Influence of deployment architecture on performance optimization of PLM systems in auto industry

M.S. Gopinatha, Dr. Vishnukanth S. Chatpalli, Dr. K.S. Sridhar

***Abstract*— Product Lifecycle Management (PLM)** is a software application based business approach to support the collaborative creation, management, dissemination and use of product definition information. PLM applies to product information from concept to end of life—integrating people, processes, and information. **PLM technology** has become the backbone of product design and development for many of the auto companies across the globe.

System Performance is the "Transaction time as perceived by the end-user - conforming to requirements". **Optimized performance** of all the IT components of the PLM system will ensure successful product design and development in auto industry. Optimization will generally focus on improving just one or two aspects of performance: execution time, memory usage, disk space, bandwidth, power consumption or some other resource. This will usually require a **trade-off** - where one factor is optimized at the expense of others.

A typical **PLM system** in automobile industry consists of the **PLM Application server** which hosts the PDM system, CAD system, Digital Simulation system, BOM system etc. **Database server** stores all metadata of PLM system while the bulk data is stored in the file volume server connected to SAN storage. **Web server** helps in connecting the PLM Application server with client machines, ERP systems, CRM/SCM systems, legacy systems etc. [Ref. 4 to 7]

An effort is made in this research paper for analysis of the **deployment architecture** component of the PLM system, which greatly influences the optimization of system performance. [Ref. 6] Present day automotive industry is global, both in terms of customers spread across multiple regions (from mature and emerging markets) and manufacturers and suppliers scattered around the world. Combined with developments in the global network infrastructure it is now possible to consider Global deployment of PLM system based on a few or even only one central site. For many customers this is already a becoming a reality. Global deployment offers saving in administration and closer co-operation but also new challenges.

Manuscript received Sep 12, 2014

M.S. Gopinatha, Research Scholar, PES Institute of Technology (VTU), Bangalore, India, Senior Expert Specialist, PLM Competency Centre (Asia Pacific), Siemens PLM Software

Dr. Vishnukanth S. Chatpalli, Professor and Registrar, Rani Channamma University, Belgaum, India

Dr. K.S. Sridhar, Prof. of Mechanical Engineering, PESIT, 100 feet Ring Road, BSK III stage, Bangalore, India

Centralisation of services relies on the availability of central site. The availability of the site depends on every component of the PLM system deployed. All CAD and other File system items are required to be shared across the all user sites spread across the globe. Actual data is stored at Datacenter SAN storage and is to be shared across remote locations. Network Latency and Bandwidth between site locations play a major role on the overall system performance of the global PLM system.

I. Overview of PLM system Deployment Architecture

Focus of defining the PLM system deployment architecture is to maximize the usage and architecting the deployment within the constraints of network. Upgrading network infrastructure is practically very difficult and expensive. Consider end user's distribution and define the optimal deployment architecture based on available bandwidth and latency of the network.

Deployment Architecture to be deployed should define the following:

- Types of clients to be used at the various sites such as 2-tier, 4-tier, rich client, thin client, etc.
- Location of servers – Resource Tier, Enterprise Tier, Web Tier
- Location of files – Volumes, File Caches, Logs, etc
- Failover servers including clustering options
- Load balancing architecture for each layer of PLM system
- Options possible with virtualization of servers and cloud computing
- Other requirements including:
  - Translation servers
  - Integration with any existing third-party applications
- Impact of Security requirements
  - De-Militarized Zone (DMZ)
  - Proxy servers
  - Firewall
- High Availability
  - Service level
  - Resource level

II. Two-tier and Four-tier Deployment Architecture

The 2-tier Rich Client user interface depicted in Figure1 is highly interactive requiring very low latency (ping time) from client to server communications. Latencies exceeding 10ms will result in unacceptable end-user response times and hence call for design of optimal deployment architecture considering the network effect on performance. Operating 2-tier Rich Clients in a WAN environment is not recommended.

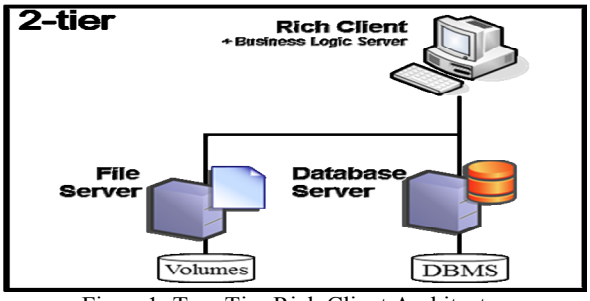

Figure1: Two Tier Rich Client Architecture

The advantages of the two-tier architecture are:

- The Business Logic server processes are distributed across the enterprise to all the users' workstations, each user's Business Logic server is local to his machine. Thus, the memory and CPU requirements are distributed, making use of workstation CPU and memory resources;
- The backend is simpler: There is no need for a web application server for rich client users.

The key disadvantage of the two-tier architecture is that it is extremely sensitive to network latency between a user's workstation and the database. The insertion of even a few milliseconds can affect performance significantly. This is due to the chattiness of the Business Logic server process with the database. The maximum recommended round-trip latency for two-tier is 10ms.

The advantage of the four-tier architecture depicted in Figure2 is that the Business Logic server processes are centralized, close to the database. The effect of network latency is mitigated, because the rich client is less chatty with the web application server than the Business Logic server is with the database. Thus, the less chatty link is the one on a WAN. However, the four-tier architecture is more complex in that now there must be a web application server to service rich clients, and there must be one or more servers configured to run the Business Logic server pool. [Ref. 2]

This 4-tier deployment option is appropriate for large scale deployments, where flexibility and scalability are the key considerations. It has several advantages in this type of environment:

The hardware for each tier can be independently configured to suit its processing load and the type of processing it is doing. Multiple host machines can be used for each tier, to support scalability and failure tolerance. These can be added or removed at run time as necessary.

Clients can connect through firewalls and across wide area networks.

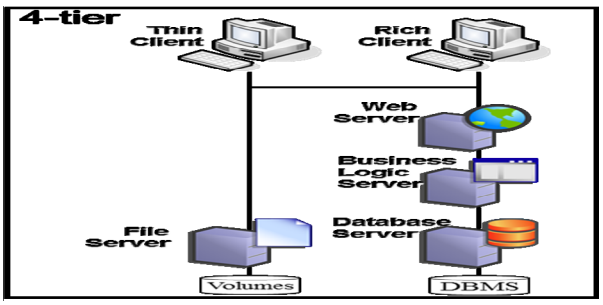

Figure2: Four Tier Rich Client Architecture

In summary, the four-tier architecture is intended to enable use in a WAN environment and two-tier in a LAN environment with very minimal latency.

III. Caching

Caching is a fundamental method of removing performance bottlenecks that are the result of slow access to data. Caching improves performance by retaining frequently used information in local high speed memory, which reduces access time and thus improves performance.

Optimized File Management System (FMS) solution design is essential to achieve good performance for PLM files management, especially in data transfer between clients and servers where the life-cycle processes of file authoring and editing are not necessary centralized in one location. Incorrect FMS solution will lead to serious performance issue and even cause data access failure. Need is to take full advantage of FMS features to improve data access performance based on limited network resources.

Figure3 shows the FMS deployment architecture components and their connections. PLM data centre will host the primary server cache which caches recently accessed files from the production volume. A secondary server cache is configured at the remote location to cache all the recent files accessed by the users located in that site. Each user workstation will host the FMS Client Cache (FCC) which will cache the recent files accessed by the specific user.

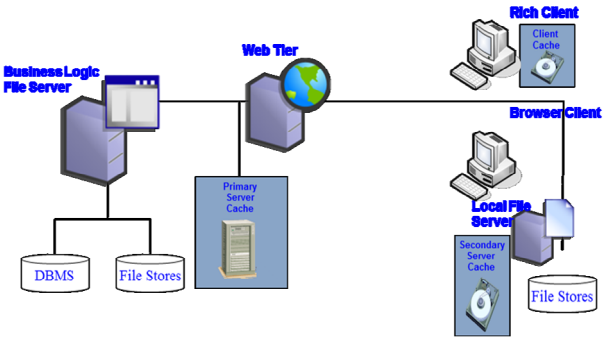

Figure3: File Caching Architecture

IV. Distributed Computing and Clustering

Distributed computing is used to improve the performance of operations that can be performed in parallel, by concurrently executing multiple operations. Operations may be distributed across multiple processes on a single CPU, taking advantage of multitasking, multiple processes across multiple CPUs, or across multiple machines. As operations are executed concurrently, ensuring synchronization between processes is essential to ensure correct results.

As the trend of increasing the potential for parallel execution on modern CPU architectures continues, the use of distributed systems is essential to achieve performance benefits from the available parallelism. High performance cluster computing is a well known use of distributed systems for performance improvements. Machines can be clustered in active-active or active passive modes as per the expected load on the specific deployment layer.

Accessing data from disk is very time consuming and every effort should be made to reduce disk access times. For large deployments, high throughput low latency SAN based file servers are highly recommended. Benchmarks have shown that for high load usage profiles, internal disk arrays may not be able to keep up, even with properly configured RAID arrays.

Multiple smaller, fast disk drives can be considered, rather than a few large drives. Data can then be spread across multiple drives. Spreading data across drives allows the OS to perform multiple drive operations at the same time, improving throughput. How many drives are appropriate is dependent on the type of drive, interface, and controller. Multiple disk controllers with several drives on each can be considered. This allows data to be spread across controllers as well, allowing data to be read and written in parallel. RAID configurations are to be configured to get improved (or at least do not degrade) throughput.

V. Load Balancing

A system can consist of independent components, each able to service requests. If all the requests are serviced by one of these systems (or a small number) while others remain idle then time is wasted waiting for used system to be available. Load balancing refers to arranging the systems in such a way that all systems are used equally. This can improve the over-all system performance.

Load balancing is often used to achieve further gains from a distributed system by intelligently selecting which machine to run an operation based on how busy all potential candidates are, and how well suited each machine is to the type of operation that needs to be performed.

Larger sites can distribute the pool of server processes across multiple hosts and include multiple HTTP servers to support load balancing. Load balance of FMS data is possible using external hardware load balancers. These devices exist within the network topology, but are not part of the FMS system. This load balancing method allows to route requests from multiple clients among a number of FSCs. The goal is to minimize overall response time at the requesting clients by routing requests to the most available server.

VI. High Availability

Given below are the high availability considerations for each layer of the PLM system deployment architecture:

- **Database.** Oracle can be clustered. Oracle in production for other applications can be clustered in an active-passive mode or in active-active mode. Such an approach would require a second Oracle server identical to the first, at each site where Oracle is running.

- **Business Logic.** It is straightforward to add a server to the Business Logic layer. In fact, it is by default configured as a cluster. For each site an extra server can be added so that if one server goes down, the load can be spread across the remaining servers.

The business logic layer also load-balances. Incoming requests to the web application tier are routed to the "least used" business logic server. "Least used" is a simple algorithm defined by (Number of Business Logic servers in use) / (Maximum number of Business Logic servers). In this way the business logic layer does not need to have symmetry in the server configurations; servers with different memory and CPU can be added to the cluster. However, each one must be configured so that the memory and CPU are not exhausted; basically the maximum number of Business Logic servers needs to be properly set.

- **Web Application.** PLM system is cluster-able at the Web Application layer; a second server would need to be added and clustered. Additionally, a load balancing front end would need to be added. On failure of a server a login would not need to be done, but the transaction would need to be re-started.

- **Volume / FMS.** An extra server would be added. It would be configured to pick up the functionality of *any* of the servers that might go down. There must be a process in place to start up the failover server processes and point the rest of the system to the changed server.

- **FSC Cache.** An equivalent can be added to act as a failover server. Since the FSC Cache will sit in individual sites it has to have a separate failover from the Volume / FMS. The FMS master configuration can be configured to be redundant, so all clients (including Rich Clients) can find the proper configuration.

- **License Server.** A second, equivalent server would need to be added.

Below is a list of methods to enhance the High Availability (HA) of system.

- Failover servers are configured for each tier, including DB server, Volume server, and license server etc.

1) Failure of one server is automatically detected by software and the failed server with all the applications are started up on another running server.
2) Network routing will handle smooth transition of subsequent transactions.

- Disaster Backup is required for database and volume server
- Implement Oracle database server cluster

1) CPU requirement of Oracle Server must be doubled to support this configuration

- Using Storage Area Network (SAN) to manage data
- Implement RAID 1+0 disk system
- Implement server cluster for each tier

1) Using Load Balancing that is supported by the third party hardware and software

- The network must have the ability to automatically reroute traffic across redundant links and devices, providing at least one alternative route around any single failed component. This way the application is always accessible.

- Every single component of deployment architecture should have appropriate redundancy.

VII. Virtualization and Cloud Computing

The use of virtualization is now an accepted IT practice. Virtualized servers appear in local server rooms, company private clouds and public clouds. Virtual machines can be used for each tier of the PLM 4 tier model. The key components to the virtual machines are the CPU, memory, storage and networking. VMWARE ESXi has the ability to provide standalone virtual machines or VAPPS that can be used to aggregate a set of machines into a specific named group. The use of 64 bit guest operating systems provides for greater resources being made available to the applications.

VMWARE ESXi provides several mechanisms to configure and adjust the allocation of CPU and memory resources for virtual machines running within it. Resource management configurations can have a significant impact on virtual machine performance. This section lists resource management practices and configurations recommended for optimal performance.

- Resource settings such as Reservation, Shares, and Limits are to be used only if needed in the environment. If frequent changes to the total available resources are expected, it is recommended to use Shares, not Reservation, to allocate resources fairly across virtual machines. If Shares are used and subsequently the hardware is upgraded, each virtual machine stays at the same relative priority (keeps the same number of shares) even though each share represents a larger amount of memory or CPU.

- Reservation can be used to specify the minimum acceptable amount of CPU or memory. After all resource reservations have been met, ESXi allocates the remaining resources based on the number of shares and the resource limits configured for the virtual machine. As indicated above, reservations can be used to specify the minimum CPU and memory reserved for each virtual machine. In contrast to shares, the amount of concrete resources represented by a reservation does not change when the environment changes, for example by adding or removing virtual machines.

- Reservation need not be set too high. A reservation that's too high can limit the number of virtual machines to power-on in a resource pool, cluster, or host.

- When specifying the reservations for virtual machines, it is recommended to leave some headroom for memory virtualization overhead and migration overhead. In a DRS-enabled cluster, reservations that fully commit the capacity of the cluster or of individual hosts in the cluster can prevent DRS from migrating virtual machines between hosts. When all capacity in the system are fully reserved, it becomes increasingly difficult to make changes to reservations and to the resource pool hierarchy without violating admission control.

- Resource pools can be used for delegated resource management. To fully isolate a resource pool, the resource pool type has to be Fixed and Reservation / Limit has to be configured. Virtual machines can be grouped for a multi-tier service into a resource pool. This allows resources to be assigned for the service as a whole.

PLM services can be impacted by the improper setting of vCPU's for the virtual machines. It is always best to start with the smallest effective number of vCPU's and add them as needed. The WEB / WEB APP / FMS servers generally work well with 2 vCPU's to split the work. The Pool servers should be monitored for CPU slow down if the number of vCPU's is greater than 3. This can be a combination of sockets and cores. The database servers such as Oracle and MS SQL will be dependent upon the guest OS that is used. Start with the smallest effective number and increase that number if the actual application load benefits from the increase.

PLM applications that rely on the network such as the 4 tier rich client, 4 tier web client and FMS are impacted by the amount of traffic that flows through the virtual machines and the host server. Adequate bandwidth, low latency and properly configured networks will reduce the overhead that users will experience. All tiers will benefit from a well configured network.

For small and medium enterprises more than any other group, costs, flexibility and performance are the key requirements when considering a Cloud computing solution as an IT infrastructure optimization alternative. Three basic categories of **Cloud Computing** services are identified: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Some advantages of Cloud Computing concept are: basically cost savings; scalability; "pay-peruse" model; independence from devices and location; efficiency; providing space for storage and control; probability and transparency of the processes; optimal utilization of resources, etc. [Ref. 3].

Conclusions

Although all PLM system component tiers can be deployed on the same machine of higher Spec-int-rate [Ref. 1] and Memory, in general performance can be optimized if each component was deployed on a separate server after proper sizing calculations for each layer of deployment. This allows each server to be tuned for that specific component.
It is always recommended to configure separate server machines for Database, PLM Application, Web application and Data Volume. This will provide greater flexibility in vertical and horizontal scalability including load balancing at each layer of PLM deployment.

In PLM implementation of J2EE Web Application Servers, the HTTP and J2EE servers are normally run as a combined process. There is little advantage to separating them, although it is possible to configure an http proxy server between the J2EE server and the clients, or to configure firewalls.

Because PLM Application can consume significant database resource, **it is recommended that no other applications are served by the database server**. If other database applications ***must*** be served from the same machine, there should be a dedicated database instance for the PLM application. This allows the designated instance to be tuned specifically for PLM deployment.

Measurements have shown that deployment of 2-tier clients over WAN results in very unacceptable performance. Any client not on the same LAN as the Web and Enterprise tiers should be deployed as 4-tier. Performance is noticeably slower if the network latency exceeds 5–10 ms between PLM application server process and database server. For optimal performance, it is recommended that the PLM application server reside on the same Gbit LAN as the database server.

References

[1] http://www.spec.org/cpu2006/results/

[2] **S. El Kadiri, M. Delattre, P. Pernelle, and A. Bouras,** "*A complementary generic architecture for PLM system to control collaborative work",* The International Conference on Software, Knowledge and Information Management and Applications, Fes: Morocco, 2009.

[3] **Emilija RISTOVA, Valentina GECEVSKA**, "*PLM AND BUSINESS PERSPECTIVES OF THE NEW PARADIGMES IN INFORMATION TECHNOLOGY",* ANNALS OF FACULTY ENGINEERING HUNEDOARA – International Journal Of Engineering, Tome X, Fascicule 3, Pp307-311, Year 2012.

[4] **Michael Greaves**, *Product Lifecycle Management: Driving the Next Generation of Lean Thinking*, NewYork, McGraw-Hill, (2006).

[5] **M.S.Gopinatha and Dr.Vishnukant S.Chatpalli**, "*Implications of Globalization on product design IT systems in Automobile Industry*" PDMA India IV annual International conference, NPDC 09 "New Product Development: Challenges in meltdown times", Department of Mechanical Engineering and Department of Management studies, IIT, Chennai, India, pp. 94-101, 17-19th Dec 2009 (www.npdc.iitm.ac.in).

[6] **M. S. Gopinatha, Dr. Vishnukant S. Chatpalli and Dr. K. S. Sridhar,** "*Survey of factors influencing the performance of PLM system in Auto Industry*", International Journal of Research in Computer Applications and Management (IJRCM), Volume no. 2 (2012), Issue no. 12, ISSN 2231-1009, pp. 47-52, December 2012, (http://ijrcm.org.in).

[7] ***M. S. Gopinatha, Dr. Vishnukant S. Chatpalli and Dr. K. S. Sridhar,*** "*Experimental research on Influence of Hardware infrastructure sizing on optimization of PLM system performance in Auto Industry*", International Journal of Engineering Research and Technology (IJERT), Volume 2, Issue 3, ISSN 2278-0181, March 2013, (http://ijert.org).