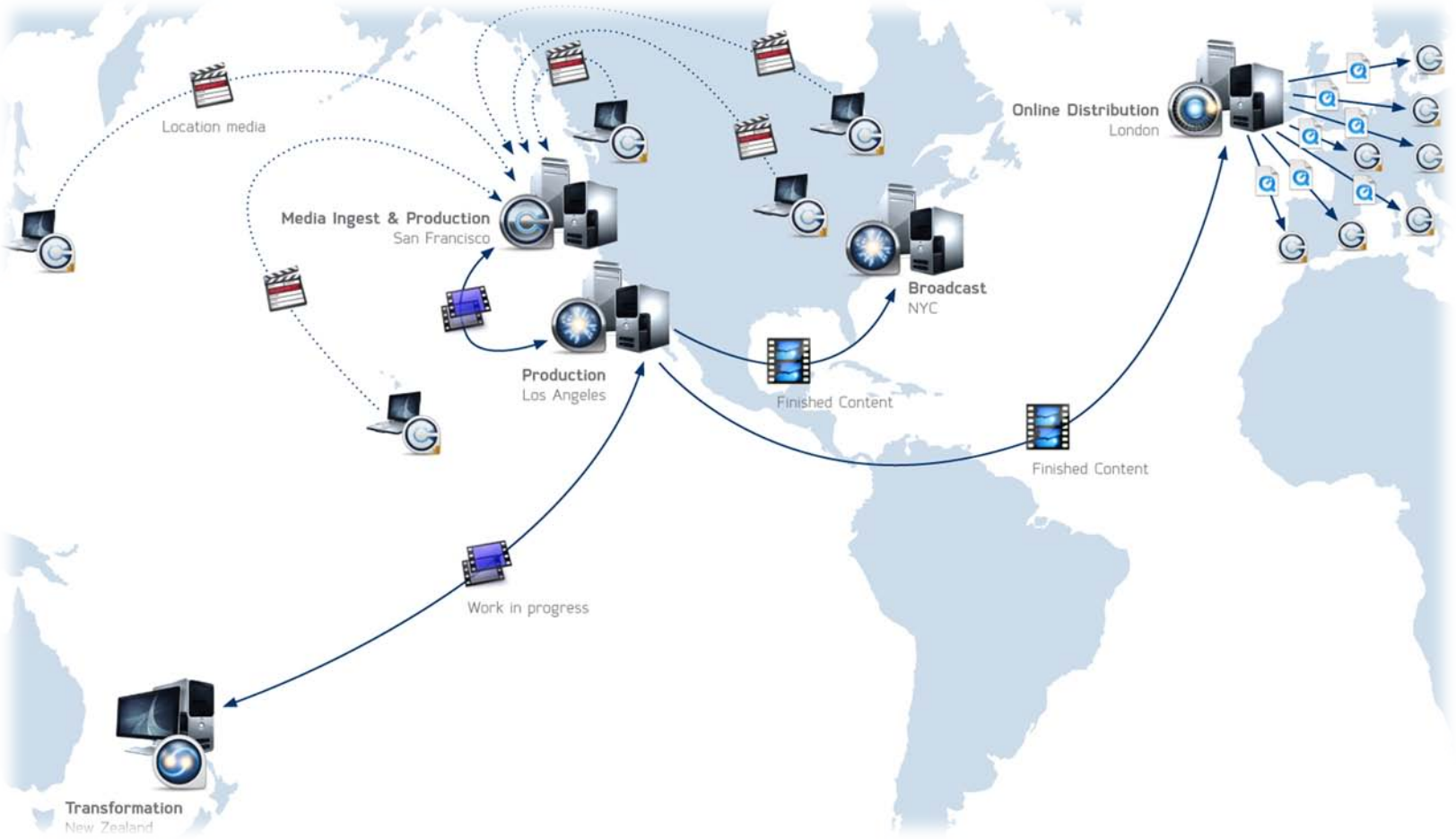
Ultra high-speed transmission technology for wide area data movement

Michelle Munson, president & co-founder
Aspera

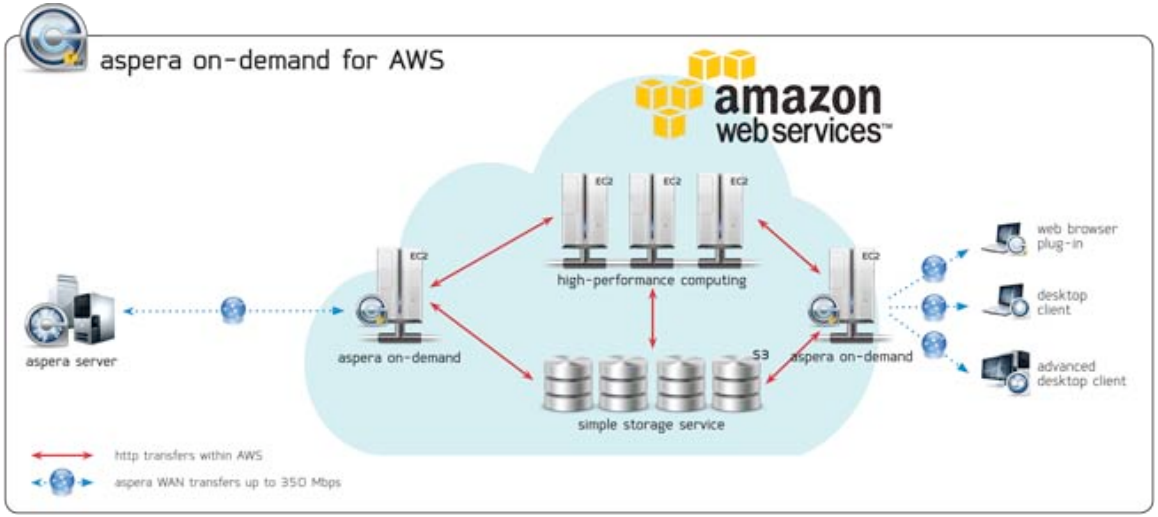
Outline

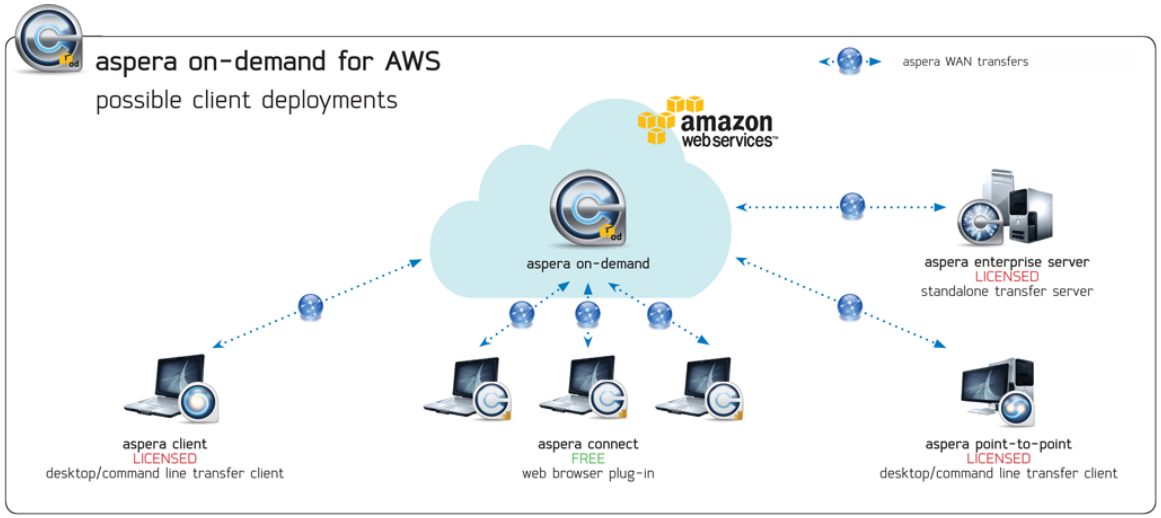
- Business motivation
  - Moving ever larger file sets over commodity IP networks (public, private, cloud)
- Technology obstacles
  - Underlying problems with standard TCP
  - Example: Copying a 1GB file over an NFS mount between AWS nodes in Europe and USA runs at <5Mbps and takes more than an hour to finish the task – 1/100th of the bandwidth capacity
- Aspera *fasp™*
  - Next-generation reliable bulk data transport
- High speed WAN throughput reveals new “last foot” bottleneck
  - From end system computer to storage

Large Data Movement Underpins the Supply Chain



Magnified To, From, and Within the Cloud





Transmission Control Protocol (TCP)

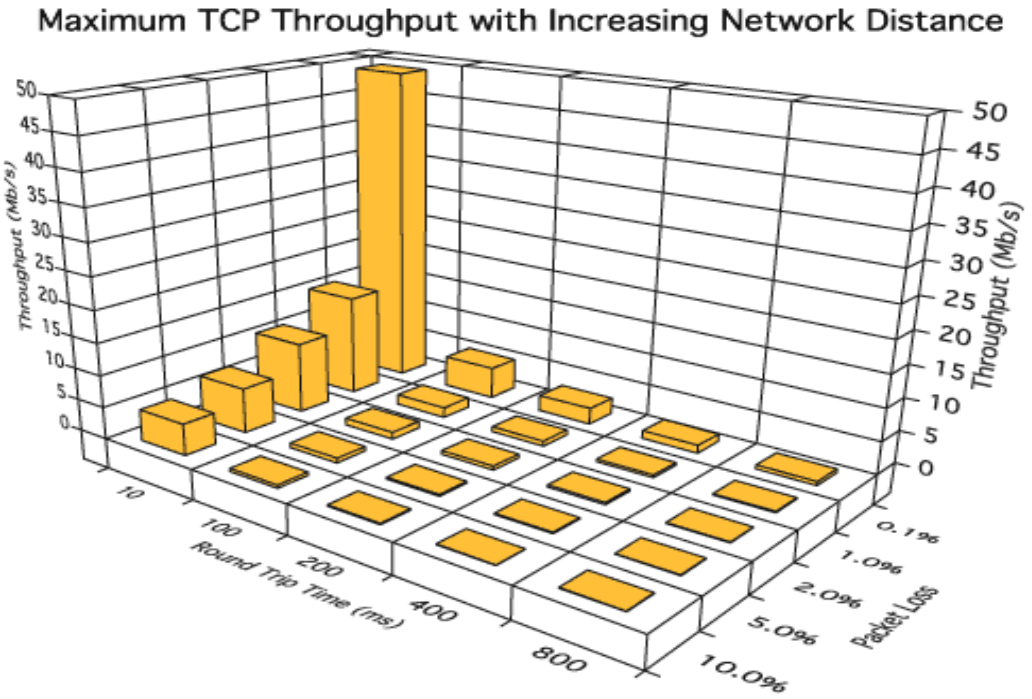
- TCP is the reliable transport that has traditionally powered file-based data transport through protocols such as FTP, HTTP, NFS, CIFS
- TCP has well-known bottlenecks on networks with high round-trip time (RTT) and packet loss rates, and most pronounced on high-bandwidth networks
- Results from the AIMD ("Additive-Increase Multiplicative-Decrease") congestion avoidance algorithm
- AIMD, by design, increases its rate until loss occurs, indirectly causing self-induced slow down
- Additionally, other sources of packet loss (such as wireless, satellite, etc.) equally affect TCP and cause the AIMD algorithm to artificially reduce its transmission rate
- This LOSS-BASED CONGESTION AVOIDANCE has a deadly impact on file transfer throughput on typical WANs

Transmission Control Protocol (TCP)

$$Throughput \propto 1/\left(RTT * \sqrt{p}\right)$$

$RTT =$ *round trip time*

$p =$ *packet loss rate*



- **Example:**

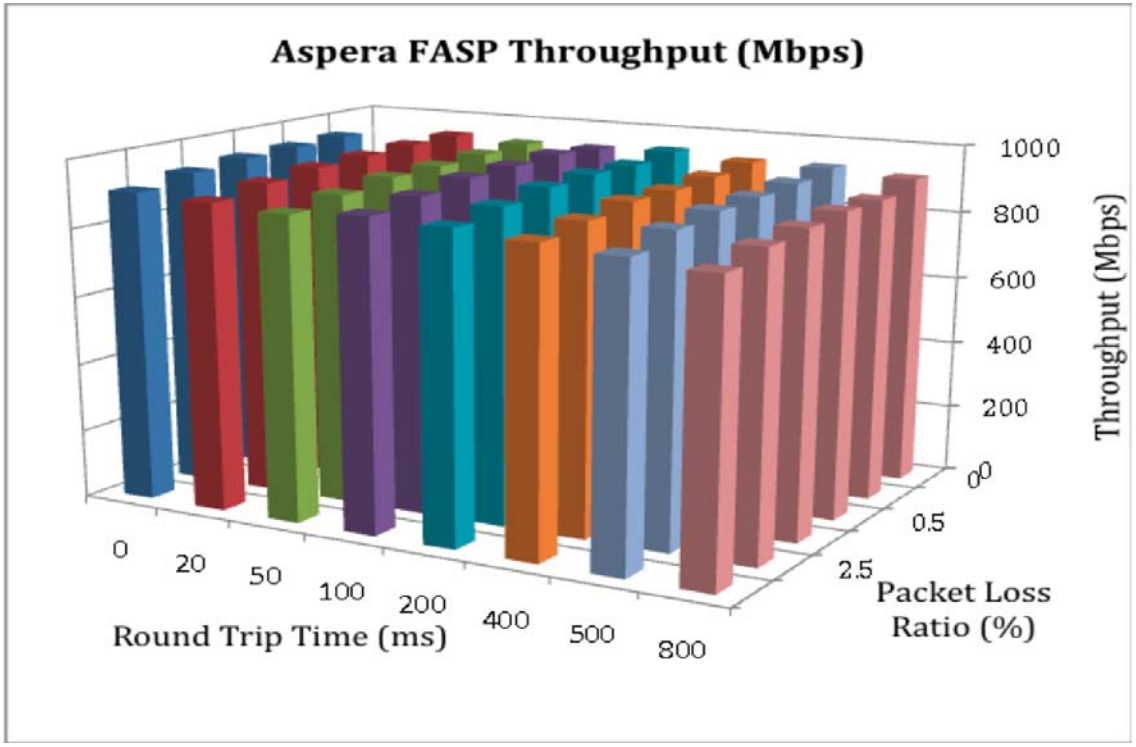
  Transfer on a WAN with 100 ms round trip time and 1% packet loss rate, assuming 1500 byte packets has a throughput of < 2 Mbps *regardless of capacity*

fasp capabilities

- A reliable bulk data transport protocol that completely separates reliability and rate control
  - Uses standard UDP in the transport layer
  - Uses a theoretically optimal approach that retransmits precisely the real packet loss on the channel
- Arbitrarily high speed
  - New packets need not slow down for the retransferring of lost packets as in TCP-based byte streaming applications
- Zero receiving cost
  - Lost data is retransmitted at the available bandwidth inside the end-to-end path, with nearly zero duplicate retransmissions
- Supports intentional bandwidth prioritization
  - Network queuing based rate control provides a virtual bandwidth-tuning handle

fasp throughput over WAN
Local storage, "cheap" computers

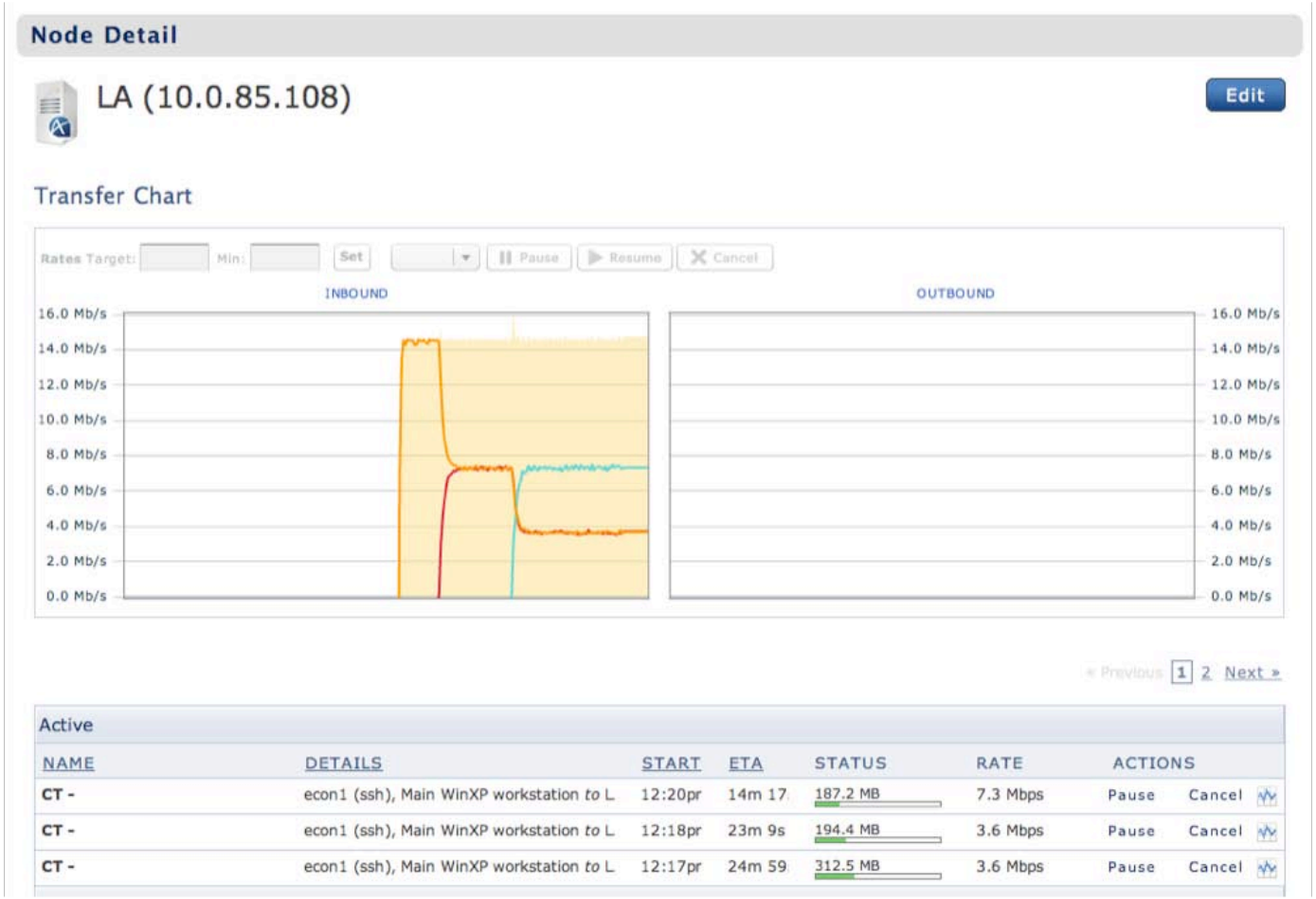
- Transfer throughput for a 5GB file over varied WAN conditions, previous Linux-to-Linux set up with 3x10,000 RPM drives



fasp adaptation to available bandwidth

- Delay-based rate control that uses queuing delay as the primary measure of network or disk-based congestion
- Aims to maintain small, stable queuing (network & disk)
  - Transfer rate adjusts up as the measured queuing falls below the target (indicating that some bandwidth is unused), and down as the queuing increases above the target (congestion is eminent)
- Advantages
  - Avoids artificial slow down on lossy media
  - Allows for stable, fair bandwidth sharing between concurrent transfers and fast ramp up for best delivery times
  - Quickly converges to a stable equilibrium with target queuing at bottleneck, bringing “QOS” experience to transfer applications
  - Virtually no sending cost introduced to the network
  - Built-in response to network queuing is a priority handle

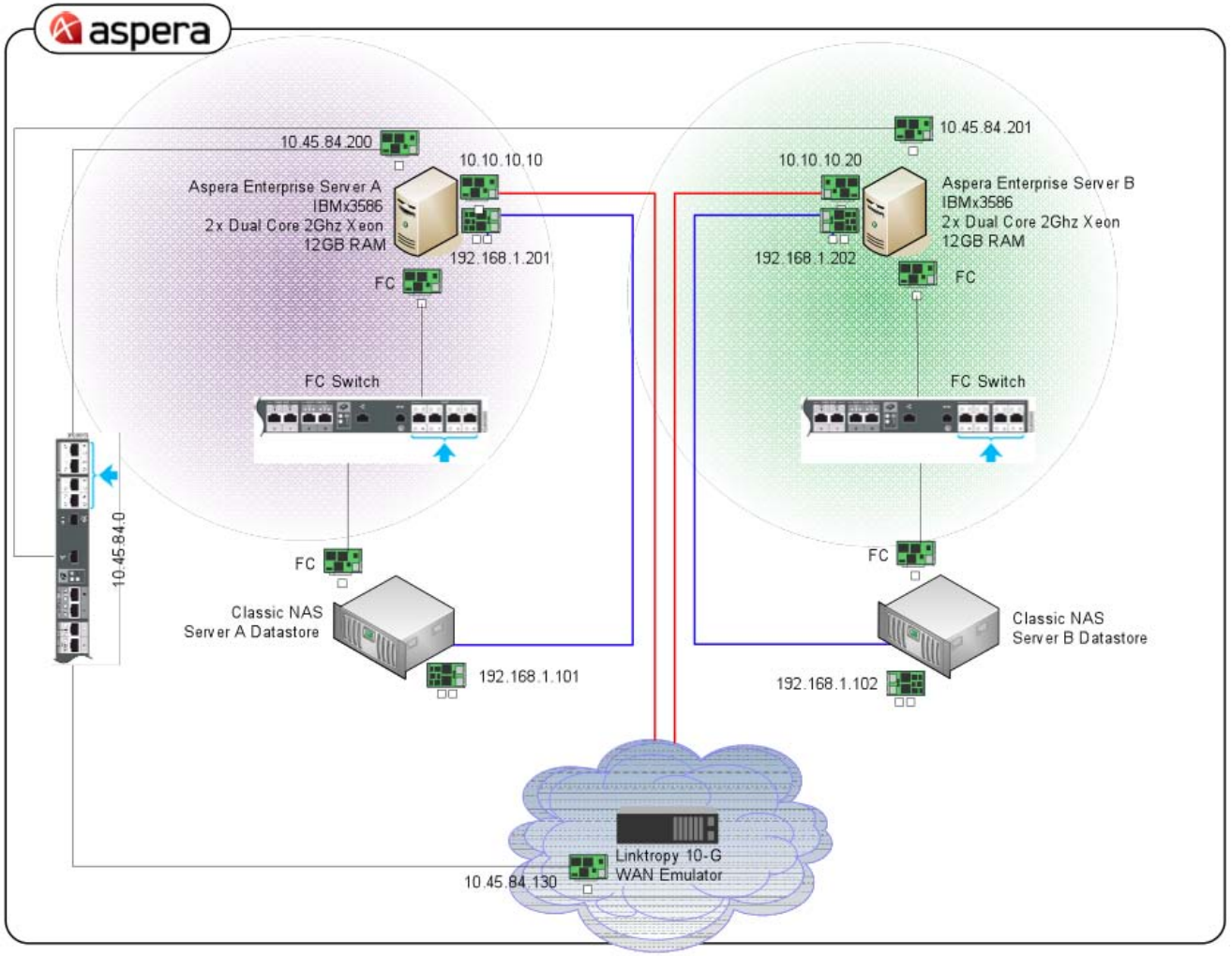
fasp Rate Control – Concurrent Flows and Bandwidth Prioritization



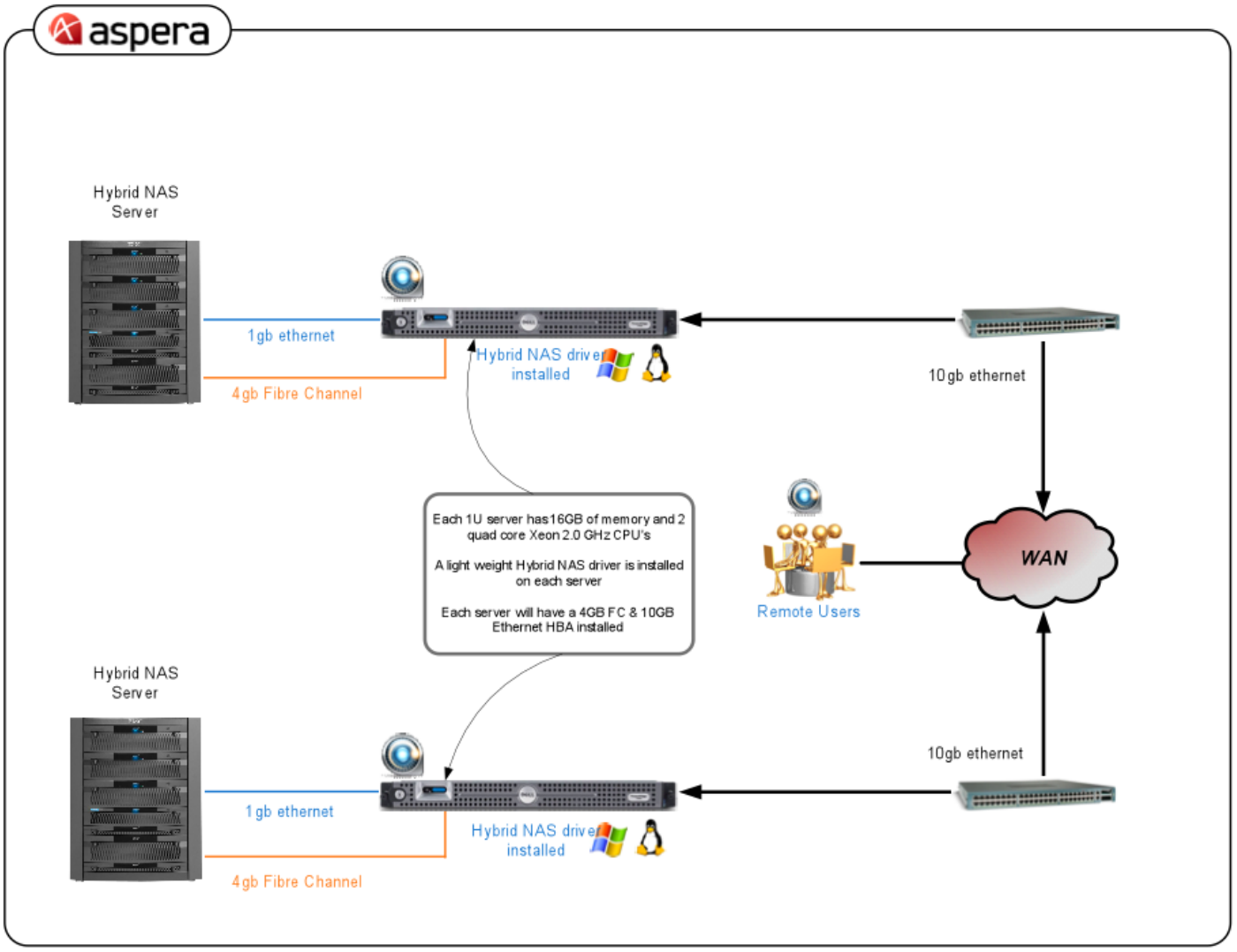
A new bottleneck to high-speed WAN transfers: the "last foot" from computer to storage

- With WAN transport bottlenecks eliminated and network capacities >1Gbps, a new bottleneck emerges
- FAST network with SLOW data storage, including computer, I/O bus, and network file systems
- Especially pronounced where 10 Gbps WANs are available
- Over the past 12 months, Aspera has benchmarked the performance of *fasp* over 10 Gbps WANs with leading storage vendors:
  - EMC, HP, Isilon, NetApp, BlueArc, Panasas
- Goals:
  - Verify that ***fasp*** can eliminate the WAN transport bottleneck for **worst-case global Internet conditions**
  - Verify capability and best practices for storage system to sustain high throughput over the "last foot"

Test Setup #1 – Classic NAS
NFS over LAN, *fasp* over WAN



Test setup #2 – Hybrid NAS
Hybrid NFS and Fiber Channel over LAN, *fasp* over WAN



Single large file (80 GB) on 10Gbps WAN
One *fasp* session

(c) Single large file, multi-Gbps WAN performance

Bandwidth (Mbps)	RTT (ms)	PLR (%)	FASP w/ NAS RATE (Mbps)	FASP w/ HNAS RATE (Mbps)	FASP w/ Cluster RATE (Mbps)	Data Size (GB)	SCP (Mbps)	RSYNC (Mbps)	FTP (Mbps)	TCP (Mbps)	1TB by FASP (hours)	1TB by FTP	Speed UP
10000	10	0.1	1670.2	1915.2	920.8	65536	17.6	20	21	49	1.3	2.1 days	39 X
10000	10	1	1649.8	1862.7	926.8	65536	7.2	6.7	8.8	15	1.3	6.8 days	124 X
10000	10	5	1626.2	1858.1	907.4	65536	2	2.7	2.4	7	1.3	14.5days	265 X
10000	100	0.1	1608.9	1813.0	914.4	65536	12	12	14.1	5	1.3	20.4 days	363 X
10000	100	1	1594.5	1805.0	909.9	65536	3.6	4.2	4.1	1.5	1.4	67.9 days	1203 X
10000	100	5	1547.2	1744.7	883.7	65536	1.9	1.6	1.4	0.7	1.4	145.4days	2492 X
10000	300	0.1	1598.9	1784.2	887.3	65536	2.4	2.3	0.4	1.6	1.4	63.6 days	1115 X
10000	300	1	1595.2	1744.4	890.1	65536	0.64	0.664	0.512	0.5	1.4	203.6days	3489 X
10000	300	5	1552.6	1697.7	876.2	65536	0.192	0.224	0.3	0.23	1.4	442.6days	7381 X

Single large file (80GB) on 10Gbps WAN Multiple *fasp* Sessions

Table 1. FASP throughput for a single large file on NAS, Hybrid NAS (HNAS) and Cluster over varied WAN conditions with 4-8 sessions

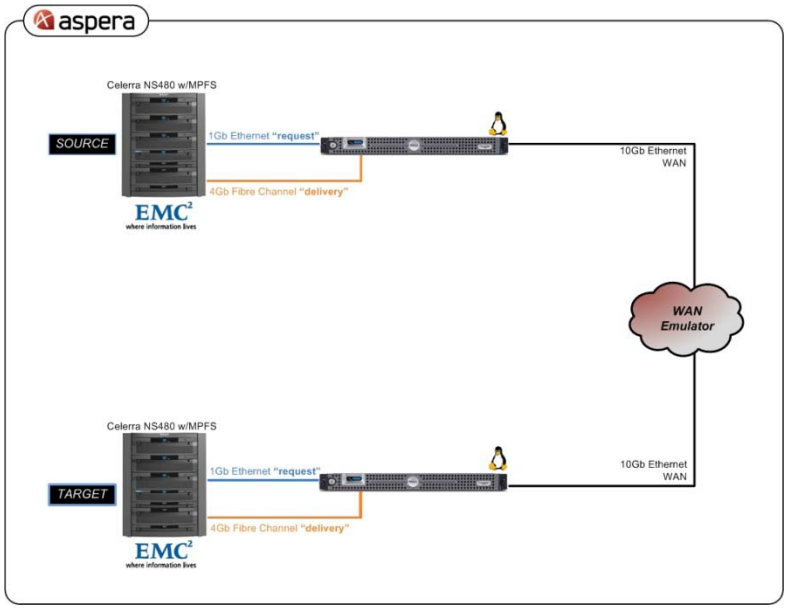
(a) Single large file, multi-Gbps WAN performance

Bandwidth (Mbps)	*RTT (ms)*	*PLR (%)*	*FASP w/ NAS RATE (Mbps)*	*FASP w/ HNAS RATE (Mbps)*	*FASP w/ Cluster RATE (Mbps)*	*Data Size (GB)*	*SCP (Mbps)*	*RSYNC (Mbps)*	*FTP (Mbps)*	*TCP (Mbps)*	*1TB by FASP (hours)*	*1TB by FTP*	*Speed UP*
10000	*10*	*0.1*	*2694.0*	*3143.5*	*2007.2*	*85871*	*17.6*	*20*	*21*	*49*	*0.8*	*2.1 days*	*64 X*
10000	*10*	*1*	*2742.1*	*3122.6*	*1978.2*	*85871*	*7.2*	*6.7*	*8.8*	*15*	*0.8*	*6.8 days*	*208 X*
10000	*10*	*5*	*2700.8*	*3072.6*	*N/A*	*85871*	*2*	*2.7*	*2.4*	*7*	*0.8*	*14.5 days*	*439 X*
10000	*100*	*0.1*	*2678.1*	*3066.9*	*1973.4*	*85871*	*12*	*12*	*14.1*	*5*	*0.8*	*20.4 days*	*613 X*
10000	*100*	*1*	*2672.2*	*2996.7*	*1943.2*	*85871*	*3.6*	*4.2*	*4.1*	*1.5*	*0.8*	*67.9 days*	*1998 X*
10000	*100*	*5*	*2626.9*	*3006.2*	*N/A*	*85871*	*1.9*	*1.6*	*1.4*	*0.7*	*0.8*	*145.4 days*	*4295 X*
10000	*300*	*0.1*	*2656.5*	*3010.2*	*N/A*	*85871*	*2.4*	*2.3*	*0.4*	*1.6*	*0.8*	*63.6 days*	*1881 X*
10000	*300*	*1*	*2658.1*	*3009.3*	*N/A*	*85871*	*0.64*	*0.664*	*0.512*	*0.5*	*0.8*	*203.6 days*	*6019 X*
10000	*300*	*5*	*2602.1*	*2992.1*	*N/A*	*85871*	*0.192*	*0.224*	*0.3*	*0.23*	*0.8*	*442.6 days*	*13009 X*

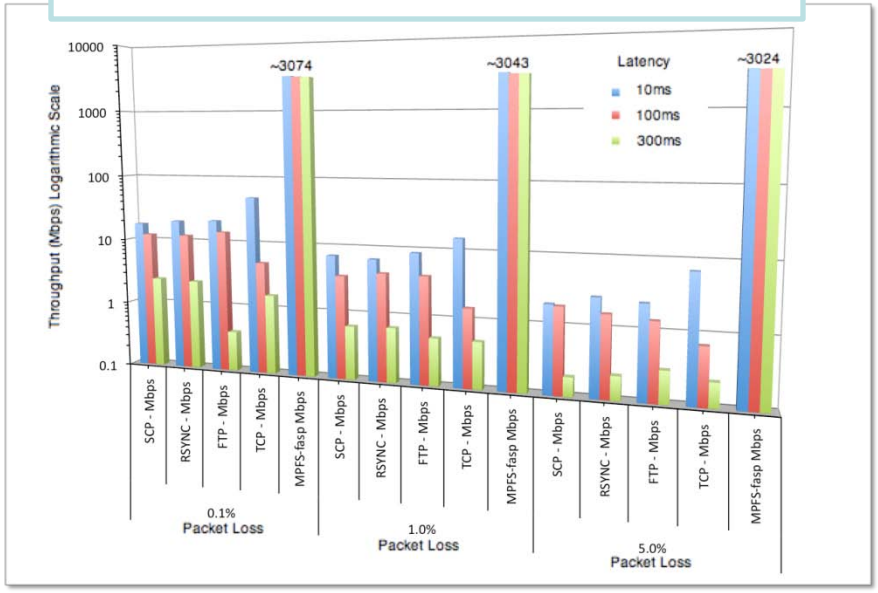
EMC Celerra and Ultra High-speed *fasp* Transport

EMC Celerra and Multi-Path File System (MPFS), with Aspera *fasp* – High Speed Bulk-Data Transfer Solution

- **Aspera + EMC = joint ultra high-speed file transfer and storage solution**
  - Multi-Gbps movement of large data over global WANs
  - Transfer uncompressed HD video, digital cinema or 3D film files around the world in minutes, using commodity hardware and public Internet.
- **High-Performance NAS Test Environment**



Throughput of FTP, SCP & RSYNC, standalone TCP, and the joint MPFS-*fasp* solution, which shows a 1250x gain



Key best practices learned
Maximizing the "Last Foot" I/O throughput

- Separating Network WAN and Storage I/O interrupts on separate CPU cores improved throughput 10-15%
- Sender-side storage system – adjusted read ahead queue and I/O depths, as well as read and write cache sizes for maximum throughput
- Receiver-side Linux OS – increased write cache size and the frequency for flushing dirty data
  - Critical to avoid stalls in flushing large chunks of data to disk, a deadly bottleneck in NFS
  - Linux system file caching management policy was configured to reduce vm.dirty_ratio and vm.dirty_background_ratio to eliminate periodic "pausing" of NFS in writing

Summary

- *fasp* provides a next-generation transport for bulk data that fills the gap left by standard TCP
- Eliminates artificial bottlenecks due to
  - imperfect congestion control algorithms,
  - packet losses (by physical media, cross traffic burst, or coarse protocols themselves), and the
  - coupling between reliability and congestion control
- Transmits at the available bandwidth, including fast-scale reaction to disk slowing and a long-scale (on order of RTT) unified delay-based congestion control for network and disk
- Retransmits dropped data with nearly zero bandwidth cost
- Delivers maximum effective file transfer speeds over commodity WANs
- In combination with high-speed local storage systems, offers a next-generation petabyte scale data moving alternative

Thank You

michelle@asperasoft.com

www.asperasoft.com