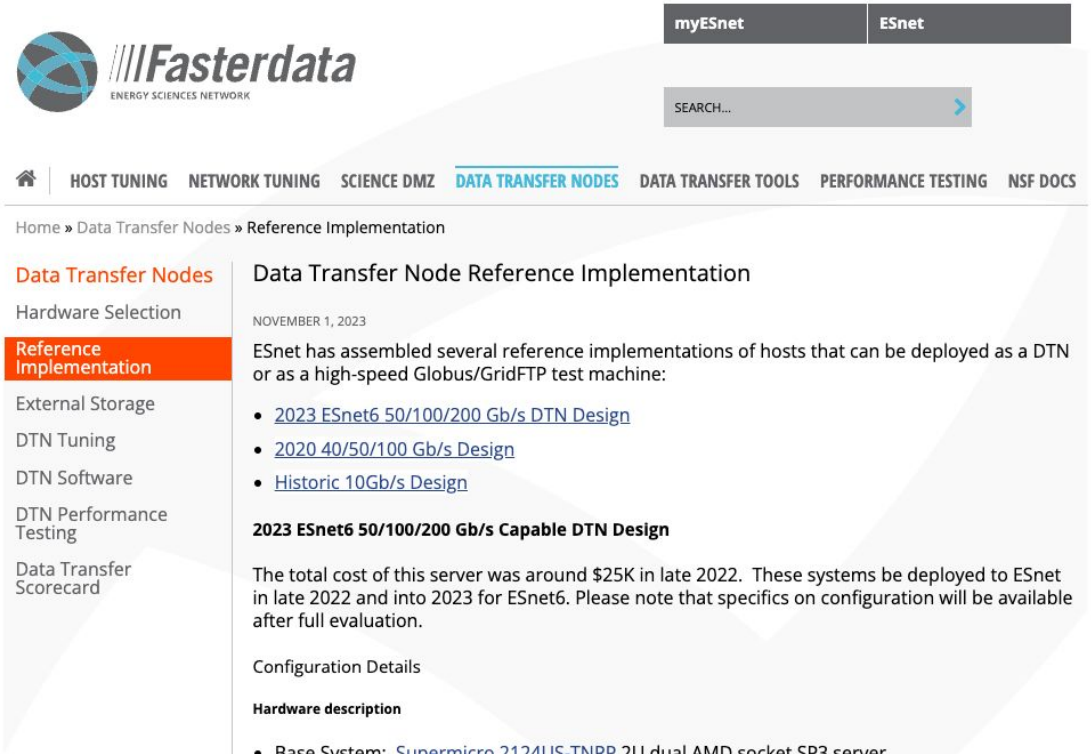
- From Bruno Hoeft, KIT
  - Significant difference in performance depending on site
  - KIT moved a lot of data with 8 DTNs and 8 squid boxes
    - Much higher per-host performance than some other sites
    - Why?
      - Host configuration?
      - Site network design?
- Potentially significant benefits from sharing
- What is the right way to approach this?
  - Share network designs?
  - Share host configs?

Knowledge Base

- One possible path forward: fasterdata
  - ESnet knowledge base: https://fasterdata.es.net/
  - Currently has host and network performance info
  - Includes DTN configuration, Science DMZ, etc.
- If there are known-good WLCG host configurations, we can document them there
  - Limitations: hardware changes fast
    - We publish multiple hardware builds, and we know they have a limited lifetime
    - Over time they change from a host specification to a design example
  - Advantages: share what others have found useful

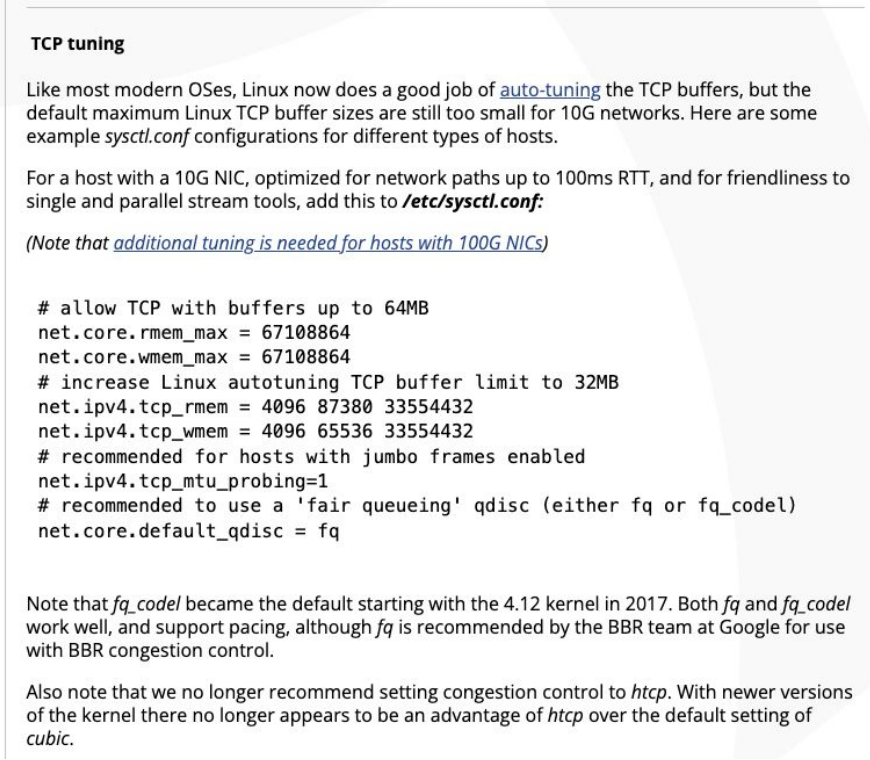
Knowledge Base: Host Designs

- Example: https://fasterdata.es.net/DTN/reference-implementation/
- No longer current, but perhaps useful
- If the community would find it useful, we can host designs on fasterdata



Knowledge Base: Network Tuning

- Fasterdata has multiple network tuning configurations, depending on intended use
  - 10G, 100G
  - Short distance, long distance
- https://fasterdata.es.net/host-tuning/linux/
- Picture is an example
- Question: how many WLCG sites tune the network stack on their hosts?



TCP tuning

Like most modern OSes, Linux now does a good job of auto-tuning the TCP buffers, but the default maximum Linux TCP buffer sizes are still too small for 10G networks. Here are some example *sysctl.conf* configurations for different types of hosts.

For a host with a 10G NIC, optimized for network paths up to 100ms RTT, and for friendliness to single and parallel stream tools, add this to ***/etc/sysctl.conf:***

(Note that additional tuning is needed for hosts with 100G NICs)

```
# allow TCP with buffers up to 64MB
net.core.rmem_max = 67108864
net.core.wmem_max = 67108864
# increase Linux autotuning TCP buffer limit to 32MB
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
# recommended for hosts with jumbo frames enabled
net.ipv4.tcp_mtu_probing=1
# recommended to use a 'fair queueing' qdisc (either fq or fq_codel)
net.core.default_qdisc = fq
```

Note that *fq_codel* became the default starting with the 4.12 kernel in 2017. Both *fq* and *fq_codel* work well, and support pacing, although *fq* is recommended by the BBR team at Google for use with BBR congestion control.

Also note that we no longer recommend setting congestion control to *htcp*. With newer versions of the kernel there no longer appears to be an advantage of *htcp* over the default setting of *cubic*.

Collaboration Before Next Data Challenge

- What is the right way to work together on this?
  - Is this needed?
  - When is the right time to engage?
- Soon, Tier1 and Tier2 sites will be buying gear that will be in production for HL-LHC/Run 4
  - Should we collaborate on designs?
- Clearly, every site must have the freedom to do what they need to do
  - Vendor relationships
  - Budgets
  - Acquisition timelines
  - Local policies, designs, etc.
- Very interested in future discussion/collaboration

Cloud

- LHC experiments have done a lot of good work on understanding and using Cloud
- ATLAS and CMS Cloud Blueprint: https://arxiv.org/abs/2304.07376
- ATLAS GCP project:
  - Operational: https://arxiv.org/abs/2403.15873
  - Total cost of ownership: https://arxiv.org/abs/2405.13695
- Many other efforts over the years

Cloud Going Forward

- ATLAS TCO paper makes it clear that Cloud is not currently cost effective for many workloads
- However, it is useful for some things, e.g.:
  - Bursting to handle peak load
  - Trying out new hardware
- Also valuable for the community to keep skills up to date
  - Ability to use the cloud when it makes sense to do so
  - Understanding how the space is changing over time

Next Cloud Projects

- Connecting a Cloud instance to LHCONE is listed under Future Work
- If this Future Work is undertaken, ESnet is interested in participating
  - Understand impact on existing services
  - Understand if new capabilities/configurations are needed
- Please let us know if you do this!

Discussion

- What work should we do together on systems before the next Data Challenge?
- What work should we do together on site network design before the next Data Challenge?
- What projects are coming up in the Cloud space?

U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ESnet

Thanks!

Eli Dart

dart@es.net

https://my.es.net/

https://www.es.net/

https://fasterdata.es.net/