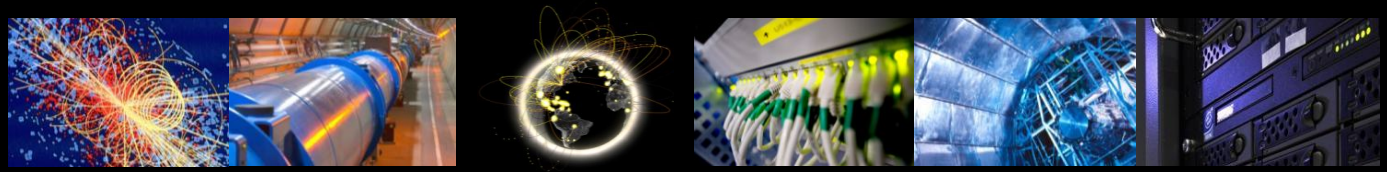
Network Traffic Optimization (Jumbo, protocols, pacing)

Justas Balcas / ESnet, Asif Shah / FNAL, Shawn McKee / U of Michigan

HTC 24

(https://agenda.hep.wisc.edu/event/2175/sessions/3176/#20240710)

Julv 10. 2024



WLCG Worldwide LHC Computing Grid

Open Science Grid

Network Optimization: Why & How

Our sites (or their host institutions) pay to have significant bandwidth, e.g., current Tier-2 sites typically have one or more 100 Gbps links to the WAN.

- However, have a 100 Gbps connection does NOT mean sites can use that capacity with any efficiency or consistency.
- Many sites are can't reliably use more than a fraction of their bandwidth
- While old or misconfigured hardware is often part of the problem, the network is typically one of the main culprits.

Our goal for network optimization is to maximize our ability to utilize the available bandwidth.

The current network toolkit for this issue is currently is comprised of:

- Jumbo frames
- Packet (traffic) pacing
- Protocols

Traffic Pacing

One way to help address the challenge for HEP storage endpoints to utilize the network efficiently and fully is traffic (packet) pacing.

- Traffic pacing means sending packets at a specific rate, corresponding to to some fraction of the total network bandwidth.
- Without traffic pacing, network packets are emitted by the network interface in **bursts**, corresponding to the wire speed of the interface.
  - **Problem**: microbursts of packets can cause buffer overflows
  - The impact on TCP throughput, especially for high-bandwidth transfers on long network paths can be **significant**.

How? Traffic pacing can be simply enabled and controlled by the Linux 'tc' application, part of the 'iproute' package.

The challenge is not in enabling the pacing so much as determining what the pacing should be for a given host and transfer…

Jumbo Frames

Jumbo frames: any maximum size bigger than 1500 bytes. 9000 bytes is the most used value in the R&D community. Note IPv4 and IPv6 frames have a **standard** maximum size of 1500 Bytes (IP header + payload)

Benefit: Reducing the relative size of the IP header over the payload can reduce the load on the CPU of both the sender & receiver of large data flows, thus **allowing greater throughput** for CPU intensive transfers

Risk: On the other hand, transfers between hosts using different MTUs can lead to traffic blackholing if the networks in between are not properly configured (See Fasterdata MTU page)

Some technical details in https://indico.cern.ch/event/725706/contributions/3120030/attachments/1743507/2821722/LHCONE-MTU-recommendation.pdf

WLCG has suggested a target of 50% of traffic using Jumbo Frames for DC26 and 99% by DC28 (assuming it remains beneficial)

Jumbo Frame Impact

ESnet has done testing showing:

Single stream

- Jumbo frames are 3x faster on 100G hosts
- Jumbo frames are about 15% faster on 10G hosts

8 streams:

- Jumbo frames are about 25% faster on 100G hosts
- Jumbo frames are the same as 1500B 10G hosts

Note testing in the UK has show that **some** long distance tests saw varying benefit

Source	Destination	RTT	9000	1500
SURF (NL)	RNP (Brazil)	100ms	31 Gbit/s	20 Gbit/s
Jisc (London)	BNL (USA)	100ms	14 Gbit/s	6 Gbit/s
SURF (NL)	Jisc (London)	7.2 ms	23 Gbit/s	6 Gbit/s

TCP Protocols

BBR(v3): A TCP protocol developed by Google that uses round-trip time (RTT) and sometimes throughput as indicators of congestion.

- ESnet testing at RTT 150ms: Throughput is 2-3 times better with BBR than CUBIC and parallel streams step on each other less than expected
- Might be the best and easiest option, once it is available in the linux kernel

BIG TCP: (See also https://www.phoronix.com/news/Linux-6.3-Networking-BIG-TCP)
Creates a very large TCP packet and relies up TSO (TCP Segment Offload) and GRO (Generic Receive Offload) hardware assist features in NICs to quickly fragment, transmit and receive.

- Requires 6.3+ kernel for IPv4 and 5.19+ kernel for IPv6
- Practical use for WLCG needs further testing

Plans for Prototyping, Testing and Evaluation

We have some demonstrated technologies that should improve our ability to more fully utilize the bandwidth we have available, but we need to determine how best to use it in our production systems.

The first step should be to **identify sites/users** who are willing to participate in prototyping and testing these technologies.

Once we have some sites enabled, we can incorporate them into the next mini-challenge or create a mini-challenge for specific testing

- Sites first run with their normal configuration and then with the specific technology enabled.
- We may need multiple tests depending upon how much configuration phase space we want to explore.

We will need to carefully document results and perhaps expand or redo testing based upon the outcome.

Questions To Discuss (and Try to Answer)

What benefit can we observe for PRODUCTION systems using?:

- Jumbo frames
- Packet(Traffic) pacing using ‘tc’
- Alternative protocols BBRv3/BBR-swift/BIG TCP

Who is willing to participate in testing any of the above?

What is the proposed timeline for testing each of the above?

Can we produce a document for each traffic optimization option describing the configuration, testing and analysis?

Can we summarize by creating a best practice guide (perhaps an update to ESnet’s Fasterdata page)?

How can we organize the work? (Suggestion: Use the RNTWG subgroup on pacing and rework its scope/mandate)

Summary

We have an opportunity to improve our site's ability to utilize the existing bandwidth they have via various technologies

We need to prototype, test and produce best practices so that beneficial capabilities can be available as part of our production by DC26 (and beyond)

Question, Comments, Discussion?

Acknowledgements

We would like to thank the **WLCG**, **HEPiX**, **perfSONAR** and **OSG** organizations for their work on the topics presented.

In addition we want to explicitly acknowledge the support of the **National Science Foundation** which supported this work via:

- IRIS-HEP: NSF OAC-1836650 and PHY-2323298

Backup Slides

Network Optimization Related Documents/Presentations

Spring 2024 HEPiX RNTWG report
https://docs.google.com/presentation/d/1Dk8GTVnBqGSVrSjuUTTi_IZ-mfTanJrBqhKA7_DfIiM/edit#slide=id.g8036819354_0_7
RNTWG Traffic Shaping Charter:
https://docs.google.com/document/d/1FZjHHlXy-3J-2S-PdjjclwKWMkWq_M9zv0G4O-tSL9M/edit#heading=h.kjs85ae6lo7a
BIG TCP:
https://isovalent.com/blog/post/big-tcp-on-cilium/#:~:text=BIG%20TCP%20over%20IPv6&text=Unusually%2C%20BIG%20TCP%20support%20was,be%20inserted%20into%20the%20packet.
Recent Jumbo frames survey:
https://docs.google.com/document/d/1x0c5rrJedjlEhfb6pgjCdUpQ3WWRHk3SpdWWVrH0578/edit
Improving LHCONE security & Use of Jumbo frames:
https://indico.cern.ch/event/1369601/contributions/5947381/attachments/2855497/4993734/WLCG-20240514-WS24--LHCONE-security-and%20jumbo.pdf
BBRv3 testbed spreadsheet:
https://docs.google.com/spreadsheets/d/1U0VXIfWHfpK7bX7k2ucFep4Xo_4KQ5x-7rKD7e4az7Y/edit?gid=0#gid=0
DC24-BBRv3-Jumbo-Frames:
https://docs.google.com/presentation/d/1OroaISoRdFp9cpNZ-U13nRCf8kKPiLje3dpPE8Zf6uo/edit#slide=id.p