

# A Networked Computer Architecture for Life Sciences Discovery and Commercialization

DRAFT-DRAFT-DRAFT-DRAFT-DRAFT

Perseid Software Limited  
Needham, Massachusetts

September, 2002

Table of Contents

- 1 EXECUTIVE SUMMARY ..... 3**
- 2 INFORMATION COMPLEXITY IN LIFE SCIENCES ENTERPRISES ..... 4**
- 2.1 Introduction—Supporting The Value Chain in Product Development ..... 4
- 2.2 Information Complexity in the Life Sciences ..... 5
- 2.3 Traditional Computer and Data Architectures Lack Integration ..... 5
- 2.4 A Plan for Enterprise Data Integration ..... 6
- 2.5 A Model for Real-time Life Sciences Systems Integration ..... 8
- 2.6 Systems Integration within the Model IT Architecture ..... 9
- 3 CONCLUSION ..... 12**
- 4 ABOUT PERSEID SOFTWARE ..... 13**

Figures

- Figure 1 A Traditional Architecture Lacking Integration ..... 6
- Figure 2 Life Sciences Enterprises Generate Tiers of Complex Data ..... 8
- Figure 3 A Model Architecture for Integration of Applications in the Life Sciences Enterprise ..... 10
- Figure 4 A Model Architecture for the Enterprise and Local Networks ..... 11

# 1 Executive Summary

## 2 Information Complexity in Life Sciences Enterprises

### 2.1 Introduction—Supporting The Value Chain in Product Development

The IT infrastructure supporting a life sciences product throughout its life-cycle is very complex. Developing a new chemical entity into a commercial pharmaceutical requires complex partnerships and is based on a complex regulatory process that can last a decade. The information systems supporting this process must support research, development, manufacturing, marketing, sales, clinical trials, the regulatory process and post-market data capture.

For example, examining the life sciences discovery process within the domain of central nervous system (CNS) product development, there has been an explosion of new acquisition devices and databases. In a recent article in *Bio-IT World*,<sup>1</sup> Dr. Stephen Wong of the University of California, asserts that new clinical procedures such as lab tests and neuropsychological exams, new structural imaging techniques such as magnetic resonance imaging (MRI) and angiography, x-ray computed tomography and electron microscopy, functional and metabolic imaging methods such as positron emission tomography, magnetic resonance spectroscopy, functional MRI and optical imaging, while essential, each has complicated the systems and data integration problems in CNS discovery. These examination and imaging techniques are also being combined with high throughput genomic techniques such as DNA micro arrays.

Other requirements beyond the discovery process within R&D include documentation for clinical research and trials and, most importantly—sharing of data among contractual partners in all aspects of product development and launch.

As a result, the IT infrastructure shared among partners during development and product introduction can be exceptionally complex and dispersed among many parties. Computing platforms can literally range from Apple Computers to IBM or Cray supercomputers, resulting in numerous disparate storage and computing systems.

Finally, the data and image storage management and administration problems in the life sciences can become daunting. Multiple storage types—files, records, databases, images, documents, etc. must be supported. Multi-vendor operating systems must be supported. Multiple computing platforms must be integrated and managed. Disparate storage systems, network attached (NAS), storage area network (SAN), and enterprise-level must be integrated. Data systems and structures from simple files to complex relational databases must be supported.

This white paper will examine the complexity associated with multi-vendor and -development partner support in the life sciences and propose how multi-vendor software and hardware can be integrated to manage a complex IT architecture for discovery, development and production in the life sciences.

---

<sup>1</sup> Wong, “Neuro-IT Needs Integrated Infrastructure,” *Bio-IT World*, July 11, 2002.  
[http://www.bio-itworld.com/archive/071102/horizons\\_neuro.html](http://www.bio-itworld.com/archive/071102/horizons_neuro.html)

## 2.2 Information Complexity in the Life Sciences

To design, test, deploy and manage during new drug discovery, or to re-market an older pharmaceutical, data must be integrated, managed and deployed to serve multiple simultaneous purposes. To manage in real-time requires the integration of complex enterprise information:

- ❖ Biological — Proteomic, genomic and other biological and chemical information
- ❖ Financial — Design, development, pricing, deployment and marketing data
- ❖ Research — Molecular, genetic, proteomic and pharmacological data
- ❖ Clinical — Clinical trial, side effect, outcomes and post-market effects of drugs
- ❖ Validation — Strategic planning data on the validation of the drug for the target market
- ❖ Market — Marketing information by country and metropolitan area
- ❖ Manufacturing — Integrated GMP<sup>2</sup> information on the manufacturing process
- ❖ Administrative — Healthcare claims, membership, diagnostic and treatment data
- ❖ Metadata — Information about what is contained in the these phases of development

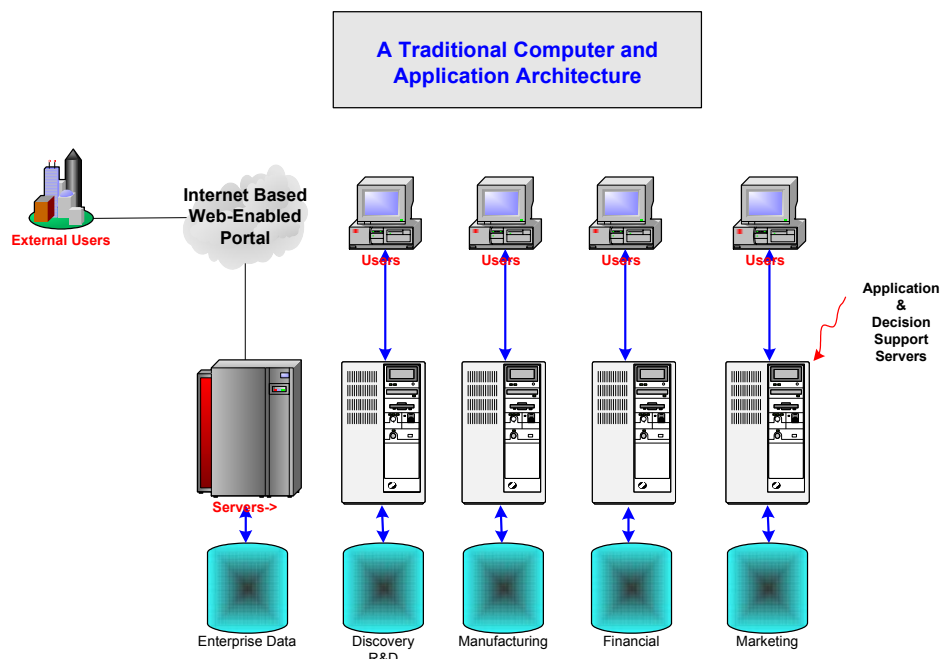
## 2.3 Traditional Computer and Data Architectures Lack Integration

**Error! Reference source not found.** highlights the “traditional” systems architecture of life sciences information systems. The computer and storage systems are not integrated and each tends towards duplication of data and systems resources. Each information system demands its own processing resources and storage architecture. The duplication of the resource is expensive and encourages a lack of data integration in the R&D value chain.

---

<sup>2</sup> Bernard P. Wess, Jr., “Building Mission Critical Document Management Solutions for Global Pharmaceutical Companies, EMC White Paper, Perseid Software, June, 2001.

[http://www.emc.com/vertical/pdfs/life\\_sciences/interstitial\\_3.jsp](http://www.emc.com/vertical/pdfs/life_sciences/interstitial_3.jsp)



**Figure 1 A Traditional Architecture Lacking Integration**

Figure 2 presents a contrary view focusing on *integration* of global management, financial, marketing and clinical data and documents. The data and documents are stored to support an integrated view of the pharmaceutical enterprise and its product development processes. Market segment research can be conducted at the population level and tracked over a period of years. Integrated clinical, biological, financial, regulatory and efficacy documentation and data are all tracked at multiple levels of the enterprise. Because of the massive size of the central repository, the “layers” of the data pyramid may be implemented as multiple physical databases. This data architecture makes it easier to reliably merge clinical, financial, marketing and patient/consumer data into an accurate central reporting system. What is needed in this form of real-time enterprise computing<sup>3</sup> is the integration of the data into a common repository, using shared “metadata.”<sup>4</sup> This improves data integration, removes redundancy and encourages regulatory and partnership data sharing.

## 2.4 A Plan for Enterprise Data Integration

Figure 2 describes the data model for the integration of information across the value chain of the life sciences enterprise, that is, a multi-vendor, multi-database architecture but one with common data integration. Two “stacks” of data are articulated in the figure:

- ❖ **Biological and Clinical** — The data and information supporting discovery and R&D of new chemical entities.

<sup>3</sup> Bernard P. Wess, Jr., “Enabling the Real-Time Life Sciences Enterprise with an IT Infrastructure,” EMC White Paper, Perseid Software, February 2002.

[http://www.emc.com/vertical/pdfs/life\\_sciences/interstitial\\_data\\_warehouse.jsp](http://www.emc.com/vertical/pdfs/life_sciences/interstitial_data_warehouse.jsp)

<sup>4</sup> “Metadata” is information in the data systems that identifies the contents of the data itself—its fields, tables, reliability and validity, for example.

- ❖ **Administrative and Financial** — The data and information supporting the processes of approval, production, distribution and post-market surveillance.

These segments of the value chain in the life sciences *require* integration. Speed-to-market of new pharmaceuticals is about having the right information at the right time for the right person in the development process. Moreover, many partners must share, selectively, data. Discovery partners need access to internal systems of the pharmaceutical company. Clinical trials partners may need to use past clinical data. Regulators may need access to a variety of clinical and market data in a confidential manner.

The need for “selective transparency”<sup>5</sup> of information systems first proposed by W. Roy Dunbar, CIO of Eli Lilly drives the need for integration of systems and data. Selective transparency would allow an employee, research partner, regulator, or any authorized entity or person access to aggregate data or the ability to execute a transaction, program or report—based on role and security, access and control criteria.

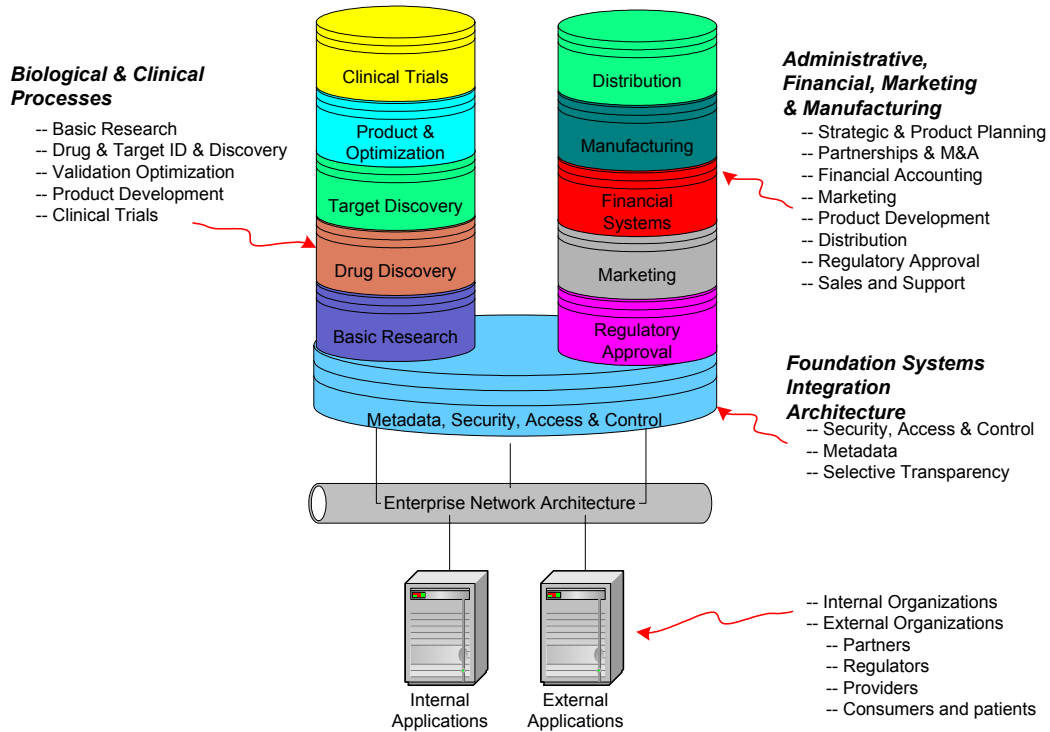
But the key to selective transparency is having an integrated information systems architecture that allows for and encourages real-time access to needed information and raw data. An integrated computer systems architecture facilitates business and operational integration by encouraging employees, partners and others to quickly register and share data. Moreover, when in place, such an architecture facilitates sharing data with regulators and others who are outside the literal boundaries of the life sciences enterprise. Given that life sciences discovery and commercialization is a national or even a global exercise, the need for registering data and sharing it effectively is profound—both in terms of costs and managerial complexity.

Below we show the dual “stacks” of enterprise data: focusing on delineating the needs within numerous areas of the life sciences value chain. These areas include the chemical, biological, clinical and validation processes associated with developing a pharmaceutical solution and the administrative, financial, marketing, manufacturing and distribution needs, once a product requires preparation for the market. Note that the foundation architecture for selective transparency is implemented as a core data integration function. Data is registered by form and type (as metadata) and then security, access and control constraints are applied to the data to register its presence in the enterprise and to provide selective access to interested parties. A key aspect of effective registration and distribution is to register a data asset once and only once as a unique version. This is a primary role of the Enterprise (ESA) and Network (NSA) Storage Architectures. These architectures avoid the proliferation of multiple versions of files and databases which is a classic means of destroying effective data integration in an enterprise.

---

<sup>5</sup> Mark D. Euhling, “The Pharma Prophets”, Bio-IT World, April 7, 2002. [http://www.bio-itworld.com/archive/040702/boston\\_it\\_pharma.html](http://www.bio-itworld.com/archive/040702/boston_it_pharma.html)

**The Data “Stack” Enabling the Life Sciences R&D Value Chain**



**Figure 2 Life Sciences Enterprises Generate Tiers of Complex Data**

The metadata of the ESN and NSA identifies who has access to data and the circumstance and means by which the data may be accessed. Thus, the systems integration process becomes very much a process of data integration—driven by appropriate security, access and control considerations and software.

The foundation in Figure 2 of the architecture is security, access, control and metadata resources that facilitate the systems integration process. New systems and data are registered in the enterprise and data access is provided through selective access controls that allow departments, partners and others selective access to data depending on their validated needs and rights. Data is secured, integrated and available for re-use at all times. Thus data integration leads to the next step—systems and applications integration.

## 2.5 A Model for Real-time Life Sciences Systems Integration

The evolution of the poorly integrated databases and applications in Figure 2 to the next generation real-time architecture is shown in Figure 3. All users pass through a security, access and control application that provides for common validation and access to production files and databases. Validation creates a “view” of the information systems for the users that is driven by their access rights and roles in the

organization. Given higher access rights and the more complex one's role—the higher the availability of applications and data.

## 2.6 Systems Integration within the Model IT Architecture

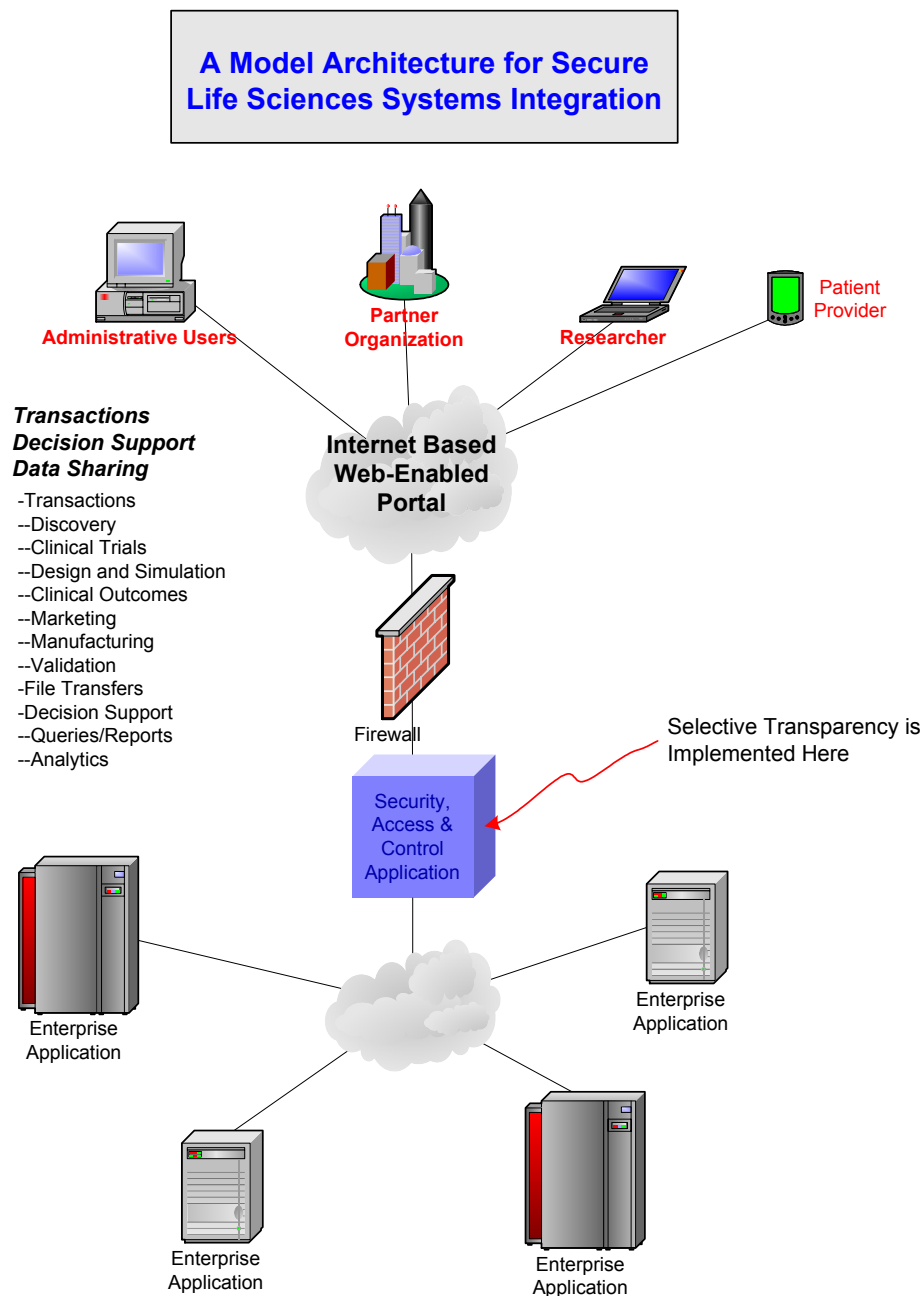
Figure 4 depicts a fully integrated foundation for systems integration of storage systems in the Life Sciences. This example is based on EMC's "E-Infostructure" which is composed of physical, connectivity and functional layers of hardware and software supporting one or more integrated database management systems with a common logical database design architecture and integrated metadata. Two forms of persistent data are present in the life sciences enterprise:

- ❖ **Record-oriented data** — Data in files and databases based on a "record" or series of fields
- ❖ **Persistent objects** — Images, documents, x-rays, and other clinical, administrative and financial "objects" that require persistent storage.

The physical layer for record-oriented data includes EMC Celera™ or Symmetrix™ for the ESN and CLARiiON™ for NSA storage systems and they provide the basic foundation for performance, capacity, availability and other physical requirements of the central repository and database management systems. The Enterprise Storage Network ("ESN") is implemented as a connectivity layer through two information connections—Connectrix™ for the Symmetrix and Celerra™—it provides a means of using all primary and secondary operating systems to connect into the enterprise application and data management platforms. These systems could include IBM operating systems, Unix operating systems, including Linux and those from Compaq, Sun, HP, and Microsoft operating systems.

The storage management layer is composed of the ESN and the Database Management System(s). EMC Symmetrix™ and enterprise storage management software are used to integrate all operating systems and database storage from Oracle™ and IBM® into a uniform central repository. EMC Connectrix™ switches for the ESN or EMC Celerra HighRoad™ software are used to handle multiple connections to the servers. EMC TimeFinder™ software provides each party with local copy of certain data to analyze, thereby increasing researcher productivity.

TimeFinder can be used to refresh data warehouses with timely information without disrupting production systems. The remote mirroring capabilities of EMC's SRDF™ (Symmetrix Remote Data Facility) software can protect enterprise databases and other critical data to avoid costly interruptions in speed to market. EMC TimeFinder is used for non-disruptive backup and data warehouse loading with EMC SRDF used for disaster recovery and information mobility.



**Figure 3 A Model Architecture for Integration of Applications in the Life Sciences Enterprise**

The EMC NSA and ESN solutions support continuous availability of reporting, decision support and web access to the enterprise repository and continuous availability of databases at the core of the federated data base management system.

The storage architecture for persistent and unique data *objects* is based on “content” addressability (CAS) solutions—each data object is given a universally unique ID. EMC Centera™ creates a unique identifier, based on the attributes of the content, which applications can use for retrieval. Centera is the world's first CAS solution designed to meet the unique requirements of fixed content management. Centera provides

fast, easy online access and petabyte scalability for a wide range of digital assets including: X-rays and MRIs, GMP business documents, marketing broadcast content, and life sciences discovery data. As an integrated hardware and software system, a single administrator can manage up to 160 TB of stored content. To assure content integrity and authenticity, Centera gives each stored object a unique address. And, there are no duplicates: only one copy and one replica of each object are stored, no matter how many times the object is used. This means one discovery object and annotation, one GMP 21CFR11 document and one form image, regardless of the number of applications and users accessing the data object.

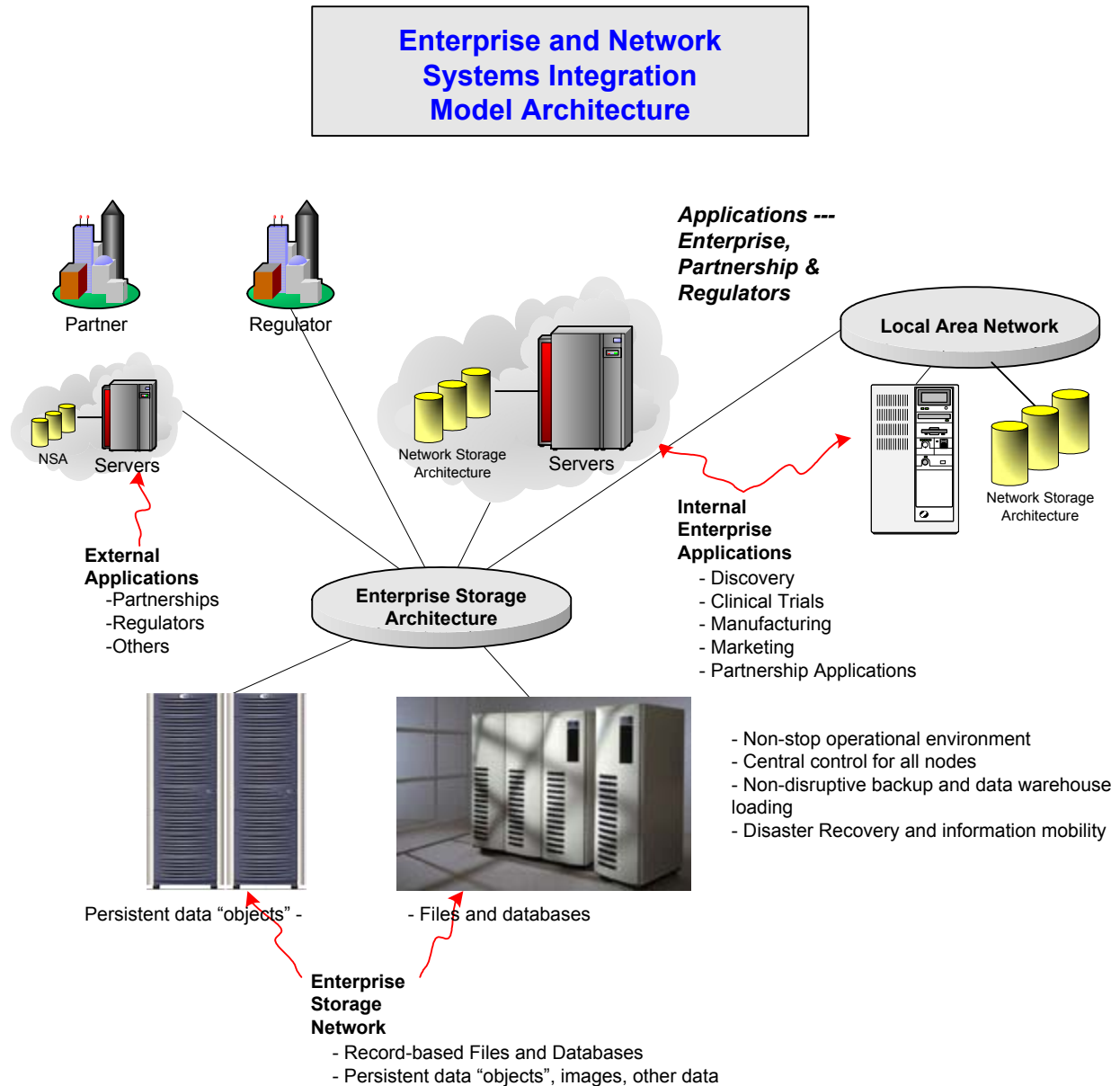


Figure 4 A Model Architecture for the Enterprise and Local Networks

The applications and databases are divided among production transaction processing and business intelligence applications and enterprise departments and partners. New data arrives from transaction processing systems in Figure 4 or from the transfer of other databases—from partners, research organizations or third-party data vendors. The systems integration architecture is responsible for ensuring that local applications and databases can submit data to the central enterprise storage network (ESN) for registration and integration or managing the local network storage architecture. Using the security, access and control solutions, data is integrated in the ESN or NSA for use by other parties that ensures that the data is registered, secured and validated so that it can be accessed by all authorized local and global users.

### 3 Conclusion

The rapid growth in the development of automated systems in the value chain within the pharmaceutical industry has resulted in a lack of systems and data organization within the extended enterprise that constitutes the modern pharmaceutical or life sciences enterprise. Not only is there a lack a standardization of data gathering and reporting, but also there has been a tendency to accumulate disparate applications and computer processors and storage without first developing a systems and computer architecture that encourages data and systems integration.

As a result, there is difficulty in moving data across boundaries within the life sciences R&D and development chain. Moreover, there must be security, access and control systems across the entire contractual and organization chain to facilitate sharing the right data with the right party under the right circumstances and at the right time.

Implementing systems to selectively allow access on a full- or part-time basis to integrated data is the solution to many of the problems plaguing systems integration in the life sciences. Using advanced operating system and storage technology from EMC, we have developed a model systems integration solution that integrates disparate data files, data bases and operating system platforms into both Enterprise Storage and local Network Storage architectures. The ESN or NSA manages, with proprietary software, access to data which is either transient or persistent. These objects may be files, databases or persistent data “objects”. Two forms of storage subsystems exist on the ESN—traditional files and databases and “content-addressable” data objects. Files and databases are well known storage objects, but content addressable storage is a new concept in storage systems. The persistent data objects are produced during discovery, documentation, patient clinical trials and other activities that create a data object that must be registered and stored once.

The model architecture, when combined with data standards that must emerge in the life sciences value chain, can become a standard for creating, storing and effectively sharing enterprise-scale data within the life sciences industry. Well formed and easily accessible data should improve the discovery, development, marketing, production and sales processes in the pharmaceutical and biotechnology industry. The alternative is “data” chaos and given the costs of modern life sciences discovery and development—clearly unacceptable to all parties in the R&D and development processes. For more than 1,000 years the construction industry has used *architectural planning* prior to development. Given the risks associated with life sciences discovery and pharmaceutical development, no less should be done within the life sciences among internal departments, partners, the regulatory agencies and healthcare care providers.

## 4 About Perseid Software

Perseid Software is engaged in providing strategic consulting and information technology design services to healthcare and life sciences enterprises. For more than 30 years, the principals of Perseid Software have been engaged in the development of mission-critical information systems and in the analysis of healthcare, disability and pharmaceutical data.

Perseid Software is not merely a strategic consulting firm. It is an engineering management and design firm focusing on database design and implementation of very large and complex life sciences and healthcare information systems. Perseid's clients include or have included some of the largest and most progressive computer, healthcare and manufacturing companies in the world.

Contact:

Bernard P. Wess, Jr., President  
Perseid Software Limited  
Needham, MA  
Direct Dial:(781) 453-2351  
bwess@perseidsoftware.com  
[www.perseidsoftware.com](http://www.perseidsoftware.com)

IBM®, z/OS™, OS/390™, AIX™, MVST™, AS/400®, AIX/HACMP™ are trademarks of the IBM Corporation  
Sun Solaris™ is a trademark of Sun Computers. Microsoft®, Windows NT™, Windows 2000/Data Center™ are trademarks of Microsoft Corporation. Centera™, Symmetrix™, CLARiiON™, Connectrix™, Celerra™, Symmetrix Remote Data Facility™, TimeFinder™ Software, EMC Foundation Suite™ and Database Edition for Oracle™ are trademarks of EMC Corporation, E-Infostructure® is a registered trademark of EMC Corporation. Oracle Open Parallel Server™ is a trademark of Oracle Corporation, Compaq® is a trademark of Compaq Computers, Unisys® is a trademark of Unisys Corporation.