A new architecture for Web Services at CERN

A. Di Meglio, F.Fluckiger, M.Marinucci, P.Hagen, A.Pace

CERN, Information Technology Division, Geneva, Switzerland

Abstract

A new architecture based on a pool of load-balanced web servers has been deployed to allow the evolution of the CERN Web Services. The new architecture introduces the concept of *logical URL address* that is translated to the physical URL through a database lookup that gives the necessary independence from the hardware layout to move web sites across servers and to avoid broken links or unnecessary redirections. Complete backward compatibility is kept with the existing www url addresses, and all *locally managed servers* that are outside the control of the central services can also be integrated in the web namespace for CERN. The new architecture allows differentiation of the various servers in the pool to offer specific or specialized services (high-availability servers for static html only, CGI servers, Gateways to AFS or Novell) if necessary. This paper gives a description of the infrastructure in place and then reviews the services offered to CERN users. The solutions retained to offer support for automated web site registration, electronic forms, database web connectors, scripting, CGI interfaces, SSL, server load balancing, authentication and access control for web authors and web developers are briefly explained.

keywords www, web, CERN, Apache, IIS, Internet Information Server, Windows, UNIX, NT, ISP

1. A new web logical namespace for CERN

In the last years, the web has become very rapidly the major publishing infrastructure of many organisations, including CERN. Unfortunately, the conventional strategy for publishing physical URLs has exposed hostnames of CERN web servers and physical pathnames. As a result, any evolution of the web service architecture without breaking existing links or disrupting the service is tricky to implement.

A solution to this problem is being proposed: All URLs published to the outside world are *"logical"* and are independent of the physical server and path that host the web site. A *redirection* service, with a database lookup will map the logical http request to the physical path.

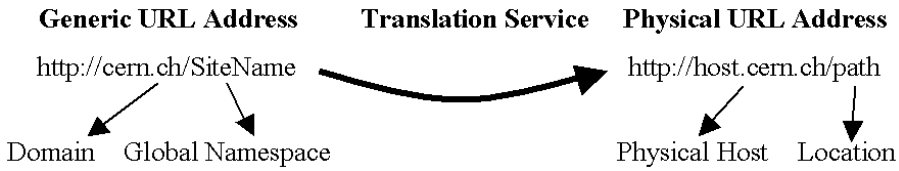

Figure 1: The base of the web namespace and its translation/redirection service

To ensure backward compatibility with the past, all requests containing physical hostnames that have been exposed in published URL in the past (www.cern.ch, nicewww.cern.ch, home.cern.ch, etc) will also be intercepted by the web redirector and mapped to the appropriate web server.

In addition to this redirection service, in order to support web-authoring tools, we had to provide a method to identify the physical web server for a given web site name. A subdomain named *web.cern.ch* has been created where all CERN known web sites have been registered. Within this subdomain, the host *xxxx.web.cern.ch* represents by convention the web server hosting site *xxxx.*

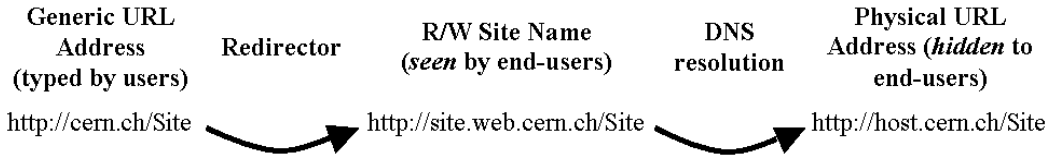

Figure 2: The web namespace hiding all physical dependencies

Finally, as the redirector is database driven, the possibility of having site alias has been implemented. A site alias is an alternative name that can be used to reference an existing site. This allows multiple names for a given site and allows site migration across servers.

The CERN web site database is aware of site hosted on both central and private servers and allows sites on private servers to be part of the CERN web namespace. The site database allows site owners to migrate easily their existing sites from locally managed servers to the central ones (or vice-versa).

Once a web site has been registered in the CERN web namespace, a wide range of URL forms will be accepted: *http://cern.ch/site*[1], *http://www.cern.ch/site, http://web.cern.ch/site, http://site.cern.ch*[2], *http://site.web.cern.ch/site*[3]. As usual, the domain name cern.ch can be omitted for access made within the cern.ch domain.

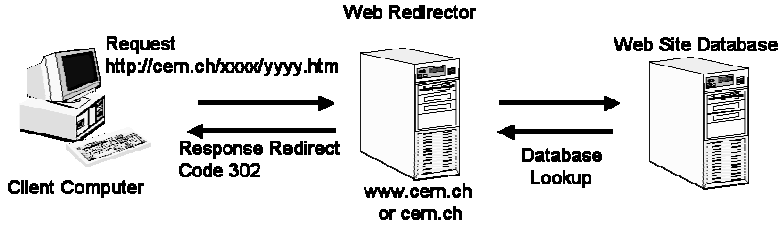

Figure 3: The CERN web redirector only sends redirect responses (http code 302)

2. The CERN web server infrastructure

Within this infrastructure where the web namespace and published URLs are independent from the physical network of web servers, several improvements have been possible. The most important is the evolution from the *single machine* architecture (www.cern.ch) to an architecture where several web servers are contributing to the service. These web servers can be identical between them (load balancing, redundancy, hot spares, …) of differentiated (different operating systems, different http servers, high reliability with static html only, development servers, access restricted from the CERN intranet only, etc).

The basic strategy taken for the CERN web architecture is to try to be as aligned as possible with the software used by Internet Service Providers (ISP) on the Internet. Using this approach, it may be possible in the future, as appropriate, to outsource parts of our services to industry at an acceptable cost. Following this approach, we are trying to be as uniform as possible and differentiate only where necessary. The HTML files (or scripts) can be physically stored on the local disks of the central web servers or on the CERN major file services (NICE-Novell or AFS).

It is currently supposed that web sites hosted on central servers are *uploaded using the http/ftp protocols only*. For this reason, we do not officially disclose the server operating system to web authors. For sites hosted on file systems, it is supposed that web sites stored

[1] The forms http://cern.ch/site and http://www.cern.ch/site are the recommended URL forms to be published or referenced outside CERN

[2] The form http://site.cern.ch/ will work only if the site owner has registered the name site.cern.ch in the central DNS service as an alias of the central redirector.

[3] The form http://site.web.cern.ch/site should be used only to author an existing web site

on AFS, are authored/published from AFS clients, while web sites stored on Novell are authored/published from a NICE client.

The entire architecture is managed from the web itself. In particular, All web site management is completely automated and delegated to site owners: They can create web sites, delete them, or delegate authoring rights without any manual intervention of the web services staff. A typical site creation request has to undergo a process in multiple steps where:

- The physical web site is created on a server of the pool selected according to the requirement specified by the owner in the web registration form.
- The new web site and its owner/creator are registered in the web site database.
- A new host is registered in the Domain Name Server (DNS) of the web.cern.ch subdomain (named xxxx.web.cern.ch, where xxxx is the site name).
- The CERN web redirector is notified and reloaded with the updated database.
- One or more moderators are notified in real time of the existence of the new site.

As those steps are automated, the whole processing time for a web site creation request is normally less than 20 seconds.

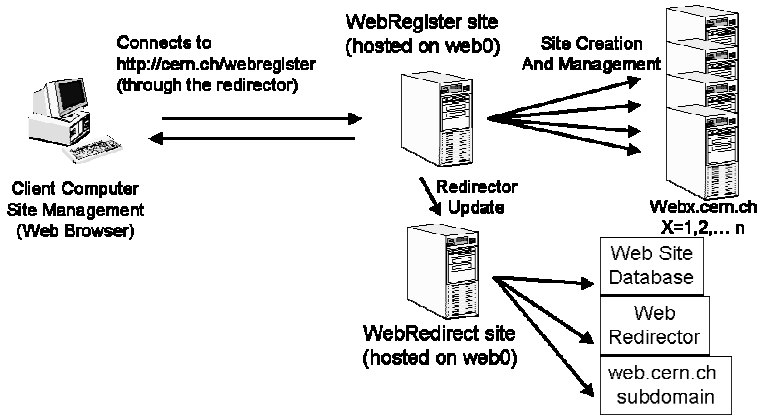

Figure 4: The automated Web Site Registration service

3. A wide range of new web services for CERN users

As the new web services are spanning on multiple machines, several new facilities could be introduced for sites hosted on the central servers (sites on Novell and AFS have been limited to static HTML publishing with some access control and CGI scripting services for the latter):

- *Support for web authoring tools* that uses the HTTP or FTP protocol for publishing. In particular, the CERN web services now support [1] Microsoft FrontPage 98, Macromedia Dreamweaver and, latest during the year, Office 2000 applications (Word / Excel / PowerPoint / FrontPage / Access). Content from look and feel separation is supported either using Cascading Style Sheets (CSS) or FrontPage Themes [2].
- Web authoring tools that can publish files only on the local file system without using the HTTP or FTP protocols are supported through the *Web Folders* feature available in Internet Explorer 5. Web authors can view, manage, move, copy, save, and rename the files and folders on a web server just as they would perform the same actions on the local file system. However, when users see the contents of a web folder, they see a list of files and folders and their associated Uniform Resource Locators (URLs). This allows seamless publishing, to a web as effortlessly as saving to a network file server [3]. For the platforms where Internet Explorer 5 is not available (i.e. other than Macintosh, Windows, Solaris and HP-UX), users can use standard FTP clients when using file-based authoring tools.

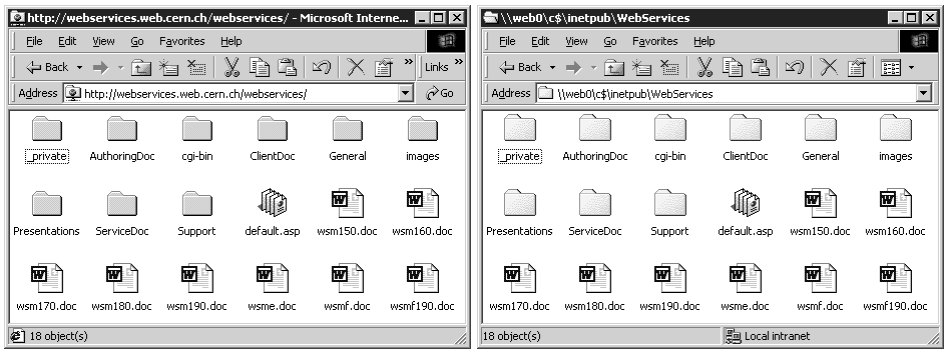

Figure 5: What is the difference ? The first access uses HTTP, the second uses the file system

- As a side effect, being the http protocol world-wide accessible from any computer on the Internet we are expecting end-users to setup web sites for personal user as *global repository for their personal working files* (home directories). This is a possible use of the CERN web services that can have a significant impact on future computing use.
- *Support of Electronic Forms* [4]. This allows web authors to create electronic forms on their web site. Form input data can be saved to a local file on the web server, added to a table in a database, or be used as parameters to generate database queries. Form input can also be sent to user-written scripts allowing authors to develop web-based applications.
- *Support for web database connector* [5] in read or write modes. Web authors can now create HTML pages whose content is dynamically linked to database content. Databases can be local on the web server (Access or Excel format) or remote (Oracle) on the network. Databases dynamic queries can be generated easily from form parameters and form input can be easily appended to databases tables. Finally, a standard scripting database interface is available for authors willing to develop web based database applications.
- *Access Control Support.* The new web services support now a wide range of access control features: from the simplest password protected documents (all users use the same password) to server side authentication (users have to login using their NICE or AFS account) that is verified against Access Control Lists (ACLs) that web authors can add to their documents to define who can read and who can write to a particular URL. Web authors can also restrict access to their web site using IP address masks and therefore restrict access to particular domains on the Internet (example: sites visible only inside CERN). Finally, for web authors requiring user-based authentication for users who are not registered in the CERN user database, the standard scripting interface can be used to authenticate external users [6].
- *Source Control Support.* For large web site authored by several persons simultaneously, the web services offer Source Control to force a formal documents check-out / check-in mechanism prior to any document modification [7].
- *SSL and HTTPS service* [8]. Web authors can now use secure HTTPS connections that are encrypted and ensure confidentiality of (authenticated) document transmission and form input submission. Authors have the choice to use HTTP or HTTPS for all documents stored in their web site as well as forcing a given document to be accessible only using secure or non-secure protocols.
- *Search and indexing services* [9]. All public readable documents stored on the CERN web servers are indexed by the CERN Search engines allowing global CERN wide searches for documents. In addition, also server-side indexing takes place on all central web servers to allow instantaneous index update whenever a document is modified and adding also server side information (hit counts, ownership, hit count per day, etc), allowing more advanced searches like “What’s new on this web site in the last 3 hours ?” or “Show me the most accessed documents in this web site” …

- *Scripting and CGI interface support* [10]. The CERN web sites now support a wide set of scripting interfaces. From the usual cgi-bin directory where end user can drop (platform specific) executables to portable scripting interfaces. The supported scripting languages are PerlScript, Jscript and VBScript for sites hosted on the central servers and Perl-Unix for sites hosted on AFS. In addition to the usual CGI Interface, sites hosted on the central servers also support server-side scripting in HTML (Active Server Pages – ASP). All these features allow web authors to develop sophisticated web Applications that clearly can go beyond the normal static html model.
- *Documentation,* in both English [11] and French [12].

4. Conclusion

An important set of pending requirements for the web services has been addressed with the new architecture. The World Wide Web is now becoming a *commodity* technology and CERN is trying to align on the same technology used by Internet Service Providers (ISP), not only to reduce costs, but also for additional reasons: on the one hand this appears as the best way to ensure a high reliability service able to cope with the fast technology evolution we are all experiencing, and on the other hand, this represents the best guarantee that web services become really usable also by computer novices or to CERN staff without computing experience.

References

[1] See http://cern.ch/web/authoringdoc/authoring.htm
[2] See http://cern.ch/web/authoringdoc/stylesandthemes/stylesandthemes.htm
[3] See http://cern.ch/web/authoringdoc/Editing/editing.htm
[4] See http://cern.ch/web/authoringdoc/forms/forms.htm
[5] See http://cern.ch/web/authoringdoc/database/database.htm
[6] See http://cern.ch/web/authoringdoc/accesscontrol/accesscontrol.htm
[7] See http://cern.ch/web/authoringdoc/versioning/versioning.htm
[8] See http://cern.ch/web/authoringdoc/https/https.htm
[9] See http://cern.ch/web/authoringdoc/searchservices/addingsearchservices.htm
[10] See http://cern.ch/web/authoringdoc/scripting/scripting.htm
[11] The Web Services team: “The Guide to WWW publishing and Authoring at CERN”, CERN-UCO/1999/208, December 1999
[12] The Web Services team: “Guide de la creation et de la publication WWW au CERN” CERN-UCO/1999/209, December 1999