Fermilab

FroNTier/CMS

Lee Lueking

WLCG Workshop DB BoF
22 Jan. 2007

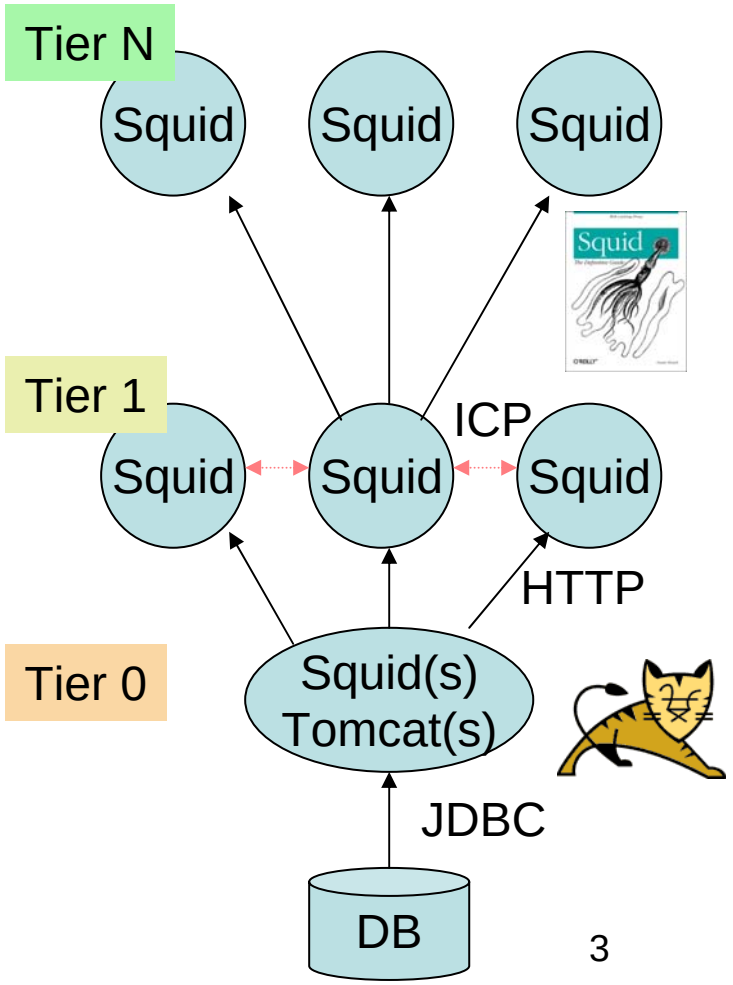
Outline

- Overview
- Features
- Deployment
- Test and performance
- Operational Experience

Acknowledgements
•Barry Blumenfeld (JHU), David Dykstra (FNAL), Eric Wicklund (FNAL)
•POOL/CORAL team

Overview

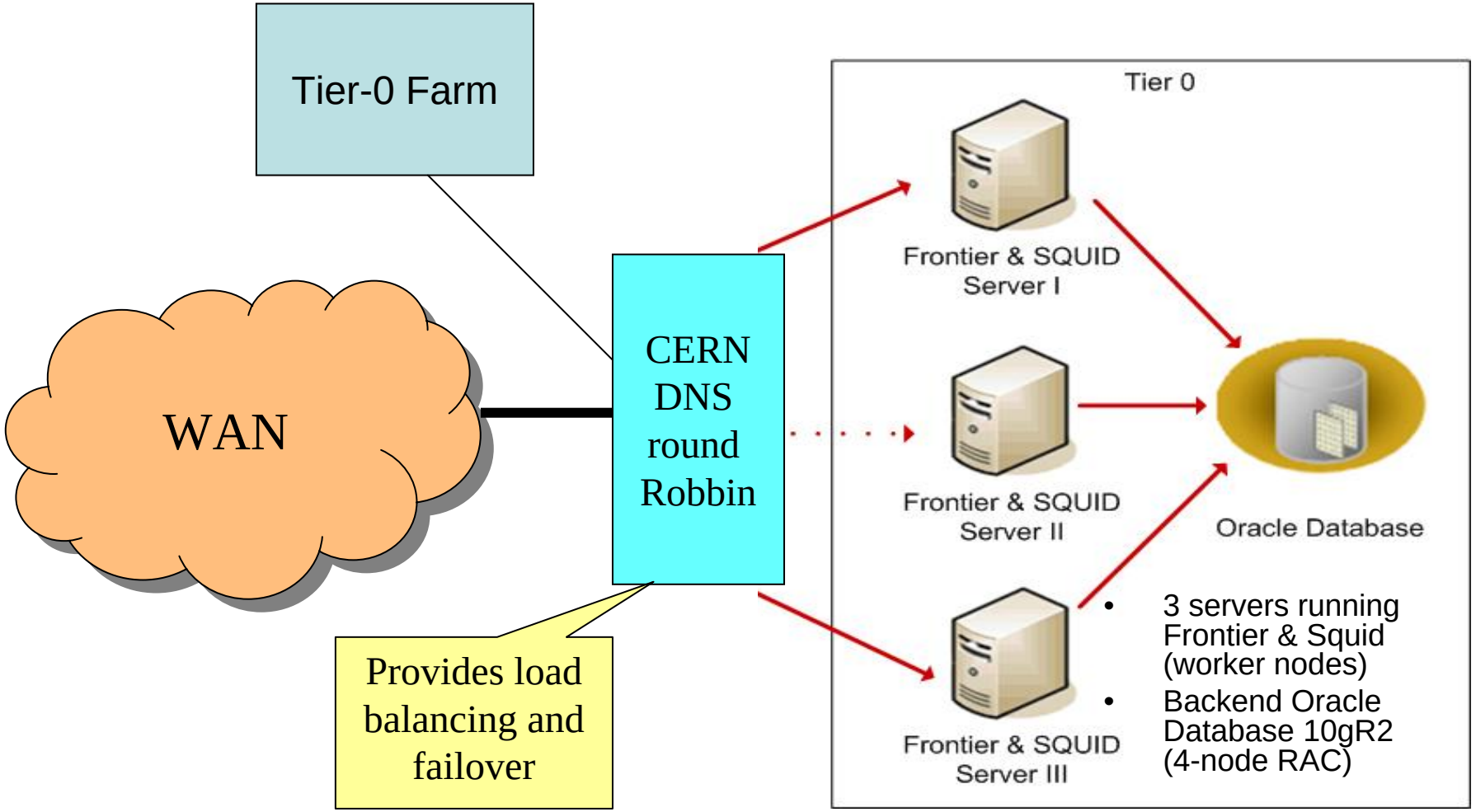
- Servlets run under Tomcat on Central Servers. DB connection is through JDBC and connection management is provided.
- Clients access servers via HTTP requests.
- Proxy caching servers (Squids) are deployed at Tier 0, 1, and Tier N sites.



Recent Features

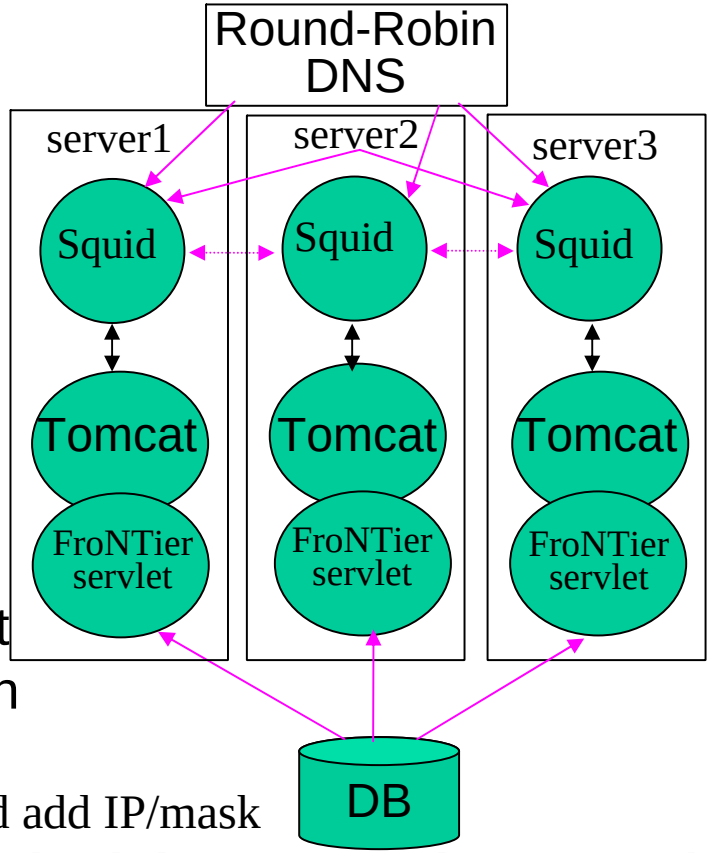
- Client can request the data be zipped by the server (compression levels 0-9)
- Keep alive signals sent to client when database is busy, avoids timeouts
- Client can use same HTTP connection for multiple requests.
- Server can insert expiration time in HTTP header for objects.
- Significantly improved client performance.
- Ported to 64-bit Linux
- Parameters can come in long parenthesized connect string instead of environment vars
- Can define logical name in long string so pool file catalog can use short name

FroNTier Launchpad Setup



FroNTier "Launchpad" software

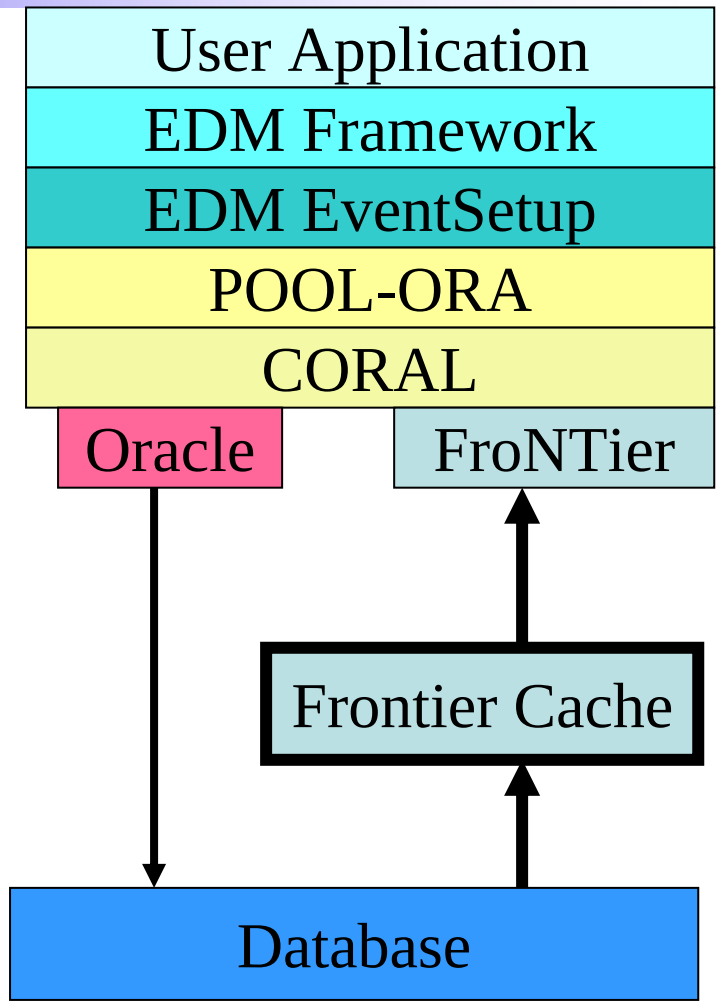
- Squid caching proxy
  - Load shared with Round-Robin DNS
  - Configured in "accelerator mode"
  - Peer-to-peer caching
  - "Wide open frontier"*
- Tomcat - standard
- FroNTier servlet
  - Distributed as "war" file
    - Unpack in Tomcat webapps dir
    - Change 2 files if name is different
  - One xml file describes DB connection



*In the past, we required the registration so we could add IP/mask to our Access Control List (ACL) at CERN. Recently decided to run in "wide-open" mode so installations can be tested w/o registration.

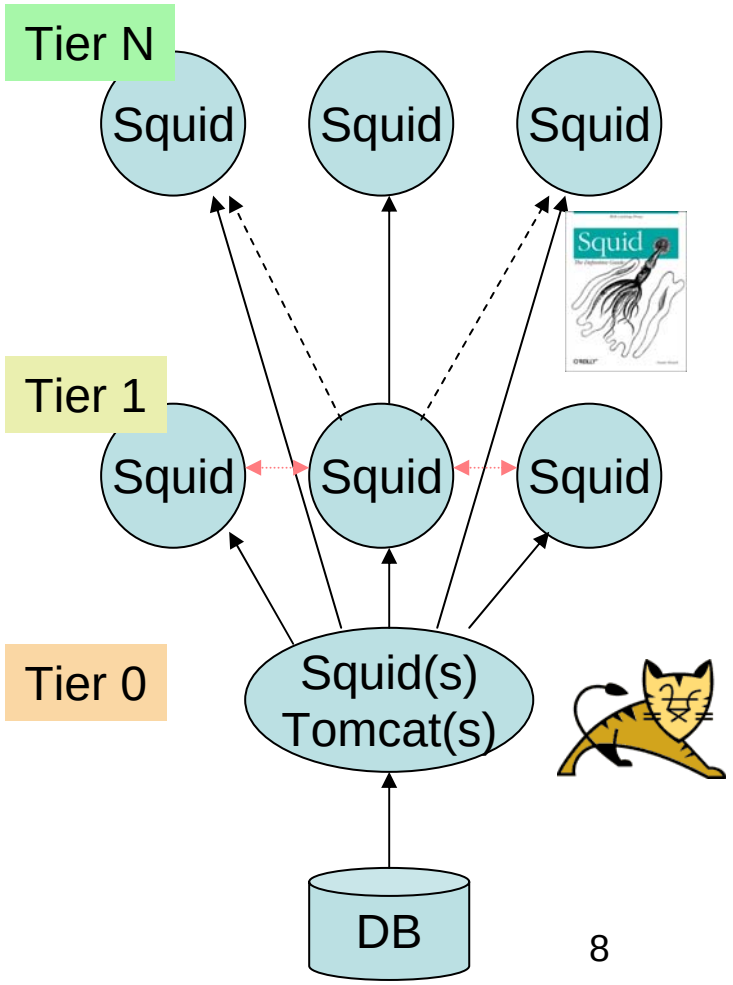
CMS Software Stack

- POOL-ORA (Object Relational Access) is used to map C++ objects to Relational schema.
- A CORAL-FroNTier plugin provides read-only access to the POOL DB objects via FroNTier.



N-Tier Deployment

- Redundant Tomcat + Squid servers are deployed at Tier 0.
- Squids are deployed at Tier 1, and Tier N sites. Configuration includes:
  - Access Control List (ACL)
  - Cache management (Memory and Disk)
  - Inter Cache sharing (if desired).
- Tier hierarchy (dashed lines) is a possible configuration change if needed.
- Site-local-config provides URL's of servers, and proxies.

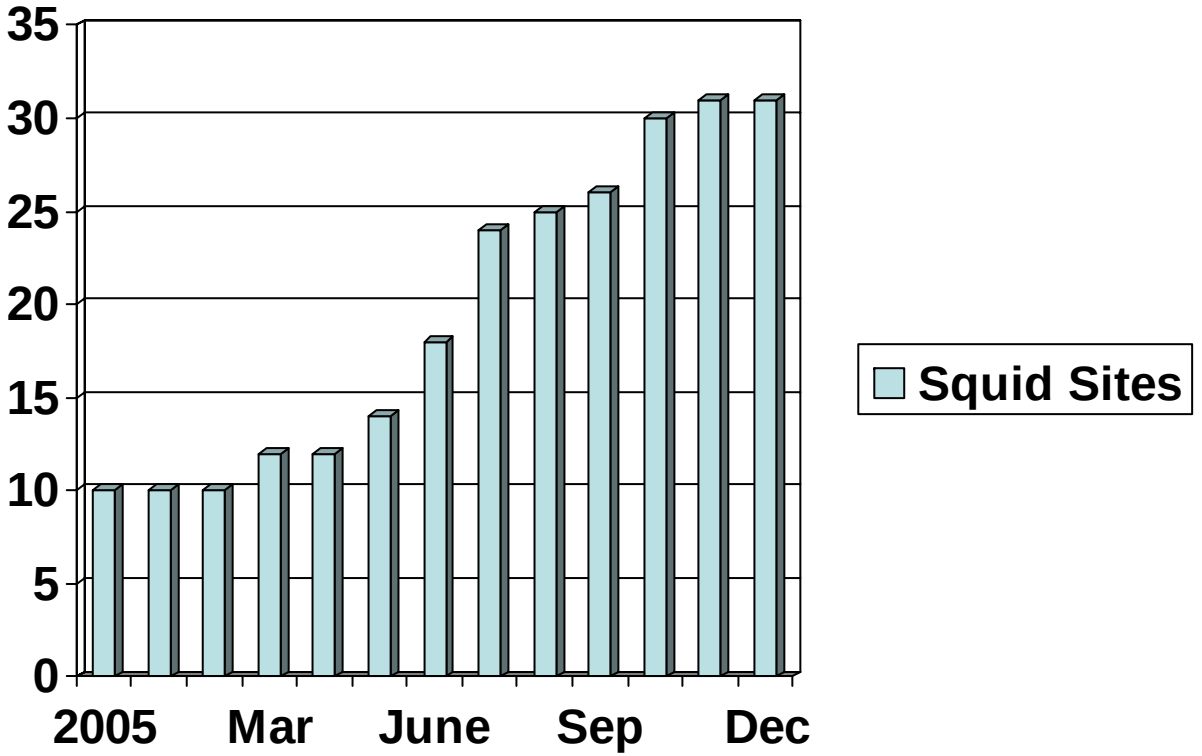


Site Squid Details

- Hardware requirements
  - Minimum specs: 1GHz CPU, 1GByte mem, GBit network, 100 GB disk.
  - Needs to be well connected (network-wise) to worker nodes, and have access to WAN and LAN if on Private Network.
  - Having 2 machines for failover is a requirement for T-0/T-1, and a useful option for T-2. Inexpensive insurance for reliability.
- Software installation
  - Squid server and configuration
  - Site-local-config file

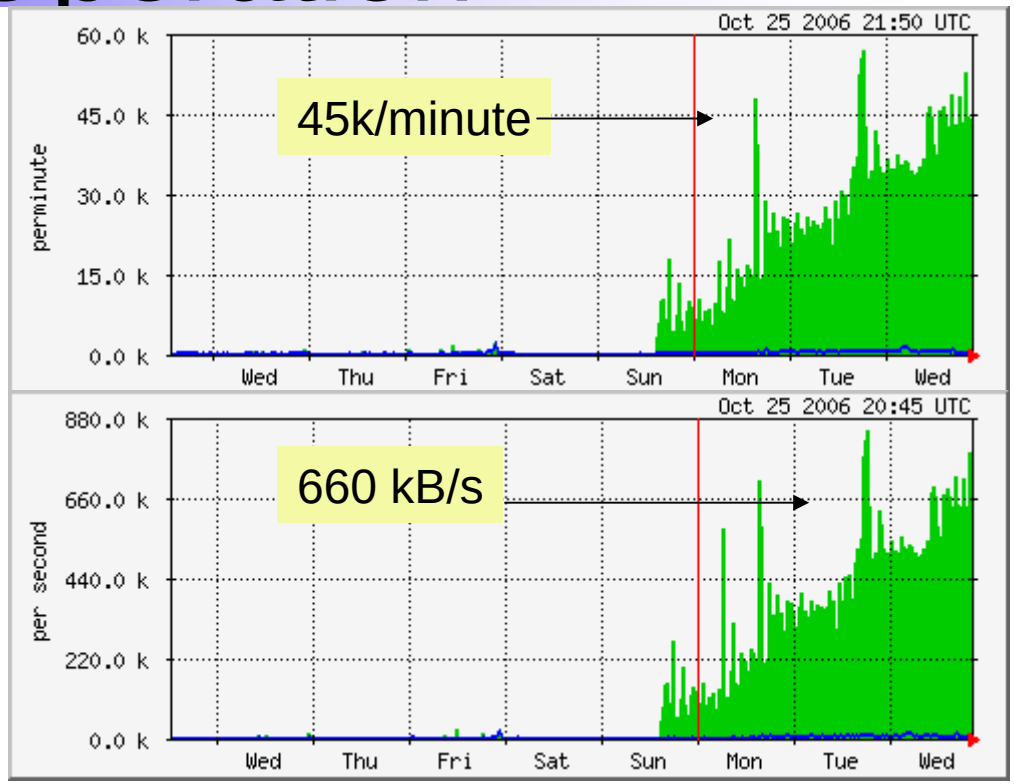
Squid Deployment Status

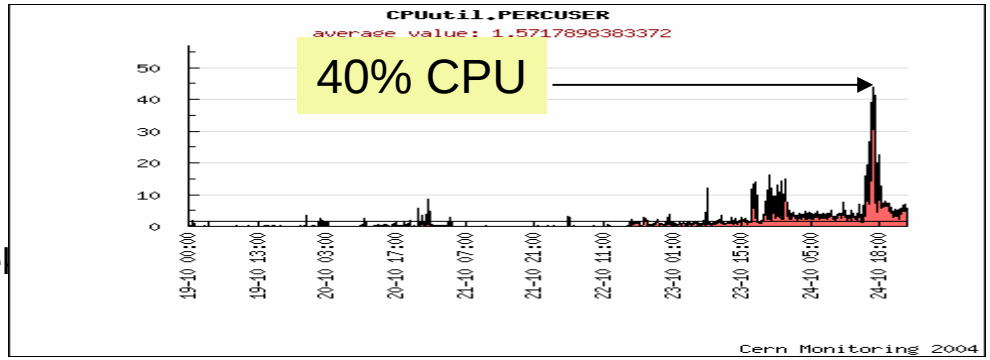
- Late 2005, 10 centers used for testing
- Additional installation May through Oct. used for Computing, Software, and Analysis challenge (a.k.a. CSA06)
- Very few problems with the installation procedures CMS provides.



Launchpad Load:T0 Operation

- One week of extensive testing during CSA06
- Ramped up farm to 1000 nodes.
- Spikes when new activity starts, and caches loaded with new objects.
- The load balancing spreads requests to all three servers (monitoring for one shown)
- Monitoring requests/minute, Bytes/sec with SNMP port on squid. Lemon provides server stats.

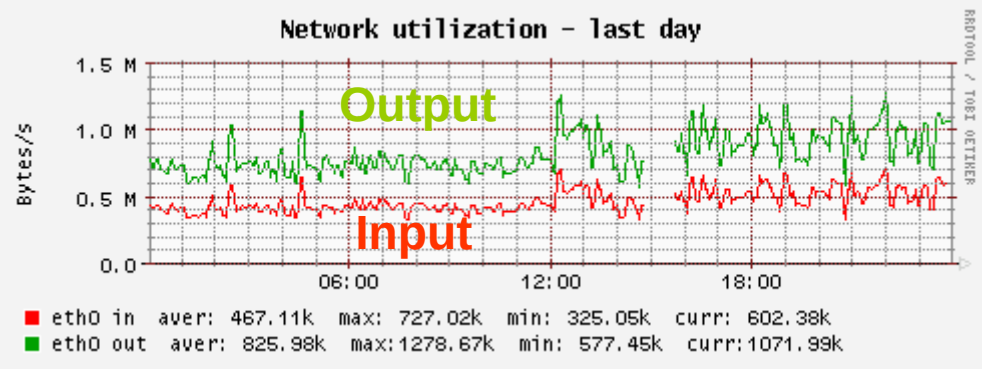




Throughput Limitations

- We discovered one CMS “object” that translates into ~28k DB requests.
- Throughput was limited to less than 1MB/s per server (3 servers).
- Squid logs were being produced at the rate of 2GB/hour.
- Clients connect to the Frontier squid with TCP. There is one TCP connection per request. The I/O overhead of the connection is about half as big as the data payload itself (128 Bytes compressed).
- Solutions:
  - “Fixing” the object storage (in progress)
  - Rotating squid logs as needed
  - Holding open HTTP connection for multiple requests.

1 MB/s →



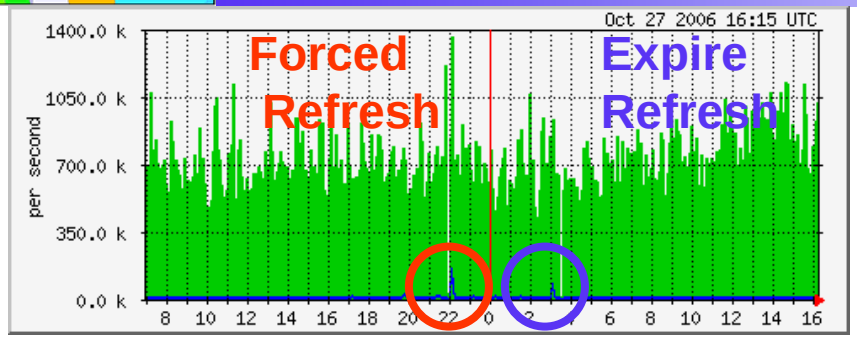
Recent Performance

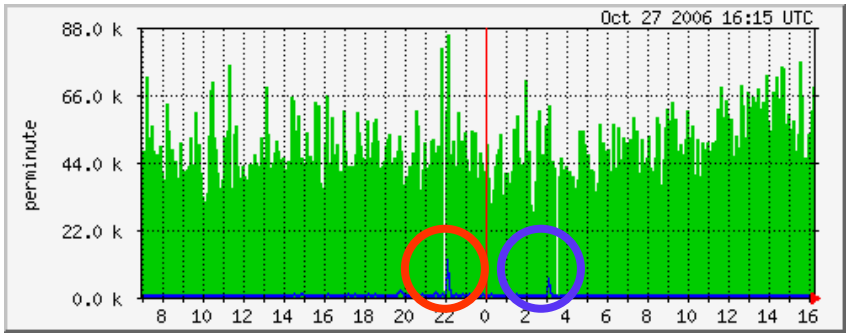
- Preparing for use with HLT farm. A large number of squids will be preloaded so the processor farm can quickly load from the cache.
- Recent testing at FNAL delivers 35MByte/s to clients
  - 200 simultaneous clients
  - Object size of ~2MB compressed
  - Squid running on quad 3GHz Xeon
  - All Gigabit network connections
  - Limitation is squid using single CPU that saturates.
- Have run 1000 simultaneous clients, just to test operation.

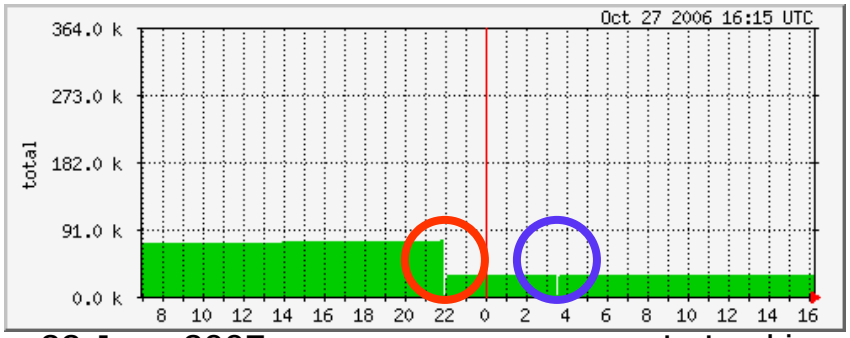
Cache Coherency

- In most cases CC is not a problem for conditions data. Managed by policy.
- In some cases, e.g.. during development, and some online uses, it can be an issue.
- Giving each cached object an expiration time is one solution that has been implemented.
- Several alternative approaches are being examined.

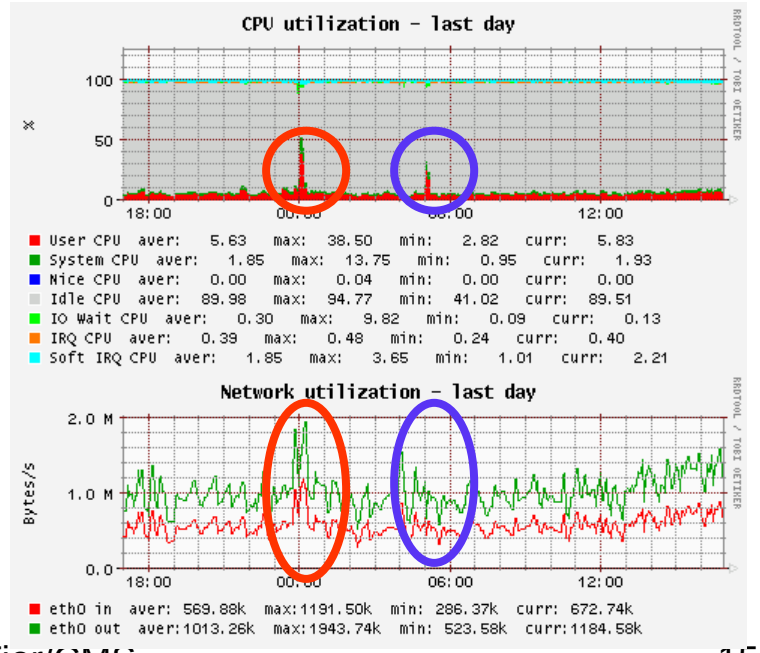
Cache Refresh @ 3 AM UTC





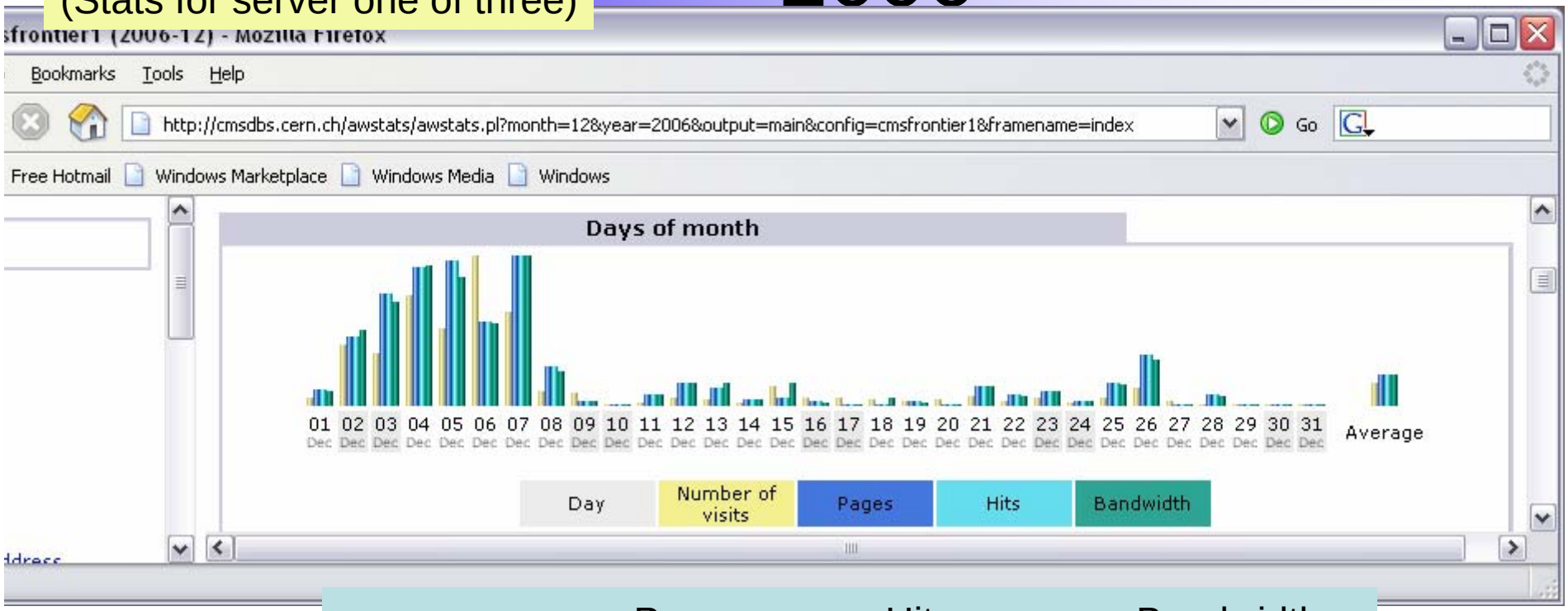


- Friday 27 Oct, started directing cache objects to expire 3:00 UTC next day (FrontierInt)
- Temporary “fix” for cache coherency.



Usage Analysis for December 2006

(Stats for server one of three)



	Pages	Hits	Bandwidth
Max 07 Dec 2006	3148109	3148109	2.72 GB
Min 27 Dec 2006	330	330	264.66 KB
Average	653634.90	653634.94	572.40 MB
Total	20262682	20262683	17.33 GB

Cache Stats for Dec 2006

(Stats for server one of three)

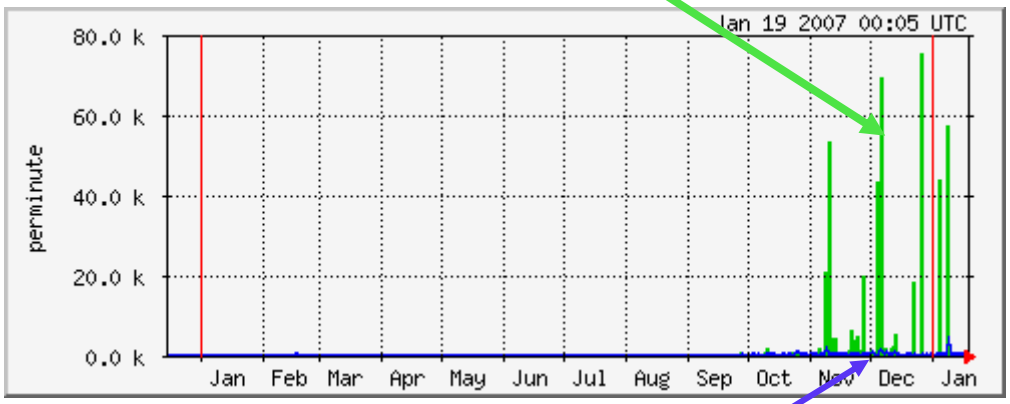
Access Status	Hits	Bandwidth
TCP_MEM_HIT:NONE (Memory Cache Hit)	16484727	13.94 GB
TCP_CLIENT_REFRESH_MISS:DIRECT (Forced Refresh)	2677225	1.86 GB
TCP_MISS:DIRECT (Request goes to Database)	1003958	1.39 GB
TCP_MISS:SIBLING_HIT (Found in “sibling” squid)	55621	68.39 MB
UDP_MISS:NONE	20961	5.80 MB
TCP_HIT:NONE (Disk Cache hit)	15302	63.05 MB
UDP_HIT:NONE	4667	1.30 MB
TCP_MISS:NONE	606	1.97 MB
TCP_CLIENT_REFRESH_MISS:NONE	12	21.66 KB
TCP_DENIED:NONE	1	1.40 KB
Total	20263080	17.33 GB

Cache Hit Rate = 82%

But Wait… There's more…

- A typical local Squid has a similar or better cache hit rate, see for example Fermilab in December.
- Thus, the cache hit rate for the global deployment is in the high 90% area.
- Only a few percent of requests must go all the way to the DB to be satisfied.
- This will improve when the conditions data becomes more stable, changing daily expiration to be less frequent, et cetera.

Requests satisfied by local Squid cache



Requests that are forwarded to CERN

Conclusion

- The FroNTier architecture is being used by CMS to distribute Conditions data worldwide.
- It was deployed to nearly 30 sites for use in CSA06.
- Throughput is limited when data objects are requested in many tiny pieces. Recent measurements show 35MByte/s throughputs for reasonably sized requests.
- Cache coherency concern has temporary solution, but needs more work.
- A high cache hit rate (high 90%) is observed.
- Overall, experience w/ the system is positive. Deployment, maintainability, reliability, and performance look good.

Finish

Squid Requirements

- Hardware min specs: 1GHz CPU, 1GByte mem, GBit network, 100 GB disk.
  - Closely networked to worker nodes
  - access to WAN and LAN if on Private Network.
  - Having 2 machines for failover is a requirement for T-0/T-1, and a useful option for T-2. Inexpensive insurance for reliability.
- Software installation
  - Squid server and configuration
  - Site-local-config file


CMS