Converged Ethernet and Storage Appliances
Where do we come from?

• The world used to be a simple place
  • NAS: NFS over UDP/TCP, SMB/CIFS over TCP, one session per client
  • Centralized Storage (Monolithic Servers)
  • Few custom TCP sessions for backups
  • (SAN: out of scope)

• State-of-the-Art:
  • Massive clusters of storage nodes
  • Interconnected using various technologies, moving towards Ethernet
  • Plethora of requirements (regulatory, DevOps, Features) met with increasing number of (internal and external) protocols
    • all behave slightly different, with complex interdependencies
    • any HoL blocking a major issue
• Used to be different physical interfaces, different physical networks
  • Mgmt
  • Frontend Storage Traffic (NFS, SMB, iSCSI, FCoE)
  • Storage specific Traffic (Backup, Replication, Configuration)
  • Backend Traffic (used to be FCP, SAS, moving to Ethernet too)

• New Traffic types
  • NVMe/RDMA (RoCE, iWARP, TCP)

• Few, high bandwidth links (n 100G)

• Segregation of traffic types and classes only via DSCP / CoS (QoS)

• 100...1000s of parallel traffic flows, various clients and traffic behavior
  • Storage Specific / Backend – often „Elephant“ Flows – high bandwidth demand, continuous
  • Frontend – Mix of burst and continuous (lower bandwidth)
  • New traffic – highly bursty, highly latency and loss sensitive, phases of very low and very high throughput
• Challenges
  • Frontend Traffic uses TCP (RoCE) with mix of different CC mechanisms
    • NewReno, Cubic, Compound TCP, ECN TCP
  • New traffic classes need different queuing response (AQM -> IETF L4S effort)
    • DCQCN, DCTCP – to be marked with ECT-1 (experimental) for proper queue response selection
  • Frequent backpressure by singular receiver via FlowControl
    • Generally, overall throughput improved WITHOUT flow control
    • Latency / Loss sensitive Traffic flows require Flow Control regardless despite lower performance
  • Real deployments exposed to unpredictable cross traffic
    • Higher loss rates, burst losses, reorderings; head of line blocking induces high delay spikes
  • Legacy QoS very complicated to set up and maintain, poor education to operators about the implications of WRR, AQM, FlowControl
    • Complex interactions, bad predictability
Ideal Solution

• Should automatically adapt to the specific environment
• Reclassify offending traffic (remove head of line blocking)
• Provide mechanisms to allow the co-existence of legacy and modern protocols, with minimal administrator interaction
• Adhere to the vendor’s selection of QoS parameters automatically (e.g. relative priorities, provide minimum bandwidth per class, ...