



Large scale robust Internet service architectures rely on caching servers to reduce network traffic, improve user wait times as well as provide additional levels of security and robustness. Tumbleweed Valicert Validation Authority 4.7 is the first product to leverage open standards to offer digital certificate status caching in a unique, flexible architecture utilizing a repeater/responder approach.

1.0 Background

1.1 Caching Internet Servers

Large scale robust Internet service architectures rely on caching servers to reduce network traffic, improve user wait times as well as provide additional levels of security and robustness. Support for caching Internet objects is built into the HTTP protocol itself as described in RFC 2616 (and the earlier RFC 1945). Traditionally caching servers are either placed next to the client (proxy mode) or in front of a web server (reverse proxy mode). Figure 1 illustrates these configurations:



Figure 1

In order to support these configurations, caching servers provide certain key features:

- On demand caching of Internet objects, and returning these objects from the cache efficiently and transparently
- On command batch updates for caching Internet objects on a scheduled basis, including refreshing objects in the cache at specified intervals to ensure that data is current and available for periods of heavy use
- Support for hierarchical or peer-to-peer distributed caching via proxy-chaining or dynamic querying of neighboring caches
- Ability to accept a Secure Sockets Layer (SSL) connection from the client and create a new SSL session with a protected web server, providing an additional barrier for web servers and applications behind firewalls

Caching servers provide many additional features that are outside the scope of this document, many of which are also vendor specific.



1.2 OCSP

The Online Certificate Status Protocol (OCSP) is an IETF protocol (RFC 2560) used to determine the current status of a digital certificate issued by a particular Certifying Authority (CA) without requiring a client to obtain and examine the entire Certificate Revocation List (CRL) issued by that CA. OCSP enables the client to query a Validation Authority (VA) either delegated by the CA or trusted by the client regarding the status of a particular certificate. The VA's OCSP Responder satisfies the client's query having previously obtained the appropriate CRL from the issuing CA. Since OCSP is a stateless transactional protocol, Appendix A of RFC 2560 describes how OCSP transactions can be implemented using HTTP as a "transport" protocol (protected using SSL or some other lower layer protocol if privacy is required).

OCSP is a critical component of any PKI deployment and the following diagram (Figure 2) illustrates the typical OCSP architecture today:

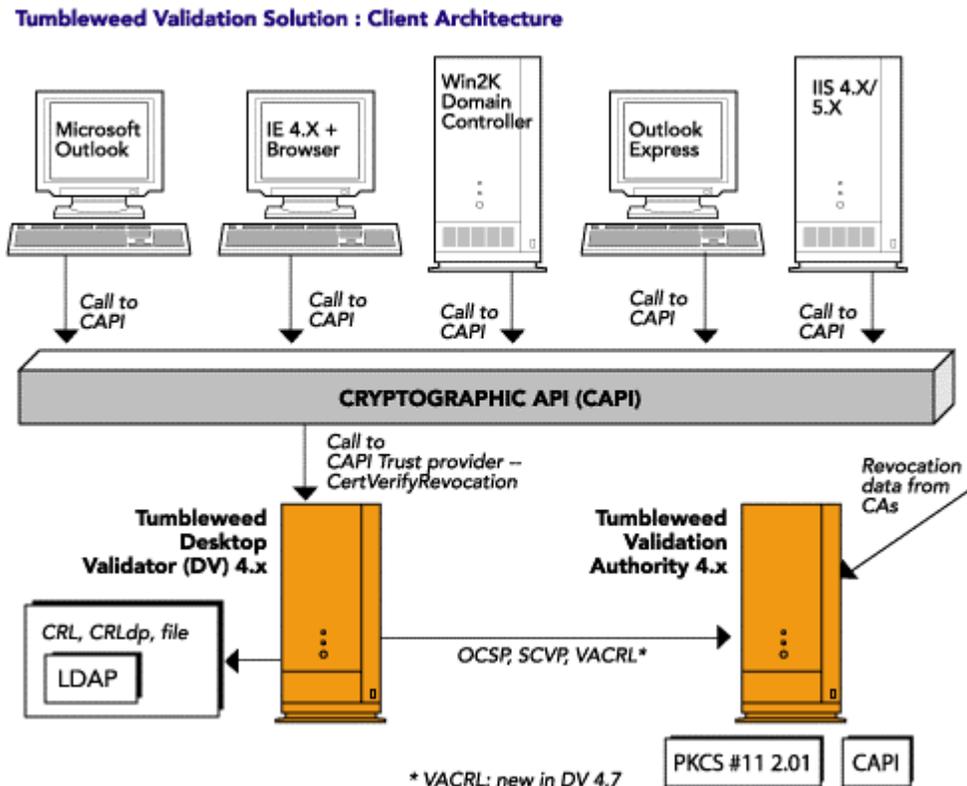


Figure 2



2.0 EVA 4.7 Repeater

2.1 Architecture

Since OCSP is implemented using HTTP, it follows that the caching can be introduced into the architecture in much the same way as it is used in traditional Internet service architectures. The following diagram illustrates the introduction of a caching proxy OCSP “Repeater”:

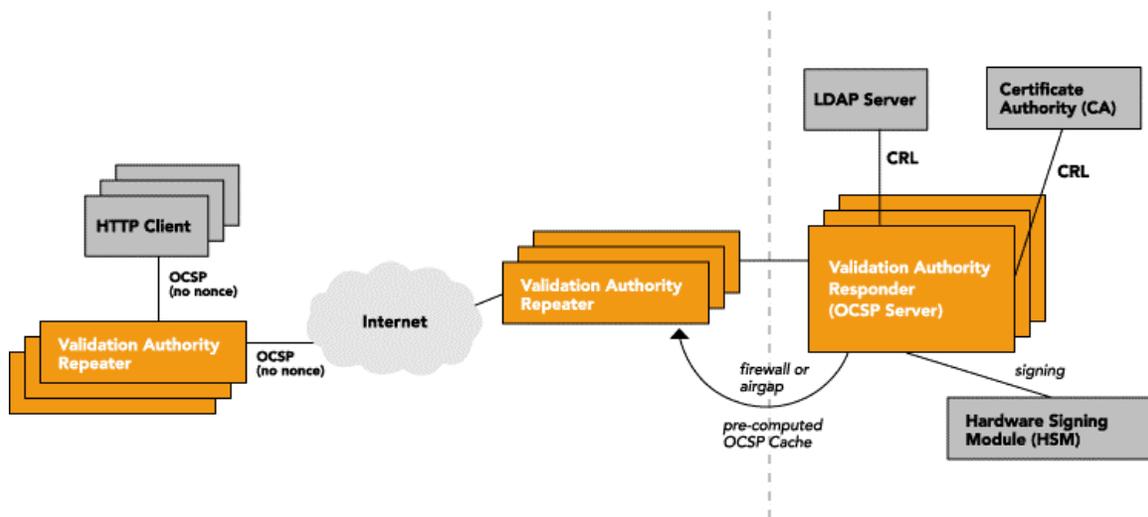


Figure 3

2.2 OCSP Proxy Caching

In order to support the new functionality, EVA 4.7 works in the following way:

- A Responder can “pre-produce signed responses specifying the status of certificates at a specified time” as described in section 2.5 of RFC 2560. The nonce extension described in section 4.4.1 which cryptographically binds a client’s request to a server’s response does *not* get utilized when pre-producing responses so that the response may be cached and used to service multiple requests.

Additionally, OCSP allows clients to request status information about a series of certificates, the response syntax as defined in section 4.2.1 allows an OCSP response to be a “SEQUENCE OF SingleResponse”. The Responder takes advantage of this syntax as well as the “nextUpdate” attribute when generating its pre-produced responses in order to improve cache performance. This will be discussed further in the section regarding **Performance**.



- A Repeater can request pre-produced OCSP responses from a Responder either on-demand or on a scheduled basis. A Repeater can accept pre-produced OCSP responses published by a Responder. In other words, the Repeater and Responder can work together in either a push or pull model in either real-time or batch mode.
- A Repeater can service client OCSP requests by returning the appropriate cached OCSP response. If the Repeater does not have the appropriate response in its cache (or if the response in its cache will not satisfy the client's request stated caching policy), it will transparently proxy the client's request to either another Repeater or the Responder, return the appropriate response back to the client, and store the response in its cache for future requests.
- A Repeater accepts an SSL connection from a client and creates a new SSL session with a protected Responder or a Repeater in reverse-proxy mode. This enables Clients, Repeaters, and Responders to all be located on different networks, protected via firewalls which restrict traffic according to policies set by each of those networks.
- Clients, Repeaters, and Responders are able to specify and negotiate flexible caching policies (expiration, use of cached response, etc.)

2.3 CRL and Delta-CRL Caching

In previous versions of EVA functionality existed to support the mirroring of a CA issued CRL or VA generated "delta-CRL" between instances of the Responder. This allowed for more efficient distribution of CRL data without requiring every Responder to access the CA and obtain a full CRL in order to maintain its digital certificate status repository. With the introduction of proxy caching in EVA 4.7, the CRL and delta-CRL can now be cached in the Repeater. This allows the distribution of CRL data to clients who cannot rely on OCSP or who wish to have a fall-back to OCSP in operational environments where real-time network access is not possible at all times.

3.0 EVA 4.7 Benefits

The introduction of the Repeater as outlined above offers three major benefits:

- **Improved response latency.** By placing a Repeater close to the client and/or in front of the Responder any network or server latency issues caused by traffic/load or bandwidth/capacity limitations can easily be addressed. This implies the overall PKI will be more resilient to potential Denial of Service attacks.
- **Improved Fault Tolerance.** Since Repeaters are stateless it is possible to have many instances running transparently behind a standard HTTP load balancer. If one Repeater fails, the load balancer will automatically redirect all traffic to another instance. Additionally since the Repeater caches Responder data, it is possible for the PKI to operate for some amount of time even if the Responder is unavailable. The increased robustness of the architecture implies the overall PKI will



be more resilient to any major outages in the backend infrastructure (e.g. at the CA, or LDAP directory).

- **Improved Security.** Responders contain sensitive private key material required to sign responses. The potential compromise of a Responder means the entire PKI is compromised since the attacker can generate false responses. By contrast, Repeaters do not have any keys and hence the potential compromise of a Repeater does not compromise the entire PKI.

Additionally reverse-proxy Repeaters enable placing the Responder on a different network, which is partitioned via a firewall from the public/client network, reducing the risk of the Responder being compromised. One can even go so far as to transfer the pre-computed OCSP caches from responder to repeater via removable media (e.g. DVD, optical media, DAT tape) to implement an air-gap defense between repeater and responder.

4.0 Performance and Security Considerations

4.1 Cache Hit Considerations

Any time the Repeater cannot service a client request out of its cache, it will need to contact the Responder for the necessary response and the client will experience an increase in wait-time. In other words, any request which results in a cache miss will have additional latency. Therefore it is desirable for as many responses as possible to be satisfied from the cache. To ensure this, it is desirable to preload the cache with as many OCSP responses as possible.

Some parties have suggested that an OCSP responder should be preloaded with the status of every certificate ever issued by a given CA! Not only would this make the cache prohibitively large in terms of computing resources, it assumes a tight synchronization between the cache and the issuing CA database is possible. Given the cache is usually in a different administrative domain than the CA, this is a very unrealistic assumption in the real world. Not only is the cache in a different domain than the CA, it is quite possible for the cache to be in a different administrative domain than the VA.

Fortunately it is possible for a Responder to use heuristics based on the CRL to pre-compute responses for pre-loading a Repeater cache. A Responder can use the first and last serial numbers appearing in the canonically ordered CRL as a serial number range to use for pre-computing OCSP responses (since a CRL does not contain serial numbers of expired certificates). This serial number range can vary in size, so the Responder allows an administrator to specify the total minimum and maximum number of OCSP responses to pre-produce based on the resources available for the cache. The Responder administrator can specify how many of the responses should pertain to certificates with a serial number less than the first serial number appearing in the CRL, and how many of the responses should pertain to certificates with a serial number greater than the last serial number appearing in the CRL allowing the cache to be more efficient if the administrator has some actual knowledge regarding the CA issuance/revocation history. Additionally, the



Responder should use the “nextUpdate” attribute to provide guidance to the Repeater and Client as to how long a response is to be valid.

If the Repeater cache is pre-loaded and routinely updated with the Responder’s pre-computed responses, the chance of a cache miss occurring is greatly reduced and overall cache performance is maximized.

4.2 Resource Considerations

The OCSP protocol allows a response to contain status information regarding a list of certificates and the Responder allows an administrator to specify the size of that list. Since each response is digitally signed by the Responder, the larger the list size, the more savings in terms of resources required to generate (signing operations), transmit (bandwidth), and store (memory/storage) the status of those certificates in the Repeater cache. However since Responder generated messages will ultimately be sent to a client (the Repeater merely caches), the potential savings must be weighed against the additional burden on the client in terms of receiving and processing a larger response. This trade-off must be made taking into consideration the entire PKI architecture.

4.3 Security Considerations

Since Repeaters are caching responses, it is possible that the information they are providing is no longer accurate. This risk can be addressed via general as well as specific caching policies. EVA 4.7 provides a robust set of configuration parameters to allow the customer to tune how often pre-computed OCSP caches are generated to address this issue. Overall it is believed that the benefits offered by caching OCSP responses outweigh the risks. However, this all depends on the security requirements of the PKI deployment. Some high security environments may not tolerate the security risks associated with not including a nonce in the OCSP request. Using HTTP over SSL can help reduce the risk associated with leaving out a nonce in the OCSP request. Such a solution would require the use of a hardware signing module (HSM) or hardware accelerators at the Repeater, and require the Repeater to manage RSA public/private keys. Tumbleweed will be happy to help customers make the appropriate trade-offs.

4.3 Deployment Considerations

It was previously mentioned that it makes sense to locate Repeaters close to the client or in front of a Responder in order to improve latency and reduce network traffic and server load. However, since a Repeater can proxy requests to either other Repeaters or Responders, it is possible to setup multi level caches using either a hierarchical or peer-to-peer topology.



5.0 Conclusion

Introducing OCSP caching in a PKI deployment offers many advantages in terms of scalability, performance, and security. This approach is based on the proven success of using HTTP caching in traditional Web service architectures. EVA 4.7 is the only product to offer OCSP caching solution based on open standards. The flexible Repeater and Responder supports numerous potential configurations, providing customers with a robust, reliable solution.