

High-Touch Analysis

Tristan Sullivan
LHCONE Meeting Manchester
March 19/25

Randall Sobie
HEPNET, University of Victoria

High-Touch Services

- High-precision, real-time visibility into network traffic
 - Process every packet of interest in real-time
 - Accurate, precision timing (ns precision / accuracy)
 - Software-defined functionality
 - Programmatically deployable and customizable
- In contrast to “low touch” services
 - Fixed function services such as IP packet routing, basic statistics
 - Optimized for speed and low cost, but not flexible
- Technology enablers
 - Software-defined networking
 - Programmable network dataplane hardware with accurate timestamps
 - High-speed packet processing libraries (DPDK, etc.)

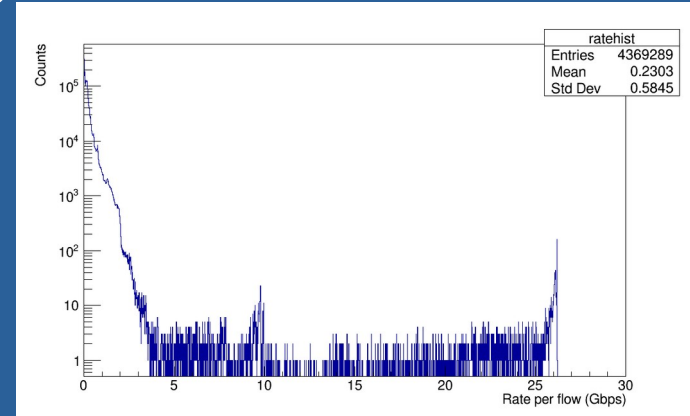


DC24 Analysis

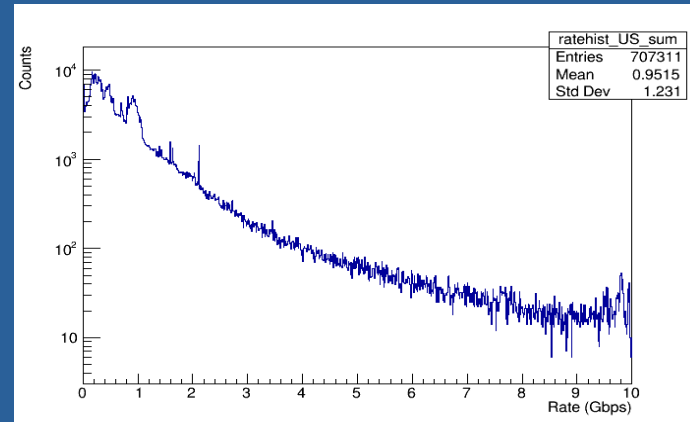
Presented analysis of DC24 high-touch data at previous LHCON meeting (Beijing, Oct. 2024): [Slides](#)

Measured per flow bandwidth ~200 Mbps perhaps lower than expected. Explanations:

- Servers sending/receiving multiple flows simultaneously
- Other flows mixed in with data transfers, e.g. job management, control channels, streaming to worker nodes
- Correlated with rtt? Couldn't be easily checked because rtt information wasn't present in this dataset



Bandwidth per flow



Bandwidth per endpoint (US only)

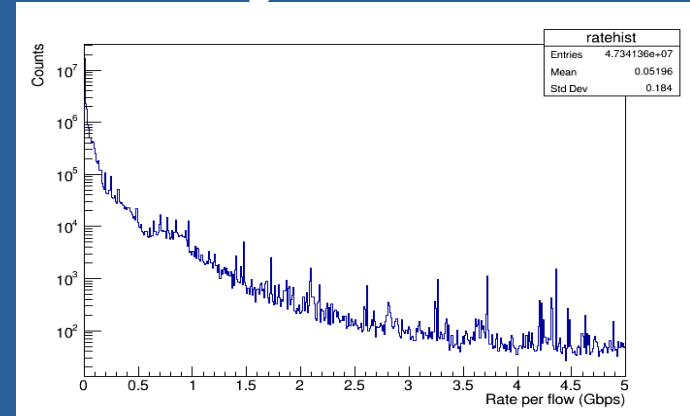
November 2024 Analysis

A new dataset was prepared from data taken on Nov. 6/24

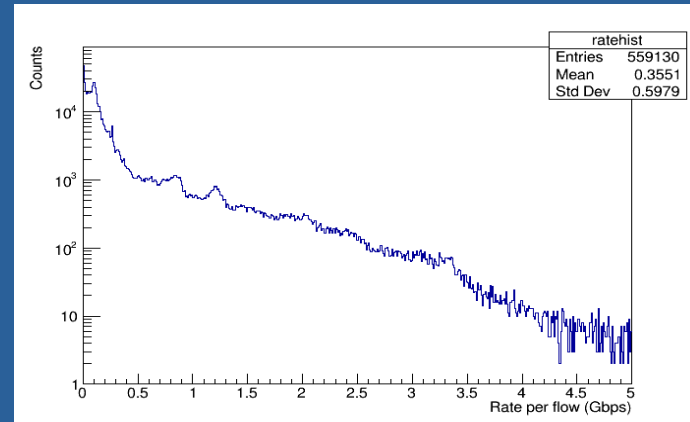
Capability to measure rtt was added

Like previous analysis, select flows that transmit > 100 MB total and remove iperf3 (port 5201)

This turns out to be a very small fraction of the total, about 0.4%

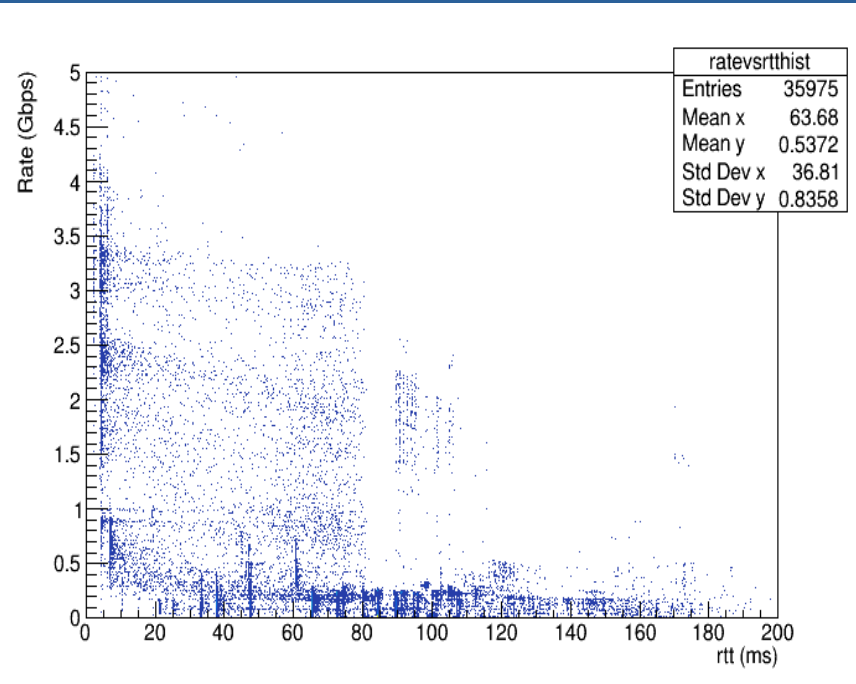


Bandwidth per flow (no size cut, first six hours)



Bandwidth per flow (> 100 MB)

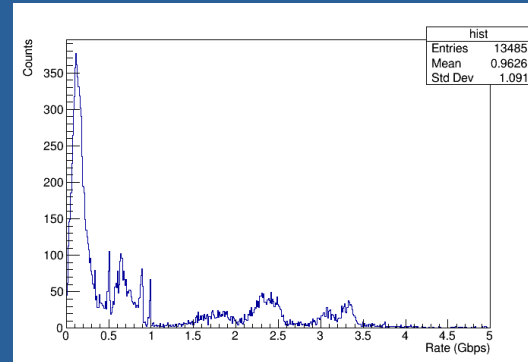
Correlation with RTT



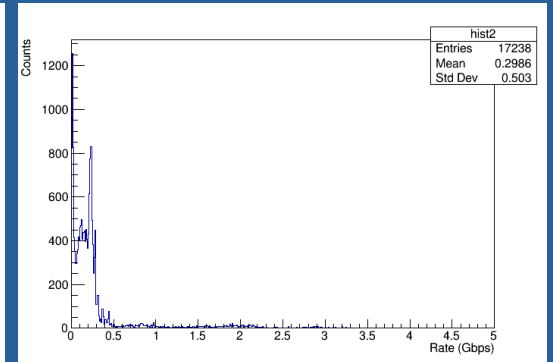
TAKE THIS WITH A TABLESPOON OF SALT

Note the number of entries: < 10% of flows have rtt info

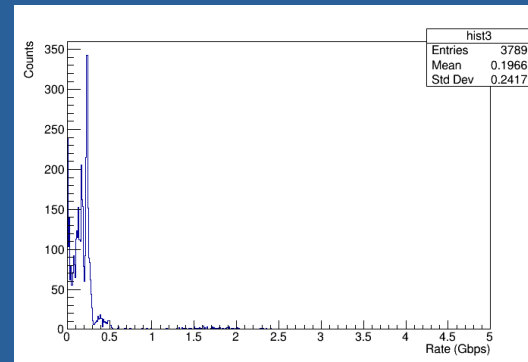
Reason unclear at this time



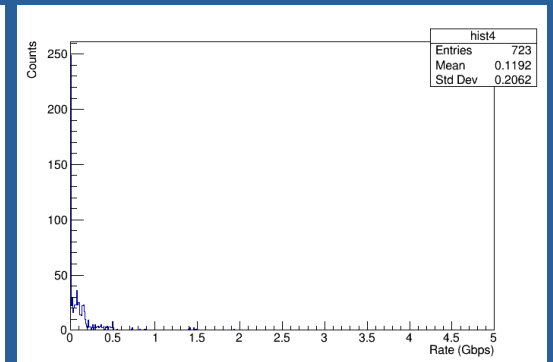
rtt < 50 ms



50 ms < rtt < 100 ms

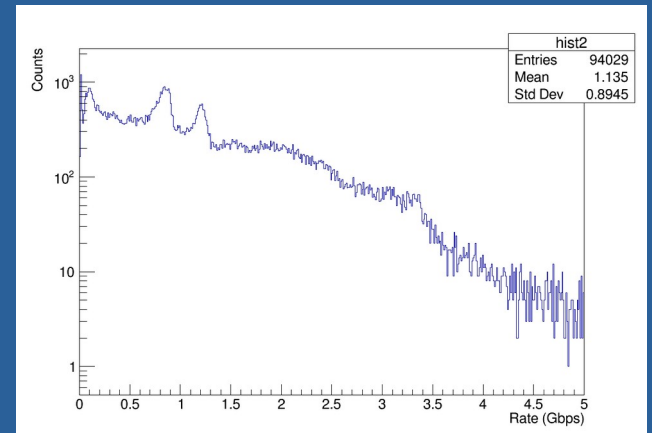
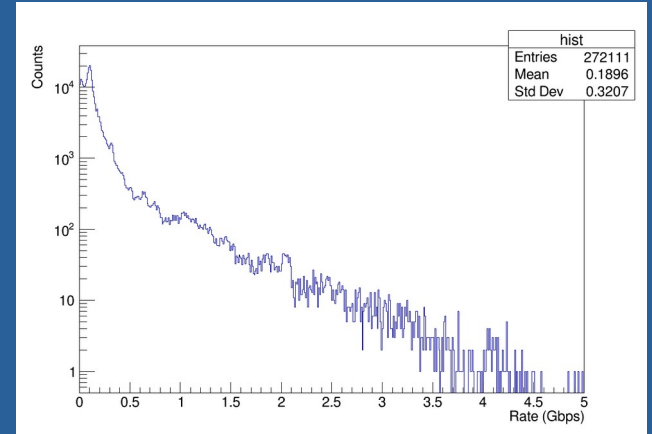
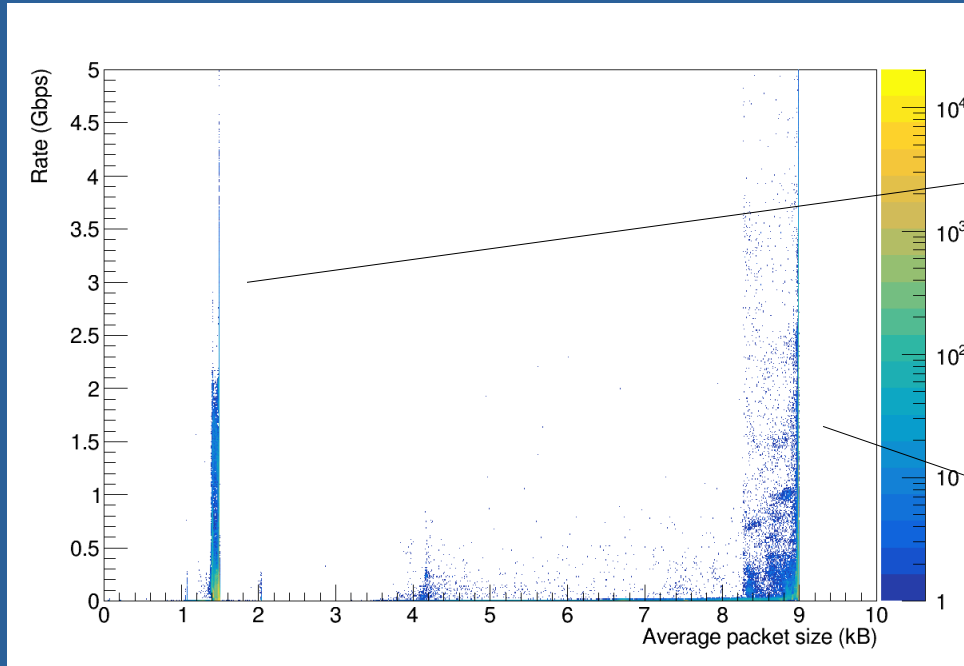


100 ms < rtt < 150 ms



150 ms < rtt < 200 ms

Packet Size



Selecting the two “stripes” may be a good way to select for real data transfers

Large impact from MTU, but must be careful about causality; data largely come from a fairly small number of sites

Top Sending and Receiving Orgs

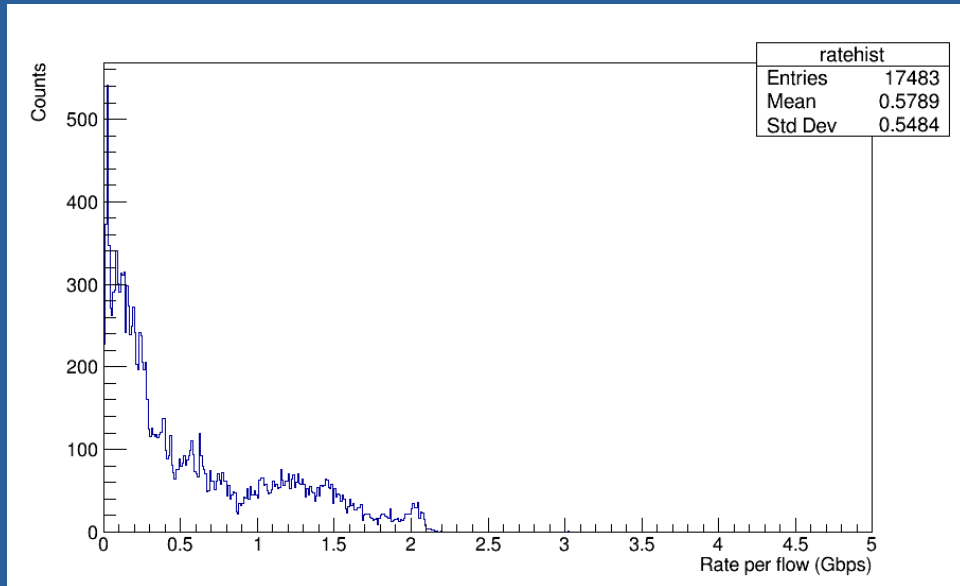
Organization	Flows (source)
U. Chicago	150,276
MIT-Gateways	80,758
FNAL	48,524
BNL	43,225
Indianagigapop	34,626
TRIUMF	15,836
KEK	15,748
University of Tokyo	3,360
Tallinn (Estonia)	1,505
NIKHEF	734

Site	Flows (destination)
FNAL	77,812
U. Chicago	46,707
CERN	43,334
GARR	37,833
UNL	32,571
JISC	31,640
BNL	26,650
U. Wisc. Mad.	20,935
	6,423
NORDUnet	1,031

Total: 559,130

These are the organizations reported in the data – some represent multiple sites (like Indianagigapop), most are one site as far as I can tell

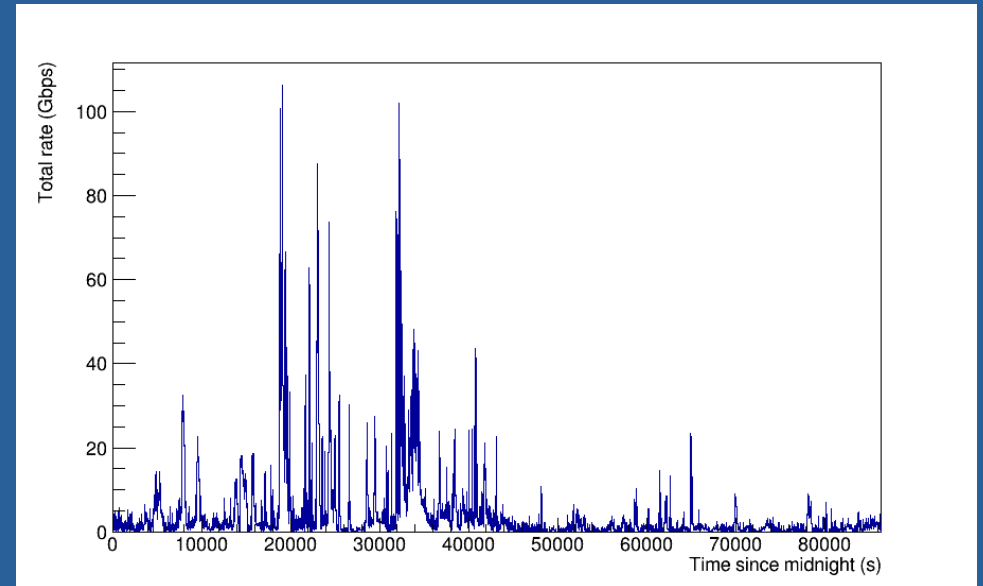
BNL dCache as Source



Average is higher, but still peaked at low bandwidth

Examine flows under 50 Mbps:

- About 25% overseas
- Remainder in US, mainly to nearby Tier-2 xcache
- Interestingly, this endpoint also shows up when looking at flows over 1 Gbps – slow flows not real data transfers?



Total bandwidth sent by BNL dCache

Conclusion

- Need more coverage of rtt to say anything meaningful, but available evidence suggests that intercontinental transfers are much slower; are some servers using default TCP buffers?
- MTU 9000 seems to make a big difference, but have to be cautious
- My feeling remains that the low bandwidth per flow is explained by some combination of three factors: servers performing many transfers simultaneously; transfers between servers with large rtt and possibly small TCP buffers; and flows that are not actual data transfers between storage servers