SC20

Everywhere we are | more than hpc.

A Trial Deployment of a Reliable Network-Multicast Application across Internet2

Yuanlong Tan (Presenter), Malathi Veeraraghavan, Hwajung Lee, Steve Emmerson, Jack Davidson

Nov. 12, 2020 • INDIS 2020

Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

Contributions

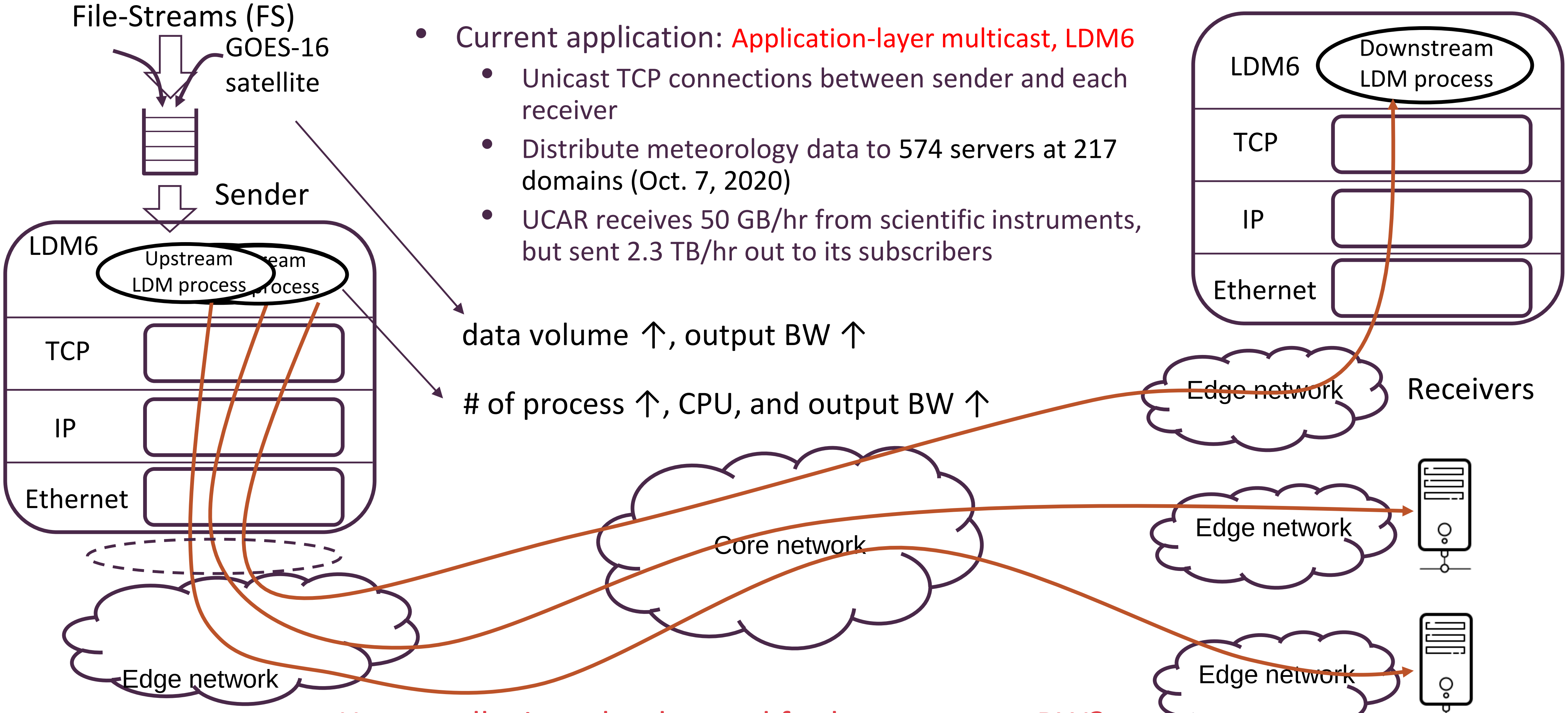
- Discussing a cross-layer architecture (DRFSM) for scientific file-stream distribution, specifically for Local Data Manager (LDM)
- Designing and implementing a performance monitoring system
- Implementing and evaluating the discussed architecture over a multi-domain trial deployment with the performance comparison with the current solution

Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

Background -- UCAR Unidata IDD project

- Current application: Application-layer multicast, LDM6
  - Unicast TCP connections between sender and each receiver
  - Distribute meteorology data to 574 servers at 217 domains (Oct. 7, 2020)
  - UCAR receives 50 GB/hr from scientific instruments, but sent 2.3 TB/hr out to its subscribers



data volume ↑, output BW ↑

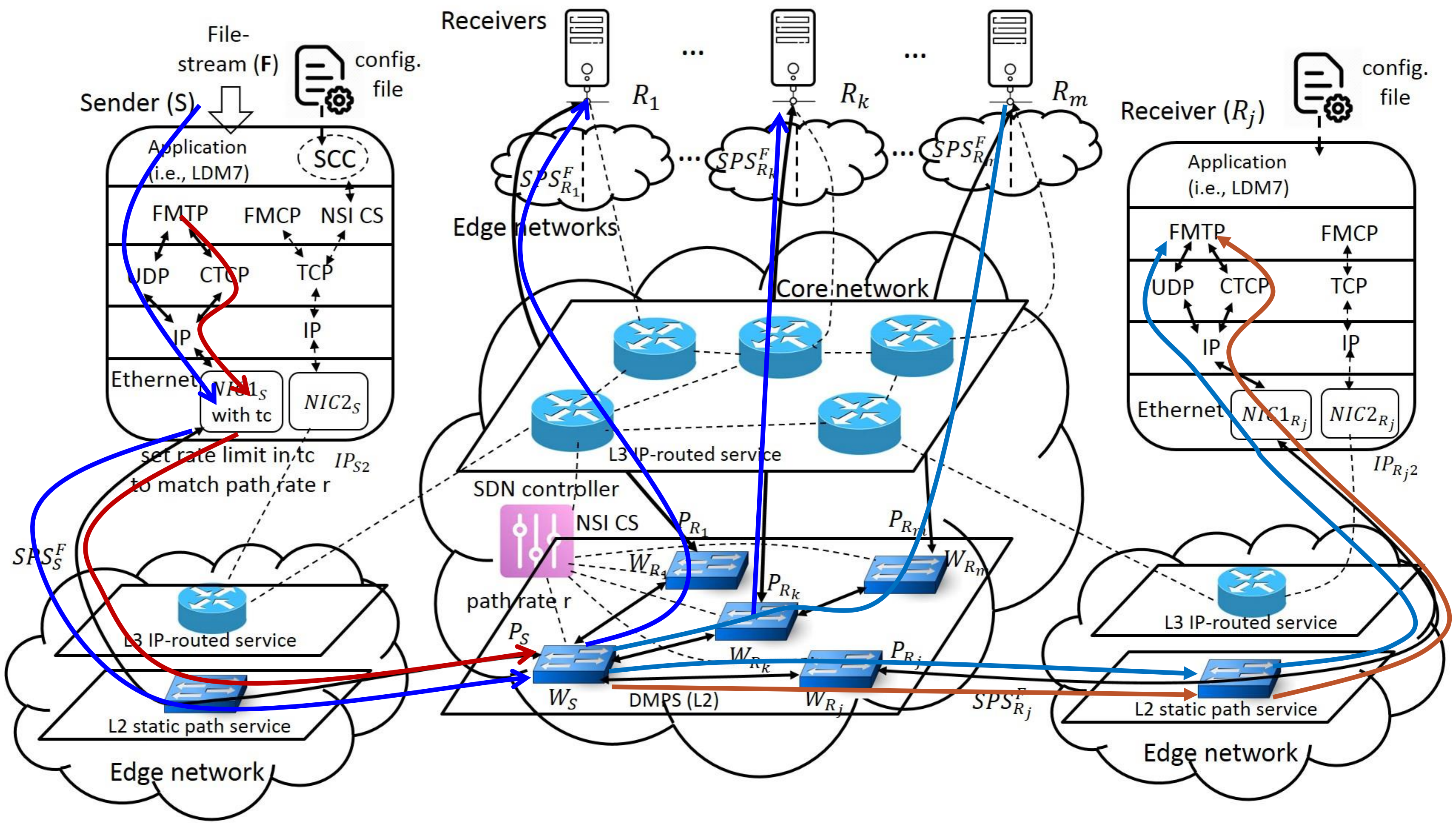
# of process ↑, CPU, and output BW ↑

How to alleviate the demand for larger output BW?

Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

Cross-layer architecture & LDM7



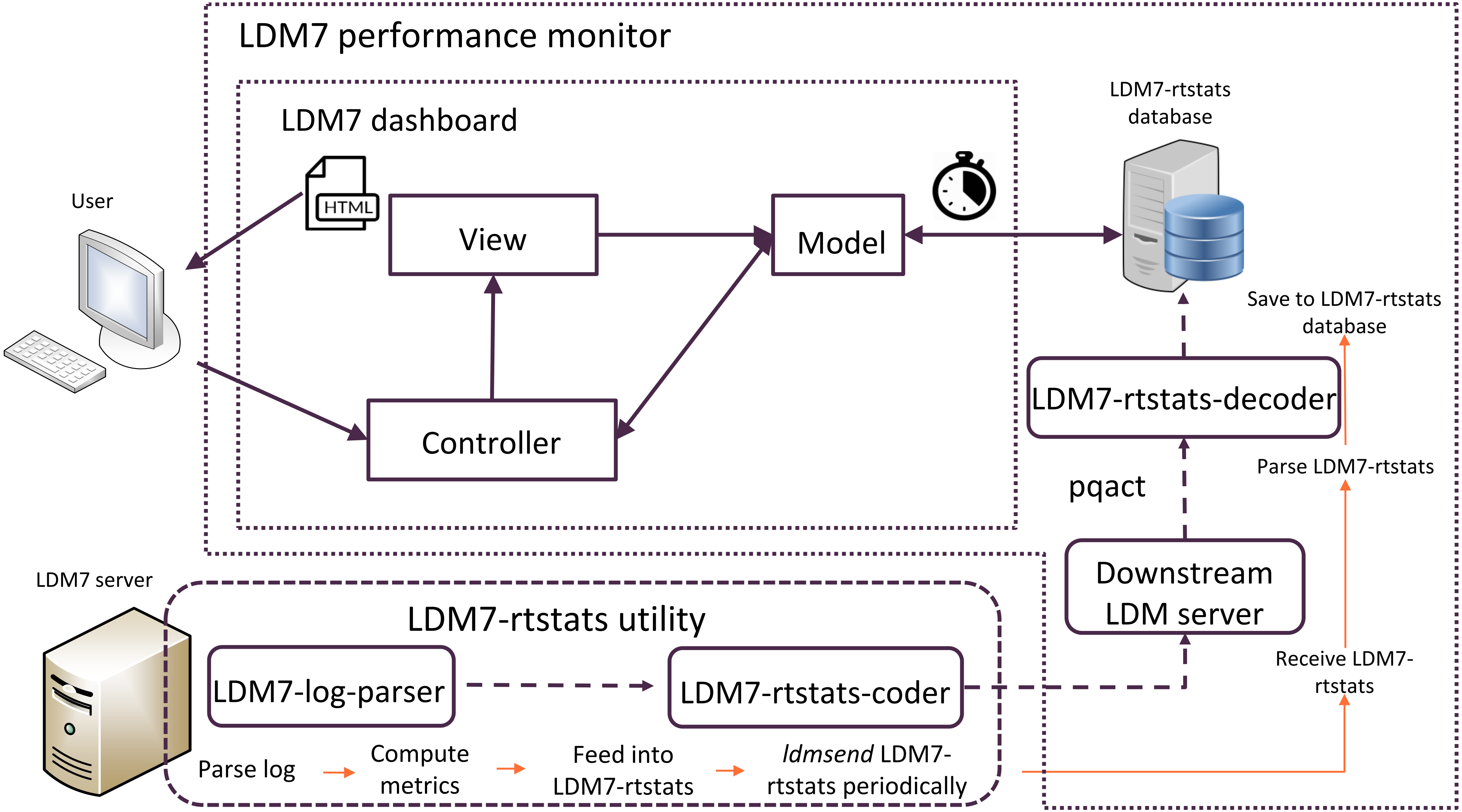
SCC: SDN Controller Client; FMTP: File Multicast Transport Protocol; FMCP: File Multicast Control Protocol; NSI CS: Network Service Interface Connection Service; tc: traffic control; SPS: Static Path Segment; L3: Layer-3 (IP header-based forwarding); L2: Layer-2 (VLAN/MPLS); DMPS: Dynamic Multipoint Path Service;

- Network Multicast
  - L2 path service simplifies
    - error control, flow control, and congestion control
  - A transport protocol, FMTP, used for reliable multicast
- Two types of network:
  - L3 IP-routed service
  - L2 path service
- Two types of traffic:
  - L3: control-plane messages
  - L2: scientific data distribution
- Provision L2 multicast tree before disseminating data

Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

LDM7 Performance Monitoring System



rtstats: Real-time statistics reporting system

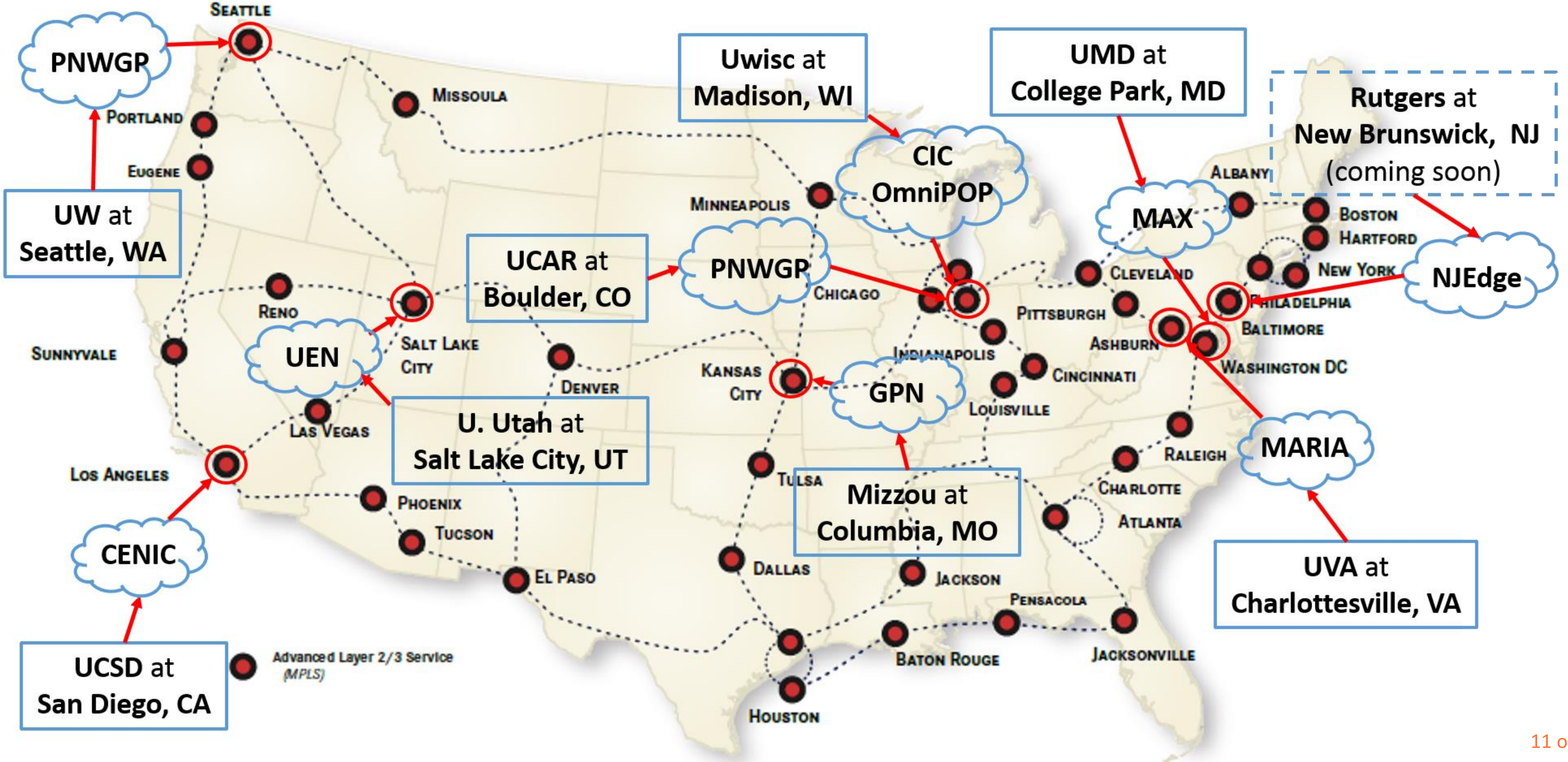
ldmsend: program to send data using LDM protocols

pqact: program to process products in a Unidata LDM product queue

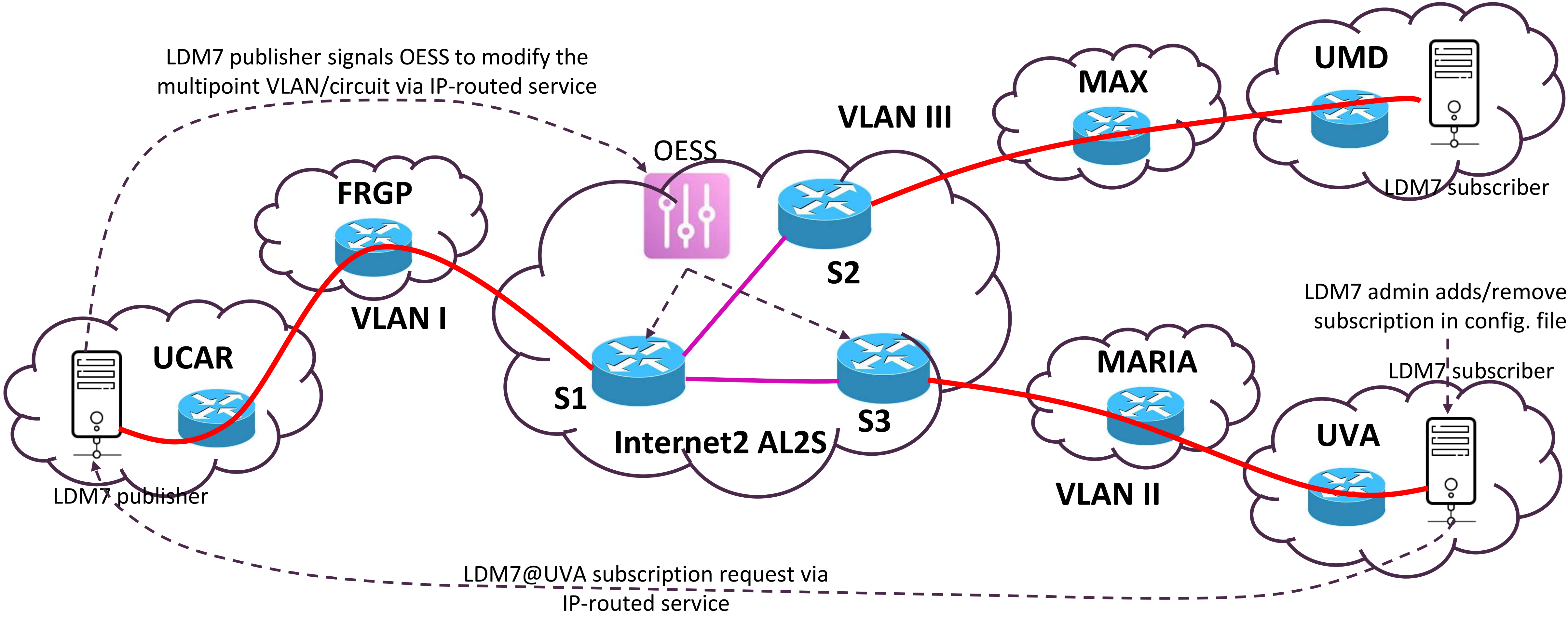
Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

Trial Deployment of LDM7 across Internet2



Control-plane procedures



LDM7 publisher signals OESS to modify the multipoint VLAN/circuit via IP-routed service

LDM7 admin adds/removes subscription in config. file

LDM7@UVA subscription request via IP-routed service

— Static provisioned VLANs — Dynamic VLANs provisioned by OESS - - - → Control message over IP-routed service

OESS: Open Exchange Software Suite (OESS) FRGP, MAX, and MARIA: regional R&E that provides Internet2 access for UCAR, UMD, and UVA

Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

Experimental setup and execution

UCAR

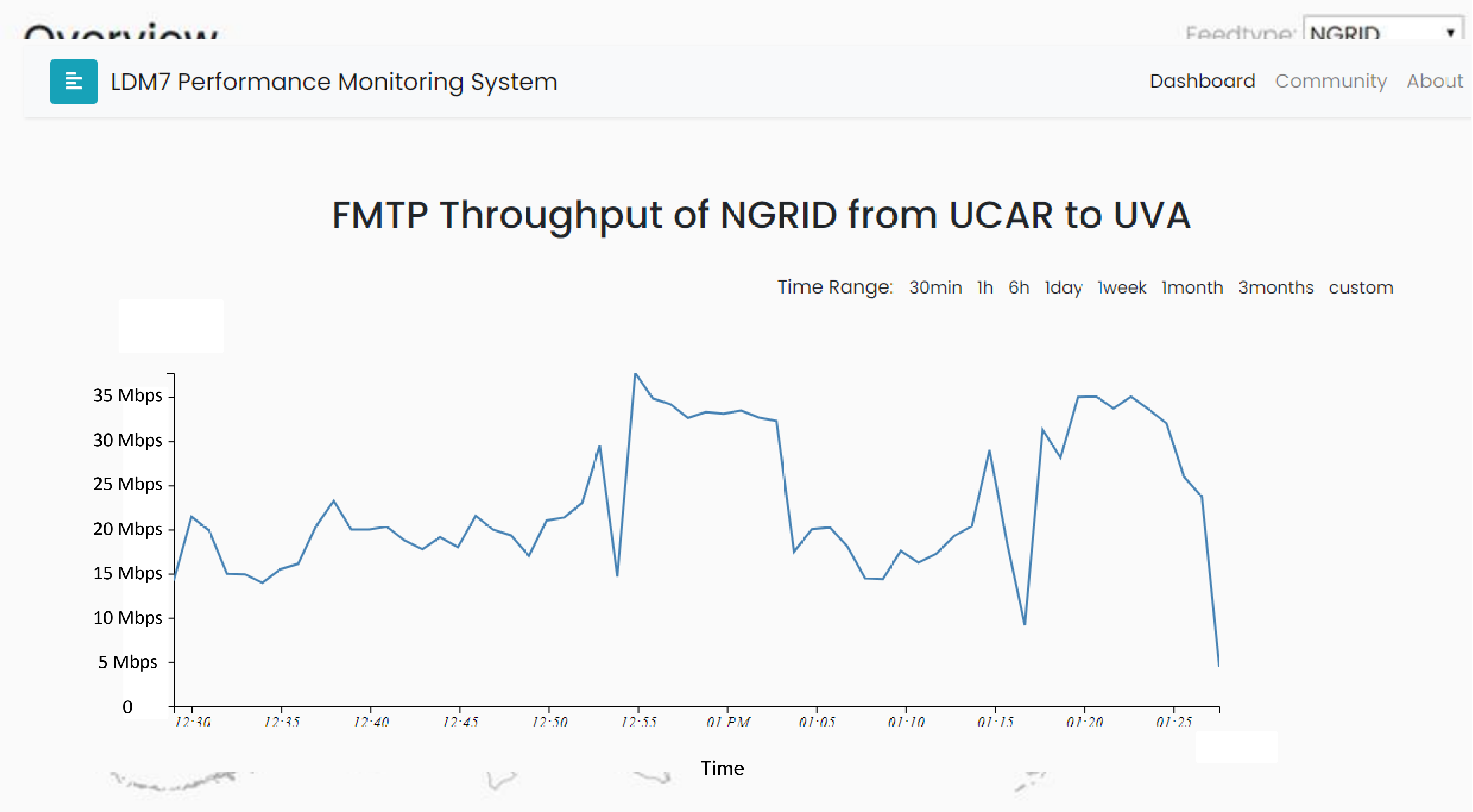
- Trial deployment across Internet2 at 8 campuses/institutions
  - UCAR, UVA, UWisc, UMD, U.Utah, UWash, UCSD, and U.Missouri
  - Each server has
    - At least 64 GiB RAM and 500 GB disk space
    - Two network interface cards, a GE NIC for control-plane and a 10 GE NIC for data-plane
    - CentOS 7 Linux distribution, but kernel version varies slightly.
- Linux network traffic control utility, *tc*
  - Used for queueing discipline (qdisc info)
  - Created two queues, one for multicast packets and one for unicast packets (retransmission)
    - ➢ Each queue was 600 MiB
    - ➢ The queues shared the available bandwidth
- Execution
  - Multicasting NGRID from UCAR to other subscribers
  - Three sets of experiments to evaluate LDM7 and compare performance with LDM6

Metrics

- Throughput
  - Per-file throughput: $T_{per} = {}^{s_i}/_{t_i}$
  - Average throughput: weighted harmonic mean -- $T = \frac{\sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} \frac{s_i}{T_{per}}} = \frac{\sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} t_i} = \frac{S}{T}$
- FMTP File Delivery Ratio (FFDR): the success of file delivery via data-plane
  - File-count-based FFDR: $F^{count} = \frac{N'}{N} * 100\%$
  - Size-based FFDR: $F^{size} = \frac{S'}{S} * 100\%$
- Multicast Packet Loss Rate (MPLR): proportion of packet loss with respect to packets sent
  - MPLR: $\mathrm{L} = \frac{B_t * (MTU - FMTP/TCP/IP\ headers)}{S'} * 100\% \Rightarrow L = \frac{B_t * 1448}{S'} * 100\%$, MTU is 1500

Dashboard

UCAR



Overview

Feedtype: NGRID ▾

LDM7 Performance Monitoring System

FMTP Throughput of NGRID from UCAR to UVA

- Colored points:
  - Green: active node
  - Red: unavailable node
  - Yellow: less active
- Publisher: UCAR
- Subscribers: UVA, UMD, UWisc, U.Utah, UCSD, Uwash, U.Missouri, and Rutgers
- Logical links

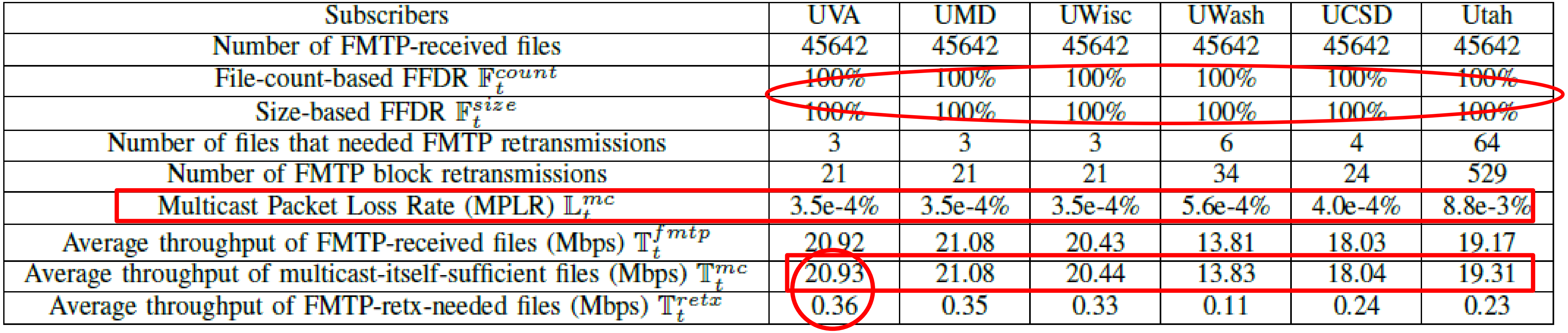
  Color: Dark → Light

  Value: Large → Small

- URL for LDM7 performance dashboard: http://idc-uva.dynes.virginia.edu:3000/

LDM7 Performance

- NGRID 2020-07-12 03:00-04:00 UTC, 40 Mbps



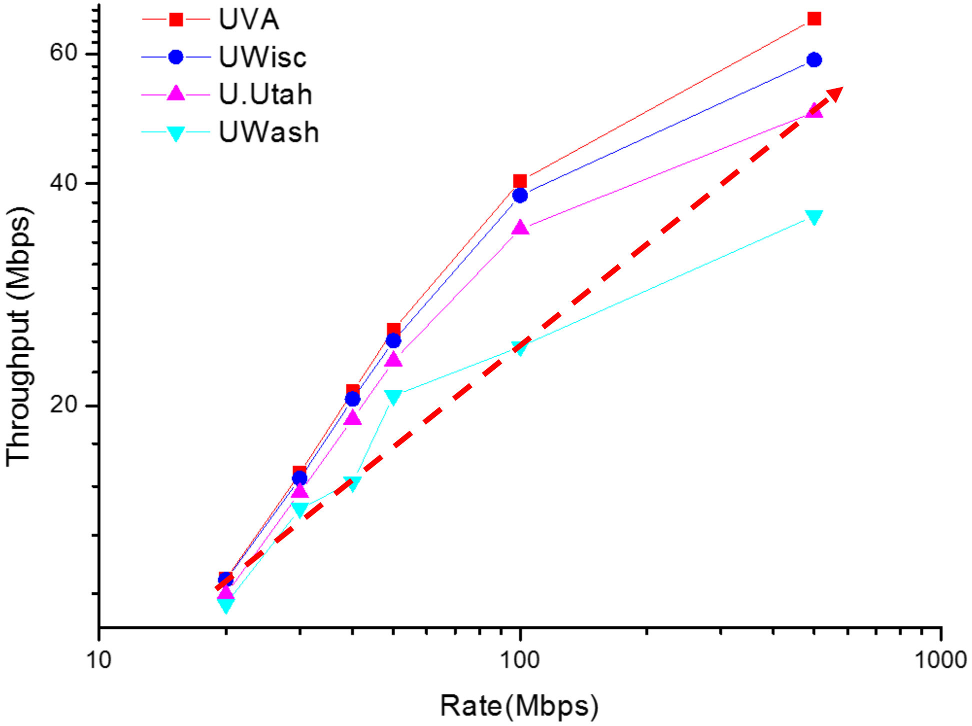
Subscribers	UVA	UMD	UWisc	UWash	UCSD	Utah
Number of FMTP-received files	45642	45642	45642	45642	45642	45642
File-count-based FFDR $\mathbb{F}_t^{count}$	100%	100%	100%	100%	100%	100%
Size-based FFDR $\mathbb{F}_t^{size}$	100%	100%	100%	100%	100%	100%
Number of files that needed FMTP retransmissions	3	3	3	6	4	64
Number of FMTP block retransmissions	21	21	21	34	24	529
Multicast Packet Loss Rate (MPLR) $\mathbb{L}_t^{mc}$	3.5e-4%	3.5e-4%	3.5e-4%	5.6e-4%	4.0e-4%	8.8e-3%
Average throughput of FMTP-received files (Mbps) $\mathbb{T}_t^{fmtp}$	20.92	21.08	20.43	13.81	18.03	19.17
Average throughput of multicast-itself-sufficient files (Mbps) $\mathbb{T}_t^{mc}$	20.93	21.08	20.44	13.83	18.04	19.31
Average throughput of FMTP-retx-needed files (Mbps) $\mathbb{T}_t^{retx}$	0.36	0.35	0.33	0.11	0.24	0.23

1. Our solution worked well and delivered 100% of files without requiring the LDM6-backstop mechanism.
2. Few multicast packets lost during the multicast, and our retransmission mechanism can handle it; throughput is lower, however.
3. Different subscribers achieved different throughput, due to their various propagation delay to the publisher.

LDM7 Performance (Cont.)

- NGRID, 2020-07-12 03:00-04:00 UTC to UVA, UWisc, U.Utah, and UWash

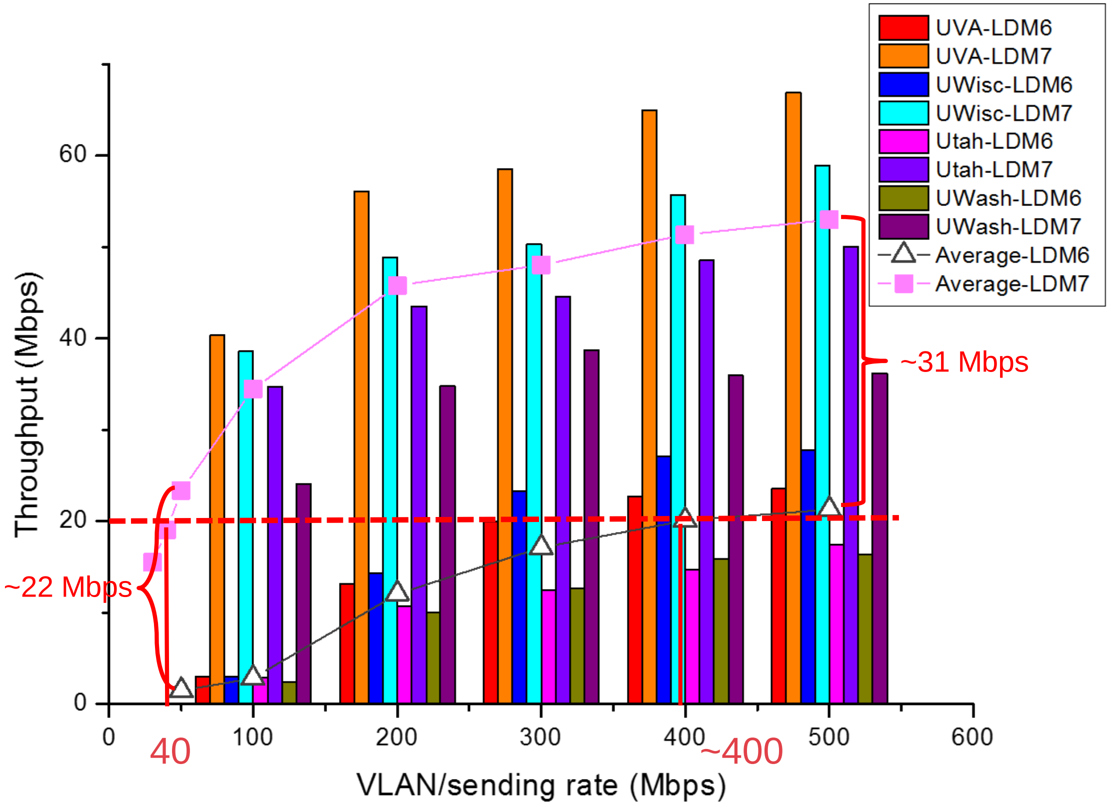
emission	sender queueing	propagation	retx	processing



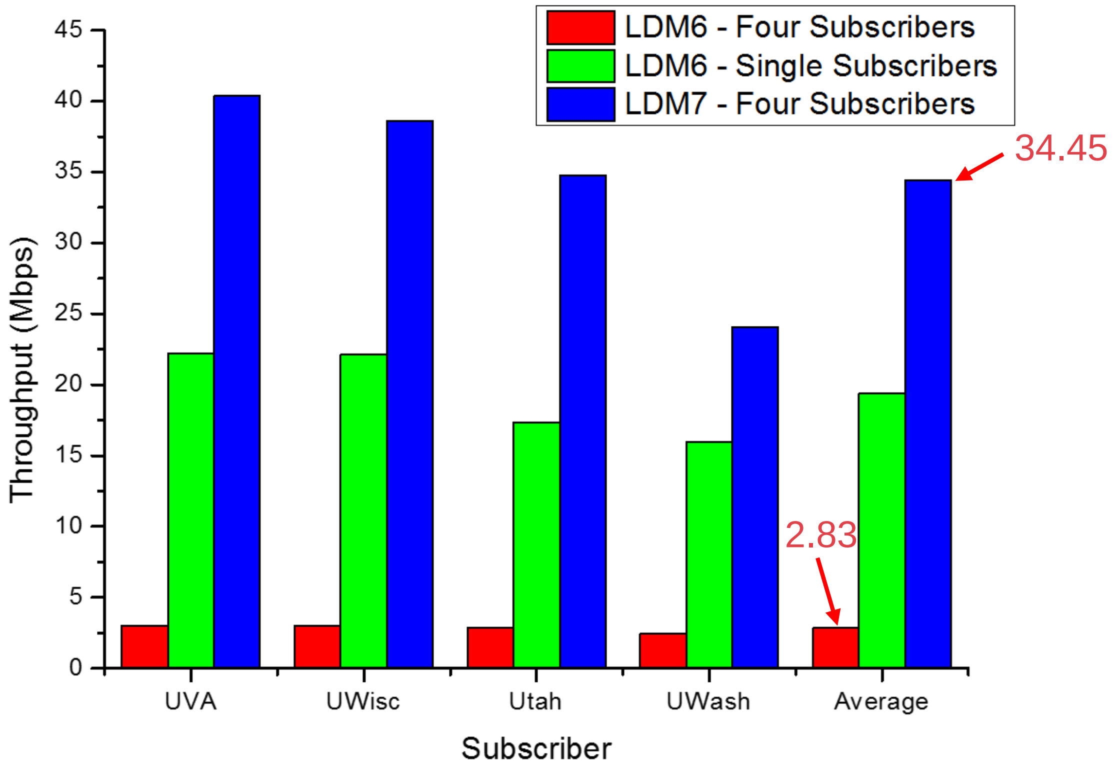
Throughput vs. VLAN/sending rate

Performance comparison between LDM6 & 7

- NGRID, 2020-07-12 03:00-04:00 UTC to UVA, UWisc, U.Utah, and UWash



Throughput vs. sending/VLAN rate



Zoom in the sending rate of 100 Mbps

Outline

- Contributions
- Background
- Cross-layer architecture & LDM7
- LDM7 performance monitoring system
- Multi-domain trial deployment
- Experimental Evaluation
- Conclusions

Conclusions

- Feasible to deploy a network multicast solution leveraging L2 VLAN/MPLS network service
- The LDM7 performance monitoring system with the key LDM7 performance metrics worked well
- LDM7 presented its advantages compared LDM6 with higher throughput at the same sending rate, and bandwidth savings when achieve same performance