

ADVANCED WIDE AREA NETWORKING

Matt Crawford, Fermilab Computing Division

CURRENT DEPLOYMENTS



Mass Storage



Disk Cache



Local Analysis



Remote Storage

— 10's of TB/day —> — 100's of TB/day —>

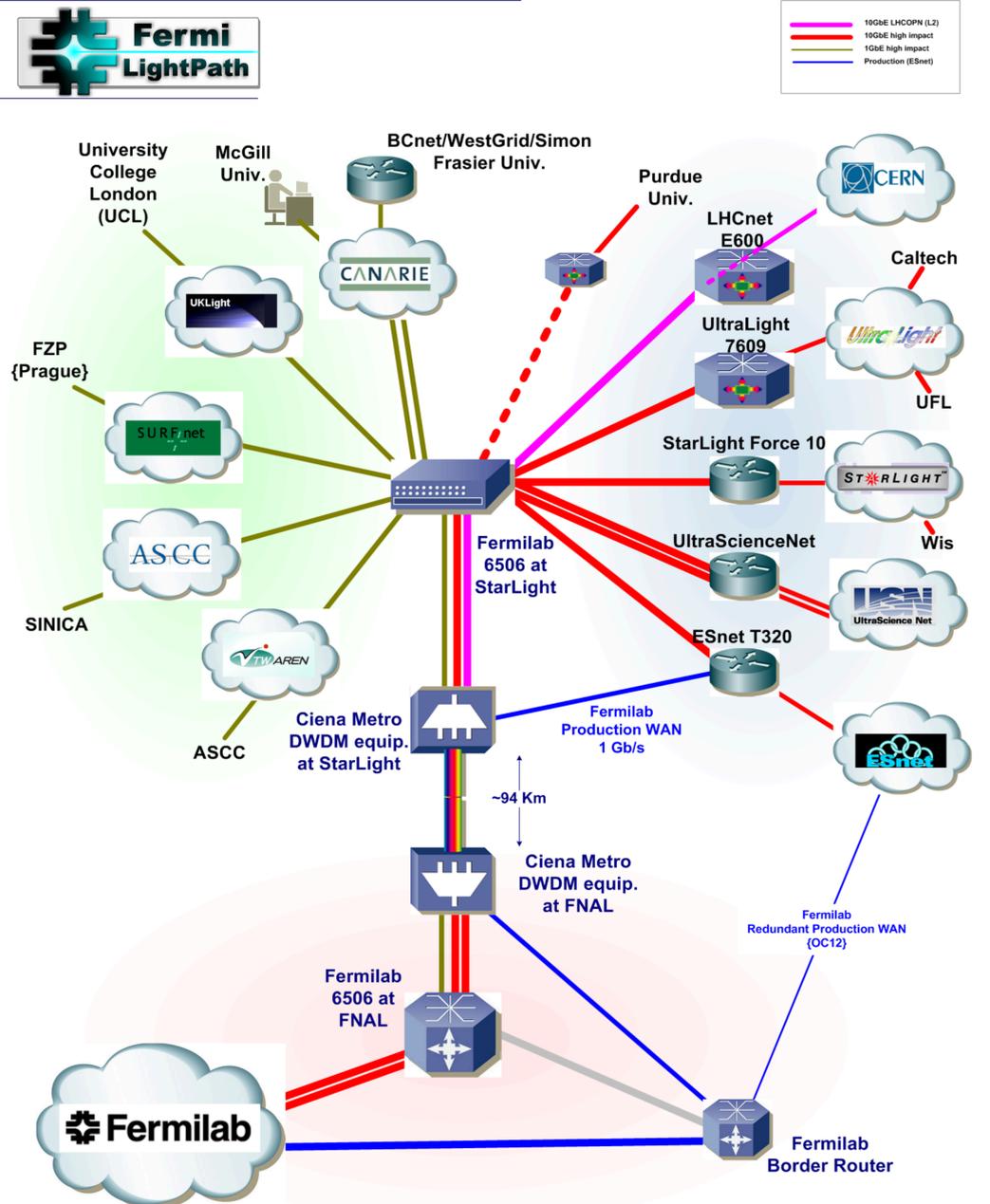
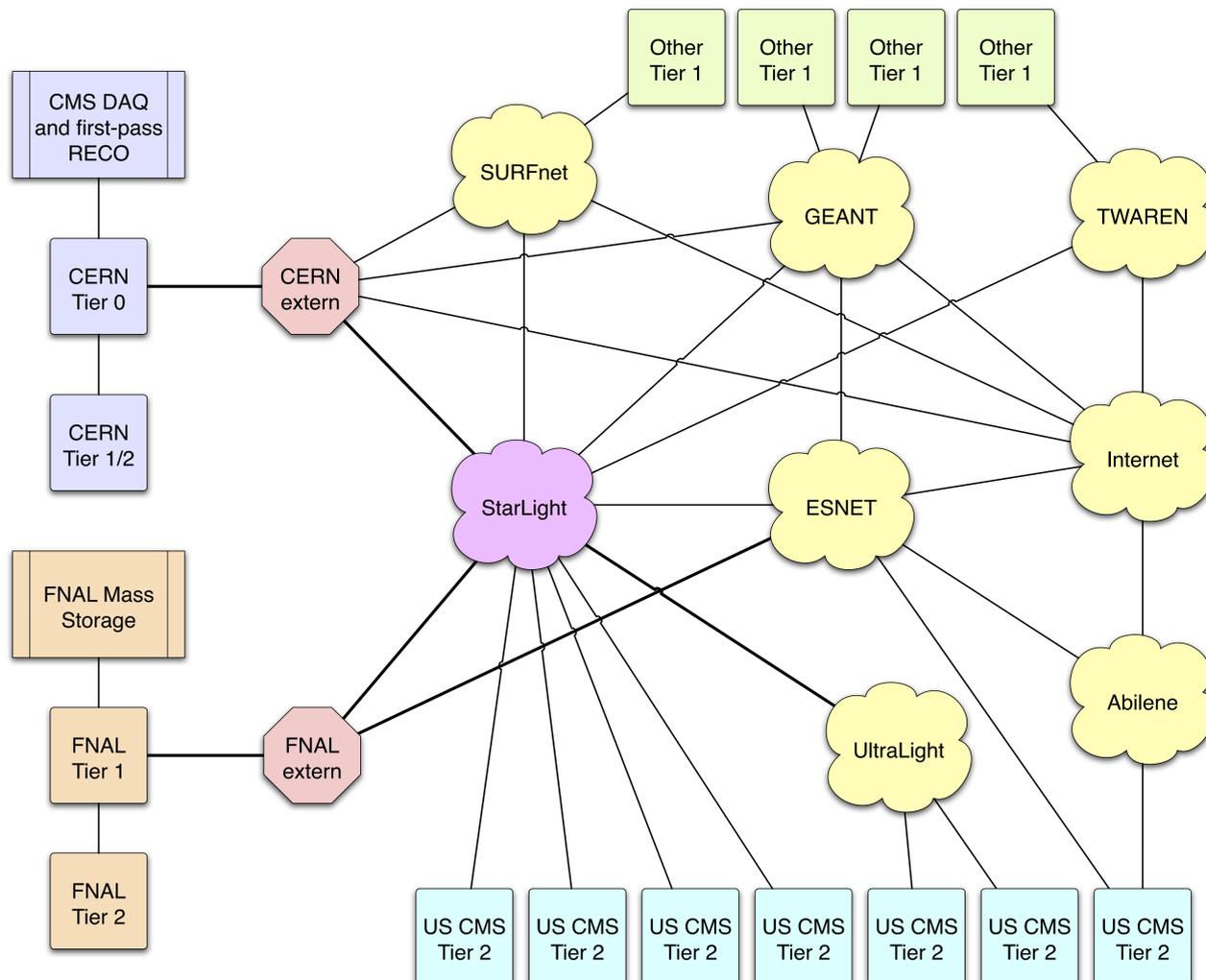
Petabytes (millions of gigabytes) of data are stored in robotic tape storage systems. Local and remote applications access the data through caching disk systems, which serve data directly or request tape mounts as needed.

Data transfer demands to Run-II (CDF & D0) collaborating institutions and from CERN (for CMS preparation) far exceed the capacity of last year's 0.6 Gb/s production link or the interim 1 Gb/s production link.

Fermilab has a dedicated fiber connection to the Optical Exchange Point in Chicago known as *StarLight*. Virtual network links, or *Lightpaths*, to our high-volume peer sites are constructed through Fermilab's switching equipment there.

ESnet, Argonne, and Fermilab are constructing a Chicago Metropolitan Area Network which will provide multiple production and backup wavelengths with 10 Gb/s capacity each, and more wavelengths to ESnet's Science Data Network (SDN) for high-impact applications.

CMS – A Worldwide Network Exercise

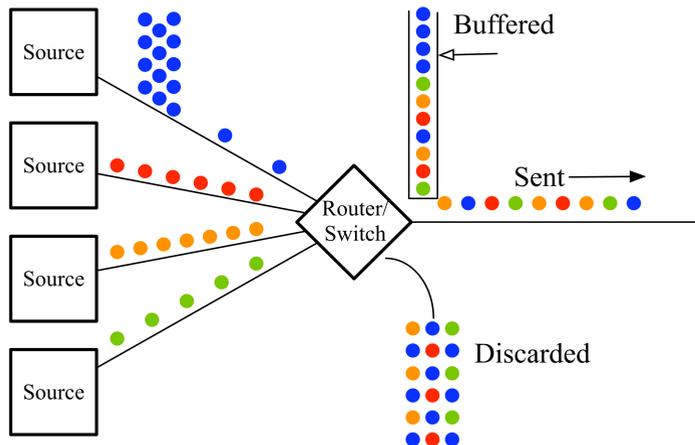


Research and Development in Wide Area Networking

Why Lightpaths?

TCP packet transmission tends toward “bursty” behavior. Every shared link in a path is a potential bottleneck in the network, and has limited buffers at the router or switch ports it connects. Hosts of even moderate CPU and memory capabilities can easily overwhelm the buffer space of an individual port, causing long sequences of dropped packets. This leads to TCP throttling its sending rate to very low speeds.

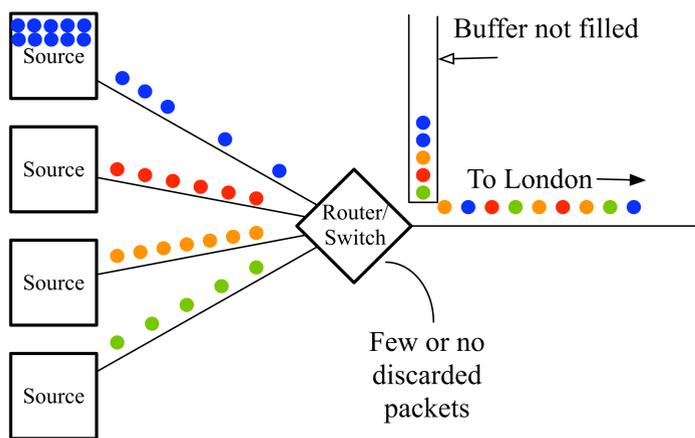
Lightpaths reduce or remove contention on links, reducing the incidence of bursts and making performance more predictable.



Taming Bursts With Host QoS

QoS, or Quality of Service, refers to a set of queuing strategies and packet-pacing rules which can be applied by network elements to deliver a certain level of service. Some operating systems, including recent versions of Linux, can apply the same methods to regulate or “shape” the traffic they send. The QoS rules may vary on a per-destination basis so, for example, data destined for a low-capacity transatlantic link may be shaped without limiting transmissions to other destinations.

We are studying the applicability of host-based QoS to regulating packet flow over international network circuits, with CDF file transfer to University College, London as a test case. Previous performance improvements were gained by doubling the buffer memory at the bottleneck (1 Gb/s) link.



CDF File Servers

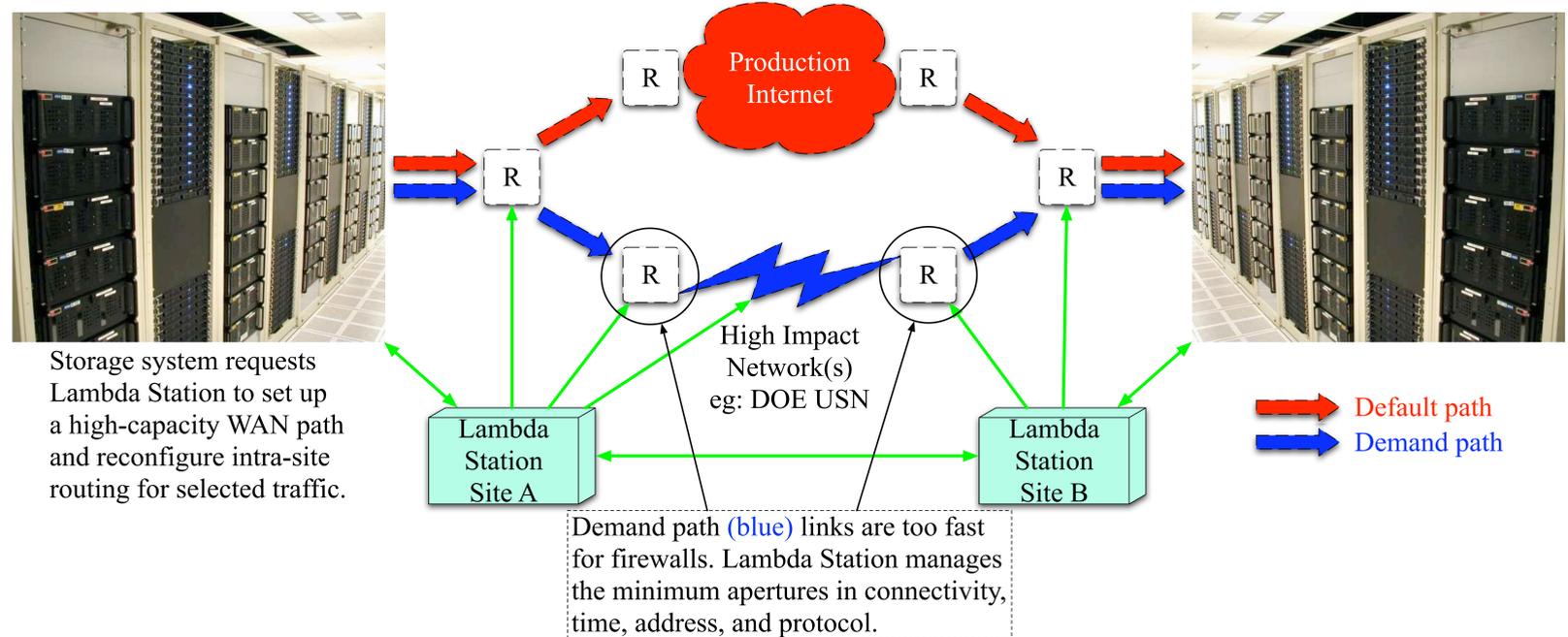
Lambda Station Project

Problem space

- “High impact” data movements benefit from special WAN links
- Other traffic should not be disturbed
- Hosts should not have to know network details
- Site/system network infrastructure should not be replicated!
- Network administrators must retain control of site network

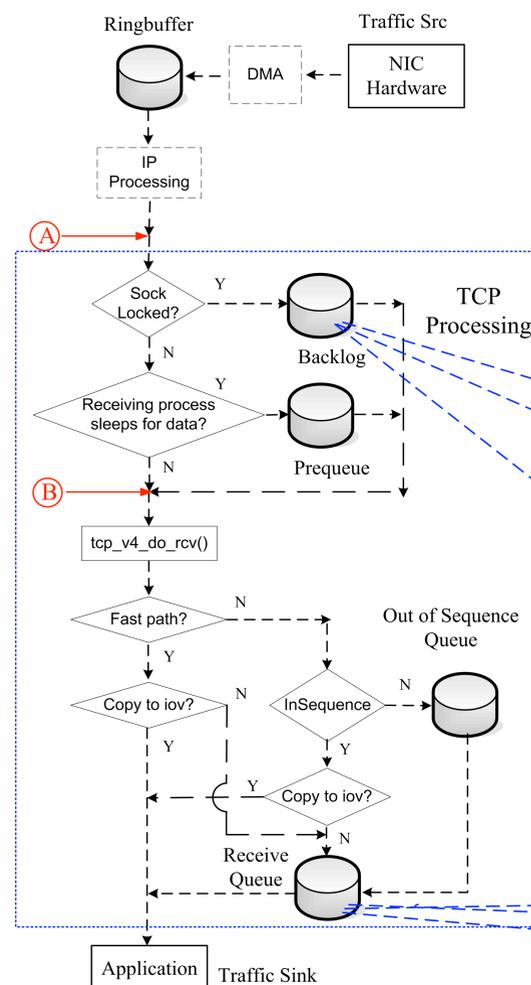
Lambda Station's function

- Authentication, Access control, Accounting
- Allocate network paths for high-impact applications
- Perform any required WAN reservation & setup actions
- Adjust site-edge access and internal site routing
- Steer selected flows on the most granular possible basis



Linux Kernel Bottlenecks

Packet input paths – TCP



Recent versions of the Linux kernel are preemptible, meaning a given process can lose the CPU while it is running inside the operating system. However, there are sections of the TCP network receive code in which the process' data-receiving structures are locked against changes. Packets received while the process is suspended in this section are *not* handled by TCP until the process is resumed – possibly many hundreds of milliseconds later if the system is busy with computational tasks. By that time, the sender assumes the packets are lost. Universities, CMS and ATLAS Tier-2 centers in particular, often use the spare disk space on event analysis farms for file storage. Data sets are received from Tier-1 centers (FNAL, BNL) and served locally. This gives rise to exactly this problem.

Kernel Load	A-B processing delay					
	< 1ms	1-10ms	10-100ms	100-1000ms	>1sec	
Old	0	775091	2778	0	0	
Old	10	108496	1072	1046	304	
New	10	495434	1823	0	0	

