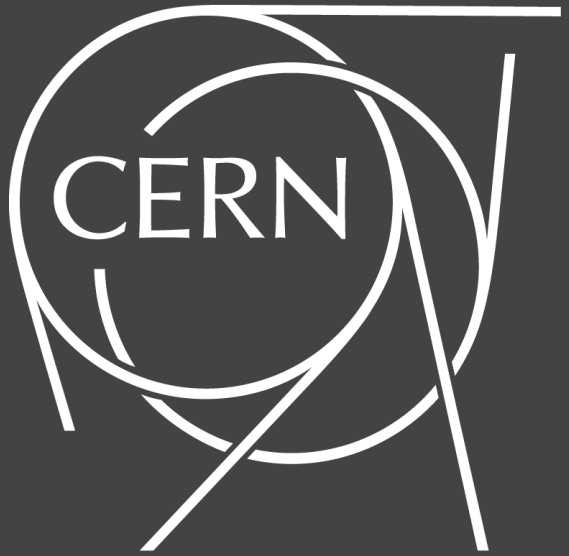

CERN

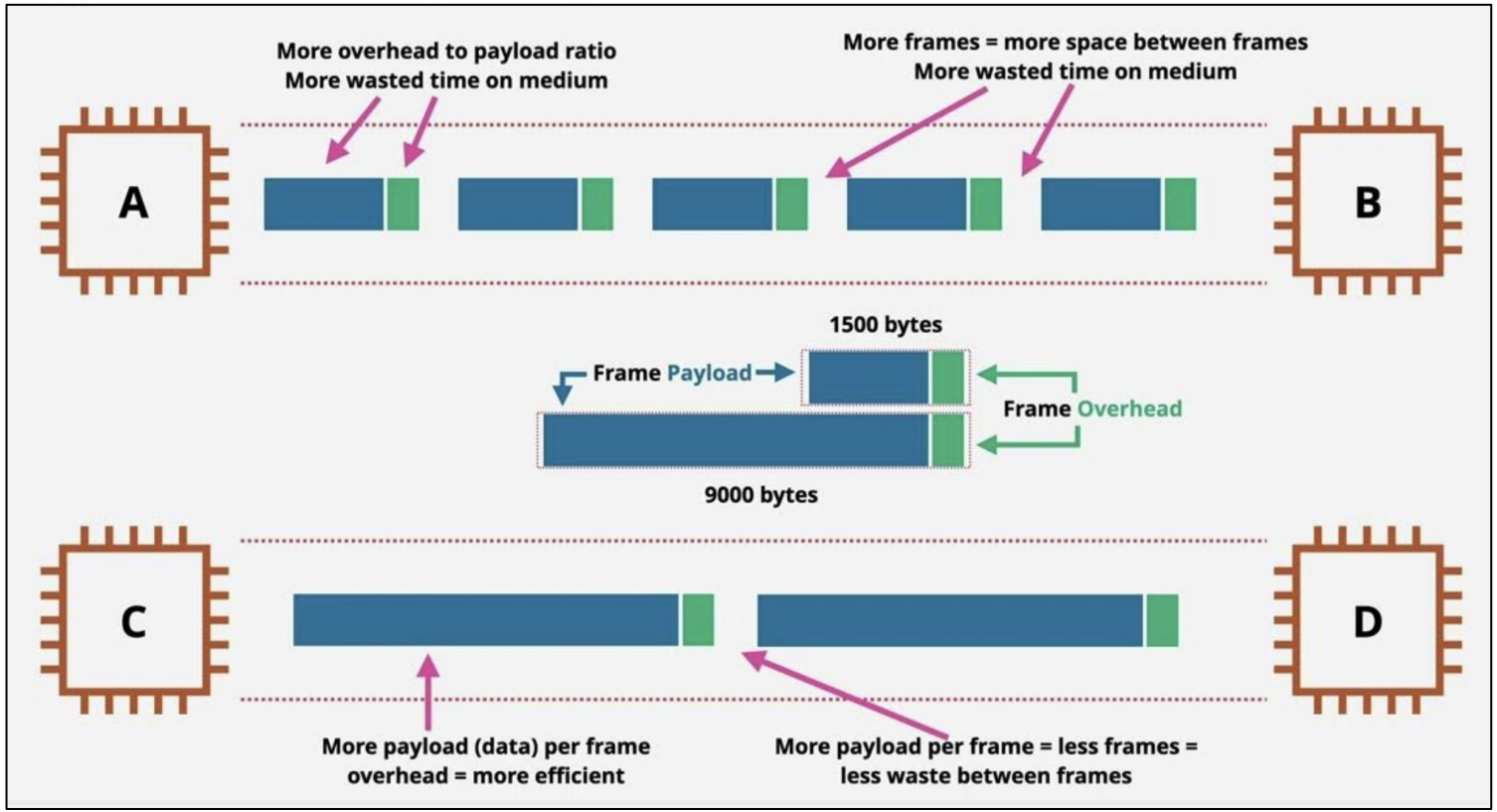
Changing MTU Packet Size in EOS Storage at CERN

Jumbo Frames (MTU 9000 bytes)

Edoardo Martelli, Maria Arsuaga Rios, Andreas-Joachim Peters
OTF 10.12.2024

Background

Jumbo Frames explained …



Switching EOS servers to MTU 9000

- Why?
  - Initial request from **CMS (**online)
- Why else?
  - Improved **Throughput**
    - lower ratio header/paypload
  - Reduced **CPU** Overhead
    - less packets to handle
  - Lower **Latency** for High Bandwidth Applications
    - less fragmentation
  - Better **Utilization** of Network Resources
    - less congestion/packet drops
  - **Optimized** for Modern Data-Intensive Applications
    - SD-WAN MPLS
  - Enhanced **Performance** in Virtualized and Storage Environments
    - support in VMWare, Hyper-V, ISCSI, NFS

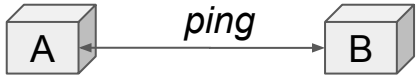
MTU Pre-testing

Switching EOS servers to MTU 9000

- How?
  on each storage server in EOSPILOT: `ip link set dev ethXYZ mtu 9000`
  → we are done 😁 !
  …

  → But it didn't work 😭 !
  EOS transfers got stuck/slow → packet loss 0.1% to 100%

- How to test that something is wrong?



  - Verify MTU 9000 setting with ping request: packet of 9000 bytes including headers
    - IPV6 headers = 40 bytes + 8 bytes ICMP payload in a ping request
    - This has to work:

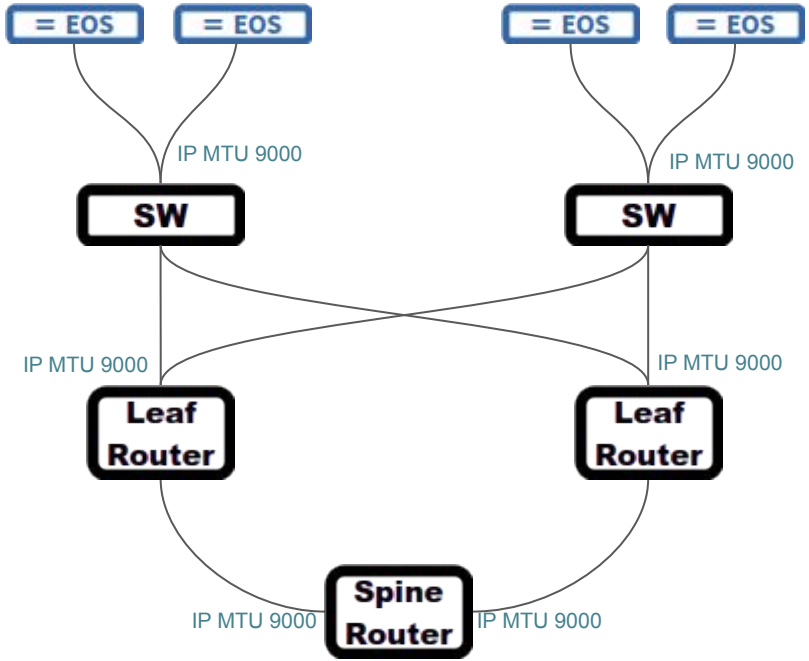
```
ping -f -6 -M do -s 8952 mtu9000-host     ( disable fragmentation )
ping -f -6 -s 10000 mtu9000-host          ( require fragmentation )
```

    - **both didn't work properly - too high/complete packet loss!**

Network Background

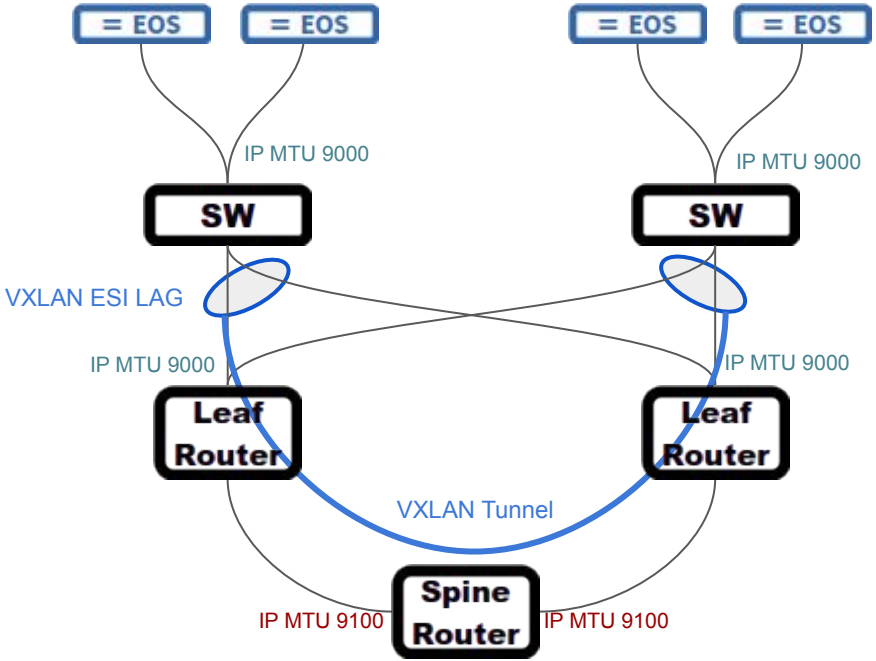
Network MTU configuration

- All network links had Ethernet MTU 9216 and IPv4/6 MTU 9000
- VXLAN ESI sends original packets over direct connections, but they get VXLAN encapsulated when they go via the Spine Routers
- 9000B user IP packets become 9050B IP packets when encapsulated in VXLAN
- These VXLAN packets didn't fit in the links to the Spine routers
- The IP MTU of the links between Spine and Leaf routers has been increased to fit VXLAN (9100B). This change has been applied to all the PDC and MDC routers
-

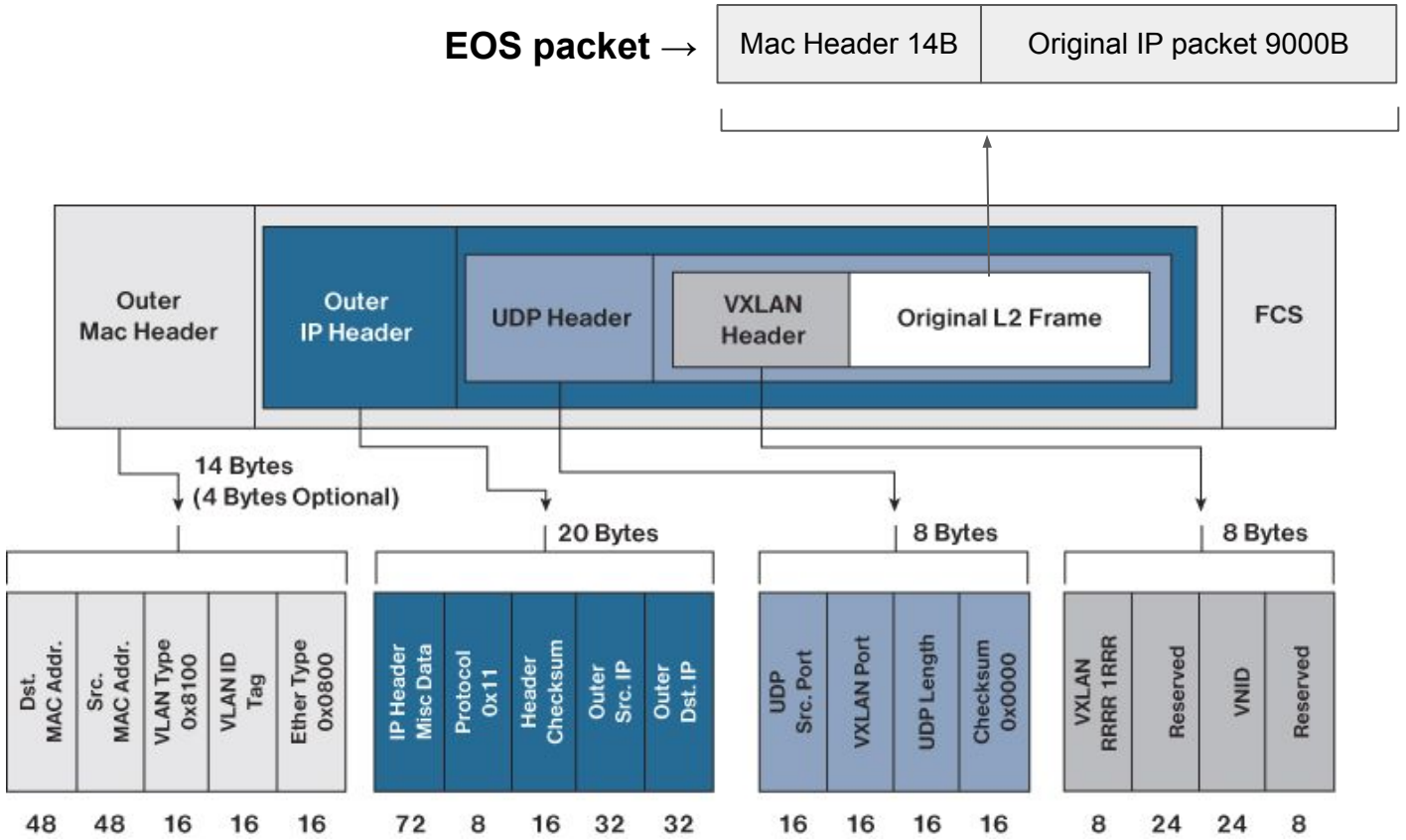


Network MTU configuration

- All network links had Ethernet MTU 9216 and IPv4/6 MTU 9000
- VXLAN ESI sends original packets over direct connections, but they get VXLAN encapsulated when they go via the Spine Routers
- 9000B user IP packets become 9050B IP packets when encapsulated in VXLAN
- These VXLAN packets didn't fit in the links to the Spine routers
- The IP MTU of the links between Spine and Leaf routers has been increased to fit VXLAN (9100B). This change has been applied to all the PDC and MDC routers



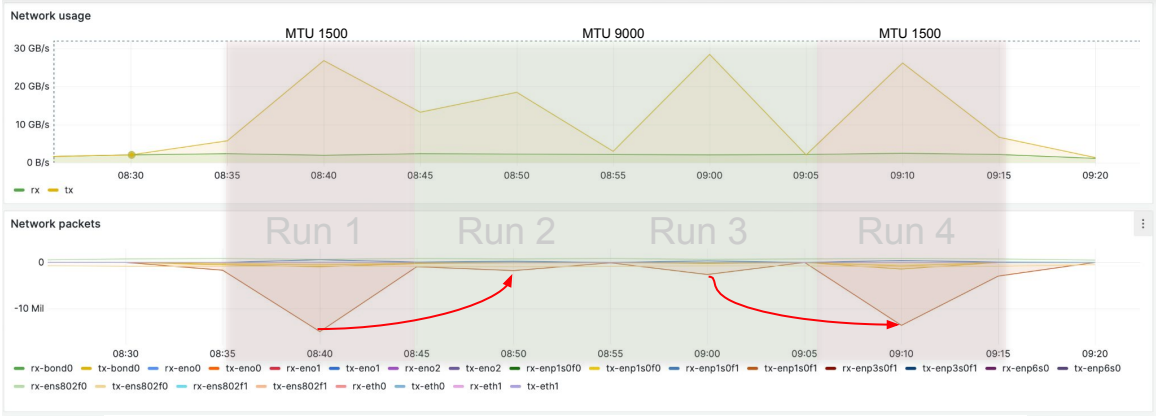
VXLAN encapsulation



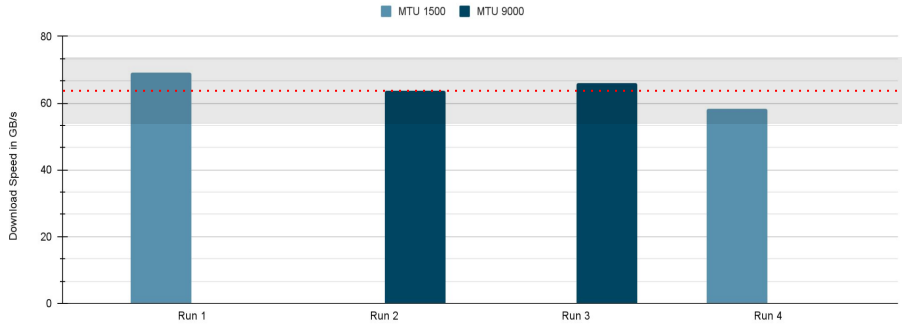
MTU Impact Pre-Testing

Benchmarking EOS servers with MTU 9000 in EOSPILOT

- Verify clients with MTU 1500 and MTU 9000 work against EOSPILOT server with MTU 9000
  - Download Benchmark 7.34 TB from 70 nodes toggling all clients between MTU 1500 and 9000
    Attention: *runtime* < *bin* size in **Monit**



Run	MTU	Time [s]
1	1500	106
2	9000	116
3	9000	110
4	1500	126



- Number of packets reduced when MTU 1500 → MTU 9000 ✓
- No obvious change in performance - normal fluctuations ✓

Planned testing with CMS

Plan for JUMBO Frame Tests for CMS

- Overview
  - Based on requests from Point5 and discussions with the network team, JUMBO frame tests for CMS are planned.
  - **Focus**:
    - Short-distance tests (Point5) and long-distance tests (FNAL).
    - CMST0 activity included to evaluate the impact on non-JUMBO frame machines connecting to EOS.
  - **Goal**: Assess performance benefits, especially over long distances, and evaluate the impact on non-JUMBO frame machines connecting to EOS.

Plan for JUMBO Frame Tests for CMS

- EOSPILOT instance
  - Tests will be conducted on EOSPILOT instead of EOSCMS.
  - Involves transfers from Point5, CMST0, and FNAL via RUCIO.
  - Safer environment to evaluate performance and impact before moving to EOSCMS.
  - Requires small configuration changes for all parties to interact with EOSPILOT instance: permission accesses and new paths.

- Key involvement
  - **From CMS**:
    - Point5: Data Transfer from P5 to EOS.
      - Proposal: Transfer **50TB from P5 to EOS** with **6GB filesize**.
    - CMS-RUCIO: FNAL transfers via RUCIO.
      - Proposal: Transfer **50TB from CERN to FNAL** to evaluate performance with **3-4GB filesize**.
    - CMST0: Simulating activity with non-JUMBO frame clients right after the ingestion from Point5.
  - **From IT**: EOS and Network teams will monitor the traffic and the speed/time for moving that data.

Plan for JUMBO Frame Tests for CMS

- Schedule

  - **January 13, 2025**:
    - Morning: Tests without JUMBO frames (Point5 & CMST0).
    - Afternoon: FNAL tests without JUMBO frames.
  - **January 14-15, 2025**: Tests with JUMBO frames enabled.
  - **January 16, 2025**: Final round of tests without JUMBO frames.

Thank you for your attention!

Question or comments?