

Using High Performance Computing Clusters to Accelerate Life Science Solutions

A Cabot Partners Executive White Paper

Sponsored by IBM

Analysts: Srinu Chari and Eleanor Haas

Life Science Innovation Makes New Demands on IT

Life science innovation supplies the market demand of consumers and providers for ever-more sophisticated health care solutions – a demand that has reached unprecedented heights. Fortunately, the recent discovery and development of digital tools for researchers and scientists has opened new doors to innovation. These new doors have been critical to the growth of systems biology, which goes beyond molecular biology's focus on macromolecules and DNA to model a living system *in silico*. Systems biology, in turn, is making new demands of computing capabilities.

These new computing capabilities need to be cost-effective as well as powerful because of the economics of health care and life science. The discovery and development of drugs and therapies has become ever more technically complex and expensive. As a result, new drugs and therapeutics drive up the cost of both care and health insurance. Health care costs already exceed sustainable levels and yet continue to rise. Using industry averages reported to the Pharmaceutical Research and Manufacturer's Association, the cost of drug development increased from \$4 million in 1962 to \$450 million - \$700 million in 2000. The current estimates are in the \$1.2 billion - \$1.5 billion range.¹

¹ Tufts Center for the Study of Drug Development, <http://csdd.tufts.edu/>

Copyright © 2007. Cabot Partners Group, Inc. All rights reserved. Other companies' product names or trademarks or service marks are used herein for identification only and belong to their respective owners.

New IT Solutions Address Life Science R&D Needs

Clusters and Scalable Parallel Computing Are the Answer

To process the massive amounts of data generated by bioscience R&D requires massive horsepower. This fuels the need for clusters, grids, and scalable parallel computing using multi-core processors, i.e. high-performance computing (HPC).

Cluster computing integrates off-the-shelf commodity computers and resources through hardware, networks and software so that all components behave as a single computer. Scalable cluster systems are tightly coupled computers with high bandwidth and low latency interconnects between processors and even storage with an optimized message-passing library, such as MPI - Message Passing Interface². Linux has become the dominant operating system in HPC, representing 52 percent of the overall high performance technical computing market and even more in departmental or divisional environments, according to IDC.³

Cluster computing goes beyond single-application parallel computing to incorporate load-balancing clusters and high availability clusters. The main benefits are scalability, availability, affordability, and performance.

Cluster systems can be expanded with standard microprocessor nodes to meet the processing needs of dramatically growing data and of the exceptional accuracy demanded by bioinformatics, systems biology, and imaging problems. In addition, the price/performance ratios of these systems can make teraflops of processing power available at a fraction of the cost of a traditional supercomputer.

Large-Scale Life Science Applications Have Special Requirements

HPC is almost always accomplished through parallelism. However, obtaining parallel computing capabilities is difficult and complex because many practical life science applications don't multithread beyond a few processes. In order to scale further, parallelism must be at a very high coarse grain level. Algorithmic approaches in life science that are based on data and problem decomposition strategies allow users to obtain the maximum advantage and scalability from parallel machines with large numbers of interconnected processors. Almost all classes of life science

² Argonne National Laboratories, "The Message Passing Interface Standard (MPI)", <http://www-unix.mcs.anl.gov/mpi/>.

³ IDC Research VP Christopher Willard, Ph.D., ISC 2006, reported in HPC Wire, June 30, 2006

applications in bioinformatics, chemistry, imaging, and systems biology benefit from cluster computing.

Careful System-Specific Tuning is Needed to Further Boost Performance and Scalability on Cluster Platforms

Parallel application development and system performance for large-scale life science applications also depend on the single processor and memory performance, the communication subsystem performance, input/output (I/O) performance, and development tools for programming, debugging, and resource management. A significant improvement in performance and scalability can be obtained for life science applications that are tuned and optimized for the specific parallel architecture. This often requires a combination of deep life science domain knowledge coupled with algorithmic and parallel computing skills.

“White Box” Cluster Solutions Are Becoming Inadequate

Many chemistry codes require cluster nodes with large memory and high performance I/O. At the same time, many bioinformatics and molecular dynamics applications require large number of cluster nodes tied together with high-performance scalable interconnects. Furthermore, a large number of applications are tied together into a computational workflow that requires additional middleware for scheduling, resource management and provisioning.

“White-box” clusters are becoming inadequate to solve the growing compute requirements for complex life science simulations. When they have hundreds of processors, they are expensive to deploy and operate. The cost associated with providing support and maintenance grows exponentially. Also, management of such diverse collections of resources is difficult, and effective software solutions that can scale are only now beginning to appear in the market. Furthermore, the electrical power and the physical facilities required to operate such large clusters are prohibitively expensive.

Integrated Clusters Are Effective in Processing Massive Life science Data

Computational life science involves solving a wide range of biological, chemical, imaging, and biological systems problems that are data and computationally intensive. Gene expression array and high-throughput sequencing technologies, along with related diagnostic imaging - X-ray, MRI, PET, etc. - generate massive digital medical and clinical data. Parallel pattern recognition and data mining algorithms based on data-decomposition and distributed query techniques on scalable parallel

systems are required to glean insights from these vast mountains of information.

Tighter Integration with Devices and Diagnostics Expands Value of Bioscience R&D

Life science data acquisition systems and clinical devices and diagnostic systems produce vast quantities of data. These systems are becoming better integrated with life science simulation and modeling software. Often these devices support desktop Microsoft Windows environments that need to be better integrated with a high-performance computing back end in order to enable computational workflows. Integration enables the effective and rapid deployment of Laboratory Information Management Systems (LIMS) for the life science. The availability of flexible and cost-effective cluster computing solutions compatible with desktop environments promises to further energize bioscience R&D.

Overview of HPC Life Science Applications

Computational and analytical sophistication is increasingly critical for drug discovery, development, and medicine. Extensive and increasing interdisciplinary studies and workflows are common in systems biology. These studies and workflows combine biology, chemistry, imaging, and statistical analysis. Together they drive up the demands on computing and data management capabilities.

Computational Life Science Problems Span a Wide Spectrum of Disciplines

Dramatic advances in information technology continue to enable new advances in *in silico* biology, high throughput screening, chemistry, imaging, and systems biology. As these disciplines mature, researchers become proficient at building sophisticated mathematical models and using modern computing infrastructures.

Most life science problems lend themselves to solutions in a clustered or parallel environment. Industrial life science environments rarely rely on just one kind of application. Instead, they often run a mix of public domain bioinformatics software and computational chemistry software largely provided by a few key commercial software developers such as Accelrys, Schroedinger, OpenEye, and others.

Many specialized programs for more sophisticated analytical applications written in-house or elsewhere in the industry are also used for emerging areas such as imaging and systems biology. The application disciplines that benefit from high performance computing solutions are biology, chemistry, clinical trial management, imaging, and business intelligence.

The life science world is experiencing explosive growth in both the volume and diversity of data. Structured sequence data is typical in many bioinformatics applications whereas unstructured text and image data is common in medical imaging and clinical development. Furthermore, emerging sub-disciplines, such as systems modeling, high content analysis, and mRNA profile analysis, are driving even more demand for computing capability. The following tables provides a summary classification of some major life science applications and solution approaches for both more established areas such as bioinformatics and computational chemistry and for emerging areas such as imaging and systems biology.

| Discipline | Solutions | Data/Application Characteristics | Major Applications |
|--|--|---|---|
| Bioinformatics – Sequence Analysis | Searching, alignment & pattern matching of biological sequences (DNA & protein) | Structured data - integer dominant, frequency dependent, large caches and memory bandwidth not critical, some algorithms are suited to SIMD acceleration. | <ul style="list-style-type: none"> ▪ BLAST ▪ ClustalW ▪ FASTA ▪ HMMER |
| Bioinformatics – Information Management | Storage, retrieval & content serving of diverse biological data | Mostly integer-oriented, some analysis tools can be floating point dominant. Structured and unstructured data including image data. | <ul style="list-style-type: none"> ▪ Business Objects ▪ Medidata ▪ Oracle ▪ SAP ▪ SAS |
| Biochemistry – Drug Discovery | Screening of large database libraries of potential drugs for ones with desired biological activity | Mostly floating point, very compute intensive, highly parallel. | <ul style="list-style-type: none"> ▪ Autodock ▪ Dock ▪ Flexx ▪ FTDock ▪ LigandFit |
| Computational Chemistry – Molecular Modeling | Modeling of biological molecules using Molecular Dynamics & Quantum Mechanics techniques | Very floating-point intensive, latency critical, frequency dependent, scalable to low 100s. | <ul style="list-style-type: none"> ▪ AMBER ▪ CHARMM / CHARMM ▪ Gaussian ▪ GROMACS ▪ NAMD ▪ NWCHEM |

| | | | |
|-----------------|--|--|---|
| Clinical Trials | Collecting, managing, and analyzing structured and unstructured data to comply with and track regulatory submissions | Mostly integer-oriented, some analysis tools can be floating point dominant. Large data warehousing with complex queries across multiple data sources. | <ul style="list-style-type: none"> ▪ Business Objects ▪ Oracle ▪ SAS |
|-----------------|--|--|---|

Figure 1: Classification of prominent established HPC life science disciplines

| Discipline | Solutions | Data/Application Characteristics | Major Solutions Providers |
|---|---|---|--|
| Molecular Imaging – Biological, cellular, and targets | High content image analysis, storage, retrieval, and decisions using statistical analysis | Large unstructured data - 32 - bit integer and floating point dominant, frequency dependent, large caches & memory and storage performance critical, many algorithms are suited to SIMD acceleration. | <ul style="list-style-type: none"> ▪ Cellomics ▪ Definiens |
| Diagnostic Imaging – Information Management | Storage, retrieval & content serving and analysis of diverse diagnostic image data | Large unstructured data from MRI, x-ray, PET, etc. - mostly integer oriented, some analysis tools can be floating point dominant. | <ul style="list-style-type: none"> ▪ GE ▪ Siemens ▪ Philips ▪ Agfa |
| Systems Biology | Integration and analysis of complex data from multiple experimental sources using interdisciplinary approaches in proteomics, metabolomics, transcriptomics, etc. | Both integer and floating point intensive, very data intensive, could be highly parallel at a very coarse grain level. Needs very highly-scalable systems. | <ul style="list-style-type: none"> ▪ ISB ▪ Medical research |

Figure 2: Classification of some emerging HPC life science application disciplines

Life Science Requires Systems Optimized for Capacity and Large Scale Coupled with Devices and Diagnostics

The solution of a single life science problem frequently requires the integration of several applications and data sources from (a) multiple proprietary and public domain databases in various formats and (b) a wide range of devices and diagnostics equipment. A tighter integration of this workflow with data from devices and diagnostic equipment enables increased scientific innovation and productivity. Bioinformatics and chemistry applications typically benefit from scalable computers, whereas software applications involved in data acquisition from devices and diagnostic equipment run on smaller systems with superior graphics and visualization capability.

Advanced IT Solutions from IBM for Life Science Organizations

IBM Collaborates with Life Science Application Developers to Drive Innovation

IBM actively collaborates on a worldwide basis with leading life science applications developers at industry and research institutions to migrate and optimize their applications on the IBM Cluster 1350⁴ in order to solve challenging problems.

Life science applications that are mapped, migrated, and optimized for the IBM cluster architecture benefit greatly from the scalability and increased performance provided by IBM cluster systems. This is true in both existing systems and technologies based on the IBM cluster architecture and future generations.

Prominent public domain, proprietary, and business partner applications are currently being optimized to run life science applications, with future plans for pre-engineered and optimized Linux cluster configurations based on sizing expertise.

Experience to date, as illustrated by the examples that follow, shows that large-scale bioinformatics and molecular dynamics problems perform well on the IBM cluster and that the system also enables smaller analysis problems that are in the workflow to achieve outstanding performance.

IBM Solutions Enable Diverse Innovations

The following IBM solutions are enabling innovations in rational drug design, medical imaging, multiple molecular dynamics simulations, genomic

⁴ The IBM Cluster 1350 is a robust, completely integrated solution, which is described more fully beginning on page 13.

sequencing, pharmacogenomics, clinical and content management, and compliance.

Rational Drug Design on IBM Clusters

IBM installed a Linux cluster that can perform more than 600 billion operations per second at St. Jude Children's Research Hospital. The hospital, founded by the late entertainer Danny Thomas, is internationally recognized for its pioneering work in finding cures and saving children with cancer and other catastrophic diseases. The supercomputer made it possible for St. Jude's to use AMBER software, which is compute-intensive. AMBER simulates the manual process that St. Jude researchers were performing in the lab to determine the best possible compound to address a specific disease. Even more important than the significant cost savings, the hospital realized a new ability to reduce the development time of a drug that could save the life of a dying child from 6 years to an unprecedented 12 months.

Medical Imaging Computing and Storage Solutions

Medical imaging is revolutionizing the drug discovery process. It promises to shorten the process dramatically and has the potential to significantly reduce the cost of drugs – all thanks to HPC.

Johns Hopkins scientists are using medical imaging to interrogate the brain's entire system of neuroreceptors in living people – something that previously was possible only in dead brains. The use of medical imaging to see and analyze neurochemical behavior requires huge computational time to automate key operations and manage data flow.

Medical imaging makes it possible to identify disease biomarkers and to develop ways of blocking negative effects of schizophrenia, Tourette's Syndrome, drug addiction and other diseases, and even of improving cognitive function. Applying these capabilities to identify and validate compounds for central nervous system drug development also represents an opportunity to get rid of clinical trials involving human beings as well as to dramatically shorten the time required for drug development. Imaging has also been effective in the treatment of cancer and cardiovascular diseases.

Virtualized hierarchical storage pools physical storage from multiple network storage devices into what appears to be a single storage device. The storage device is managed from a central console for sharing and utilizing data while maintaining the capacity to expand later.

The Dundee (UK) University College of Life science now has a virtualized hierarchical storage environment from IBM and Tectrade that makes cost-effective use of both disk and tape storage while providing a high-performance environment for image processing in the HPC cluster. It enables storage of the right information on the right media with little manual

intervention. This saves a great deal of time, keeps staffing and expansion costs low, and makes optimal use of researchers' time. In addition, because microscopy images can be easily and cost-effectively stored and accessed, the Dundee system enables new research projects to exploit historical data.

Molecular Dynamics and Systems Biology

Simulations of complex biological processes at various organizational levels, from atomic to cellular and beyond, have enabled dramatic advances in genomics, proteomics and structural genomics through virtual screening – made possible through advances in computing architecture and scale.

The Mayo Clinic Medical School is using multiple molecular dynamics simulations (MMDS) to convert genetic sequences to 3D protein structures for *in silico* protein screening. It uses a teraflop computer system designed for this purpose. The SARS 2003 epidemic was projected through this process. The projection involved 420 molecules deployed simultaneously.

MMDS have reduced the timetable for lead discovery by orders of magnitude. Sequencing took 31 days in 2003 but only a single day in 2005. These molecular dynamics simulations are helping solve the imbalance between the enormous number of target compounds and the miniscule number of leads. The imbalance results from the severe shortage of chemists who can pursue drug discovery in academic research.

Pharmacogenomics (PGx)

Pfizer teamed with IBM for a PGx data warehousing solution. The solution combines clinical patient data with genomic patient data, and allows the user to look at the two together as a basis for targeted treatment and for selecting patients for clinical trials. Only 30 percent to 50 percent of patients generally benefit from some drugs because of genetic factors. By knowing which patient has a certain expressed gene, a drug developer will know which patients can safely test a drug for efficacy. Likewise, the physician treating the patient will apply a diagnostic test before prescribing drugs that will be effective for treatment

Clinical Trial Management, Enterprise Content Management, and Life Cycle Document Management Systems

The data explosion - from drug discovery, new kinds of patient data, and regulatory requirements - drives demand for new capabilities. It becomes easier and easier to acquire biological data, so more and more data accumulates. What can you do with it? How can you keep it? How manage it? How make the best use of it?

Three additional ways in which IT is helping biomedical researchers and developers to operate more efficiently and cost-effectively are clinical trial

management, enterprise content management, and life cycle document management systems.

- Clinical management systems, such as that developed for DOV Pharmaceutical by IBM, transform raw data into clinical research intelligence and improve operational efficiency throughout the development process. They improve communications among investigative team members, collect and manage data, enable database queries, and enable compliance by tracking regulatory submissions. Reduced costs from better communication and collaboration among researchers and managers at DOV resulted in anticipated savings of \$3 million for the first year through improved productivity
- Enterprise content management systems organize and facilitate collaborative creation of documents and other content. The National Institutes of Health used enterprise content management from IBM and SAM Solutions to create genome scanning centers in collaboration with the Translational Genomics Research Institute and major universities. The purpose was to make the genomic microarray technology available to 10,000 researchers across the country, resulting in the processing, analysis and dissemination of billions of points of DNA data.
- Life-cycle document management offers life science organizations a compliance-enabled solution that extends beyond traditional document management capabilities. The IBM Solution for Compliance in a Regulated Environment (SCORE) provides a framework for managing the life cycle of multiple content types, including documents, scanned images, and medical images, and assures the delivery of that content within a secure, regulatory-compliant collaborative environment. Through the automation of critical information management, SCORE improves business processes and facilitates regulatory compliance across the entire life science value chain.

A global medical device company, with headquarters in Germany and operations in many countries, has been using SCORE since 2005. Similar to other successful global life science businesses, this company has grown organically as well as through acquisition. As a result, many of the company's operations are organized around a distributed model. The company maintains centralized management and several horizontal or shared-services departments, including information technology and quality management, which span the company. However, for practical operational management, each product-centric division maintains a high degree of process autonomy.

The company currently has 200 active users of SCORE, primarily in the global quality management department. These users are responsible for the majority of authoring, editing, and publishing of the documents managed in SCORE. They, along with users from other functional groups, are located at the company's corporate headquarters. The company's deployment of SCORE has taken advantage of its distributed organization, deploying first at the company's headquarters, then expanding to the divisions. Each division maintains its own SCORE capabilities, exchanging documentation as necessary across the company, and shares template libraries and workflows that are maintained in a global library.

The IBM Cluster Solution

IBM Cluster Solutions Are Highly Flexible and Cost-Effective

The IBM cluster is a multi-server system, composed of interconnected computers and associated networking and storage devices. Components are unified via systems management and networking software to accomplish a specific purpose. With high-speed interconnects, the IBM cluster is particularly effective for parallel life science applications that use MPI and data and problem decomposition.

The IBM cluster solution is a family of powerful clustered systems packaged as racks and blades. It ranks as the most flexible cluster platform available, with the widest choice of processors, interconnects, software, and storage. It affords clients a range of choices from deploying piece parts to deploying completely integrated systems such as the IBM System Cluster 1350.

IBM clusters can also deliver significant reductions in power consumption, cost, and space requirements through the use of innovative technologies in:

- Processors
- Advanced power, packaging and cooling
- Special and industry-standard interconnects that deliver low latency and high bandwidths
- Scalable systems and storage management based on Linux.

The performance, scalability, and flexibility of this innovative design enables the affordable solution of a wide range of life science problems.

IBM cluster systems with several hundred multi-core Intel Xeon nodes and the Linux operating system have been deployed for bioinformatics and drug discovery research at several pharmaceutical, biotech, and medical research organizations. At some of these, the computational chemistry workload

requires a large memory system, which can now be managed through the use of the IBM's Power Architecture™ instruction set.

Clusters such as the IBM System Cluster 1350 can overcome “white box” cluster limitations. The clusters provide a wide-range of pre-engineered flexible hardware options, a balanced cluster software environment with systems and workload management, and optimized implementations of major life science applications.

IBM also provides access to a wide range of HPC systems through the company's Deep Computing Capacity on Demand (DCCoD) centers. Life science applications developers and prospective users can test out and get in-depth experience on these systems in a cost-effective manner. This service can be used to study large problems and generate solutions in interdisciplinary biology as well as imaging problems on very large clusters.

IBM's partnership with Microsoft and the planned support of the Windows Compute Cluster Server 2003 (WCCS) on the IBM cluster systems make it possible to integrate devices and diagnostic instruments with back-end servers more effectively than in the past. The availability of validated, optimized life science applications on IBM cluster systems with WCCS will result in a tighter and very affordable integration of instrumentation with analysis.

IBM provides a range of flexible clustering options, each built from piece parts. HPC savvy clients benefit from the ability these options provide to better customize and optimize their life science workload. Other clients benefit from a robust integrated Cluster 1350 solution, which is described in the following section. This includes comprehensive support for pre-engineered and tailored life science applications.

The Future of Cluster Computing in Life science

Further Expansion of the Role and Value of HPC

The increased role of computing in the pharmaceutical and biotech industries and the complexity of the simulations being performed require a combination of HPC systems, massive storage systems, visualization and advanced instrumentation, applications and middleware - all connected by high-speed networks. Life science computing challenges continue to push the envelope in HPC systems, algorithms, scientific models, systems biology, *in-silico* modeling, data integration and scale. These innovations can be expected to further expand the role and value of HPC for the life science industry and related research.

Cost-Effectiveness Will Increase Cluster Penetration

The economies of scale, scope, and power offered by cost-effective standard processors from Intel, AMD, and IBM will continue to increase the penetration of clusters in life science. Furthermore, multi-core systems with increased numbers of cores per socket will enable users in the life science industry to obtain greater performance for the same total cost, including power and cooling costs.

Integrated Workflows on Clusters Will Become Common

Most life science application providers already support the distributed memory cluster architecture with scalable parallel versions of their applications. This computing model has become the dominant model in the industry. Integrated workflows on clusters can be expected to become common.

Cluster Penetration Will Increase as Differentiated Integrated Solutions are Delivered

Many life science organizations have already realized the value of building their own Linux and Microsoft clusters using commodity hardware, standard interconnects and networking technology, open source software, and in-house or third-party applications. They also increasingly realize that the expense and complexity of assembling, integrating, testing and managing these clusters from disparate, piece-part components often outweigh any benefits gained. Companies such as IBM have built integrated cluster solutions that significantly reduce these costs and complexities for most life science organizations.

Advances in power and cooling technologies, multi-core processors, virtualization, storage and data management technologies, and pre-architected life science applications solutions will further increase the adoption of these integrated cluster solutions.

Tighter Workflow Integration Among Life Science Applications and with Devices and Diagnostics Will Continue to Expand In Silico Analyses

Life science solutions from major ISVs have become better integrated with other applications and with data acquisition environments. This integration enables the effective and rapid deployment of collaborative solutions in the pharmaceutical and biotech industries. High throughput analysis and systems biology will become more prevalent as integrated workflow environments become standard HPC environments in the life science.

Conclusions: The IBM Cluster Solution Provides Unique Value for Life Science

IBM clusters perform well on a wide range of life science problems. Clusters are the mainstream platform for HPC in the life science. Many life science applications also benefit from the flexibility of hardware options, quad-core processors, superior I/O and memory performance, and high-speed interconnects such as InfiniBand.

An integrated solution, the IBM System Cluster 1350 has a wide range of pre-engineered flexible hardware options, a balanced cluster software environment with systems and workload management, and energy-efficient and optimized implementations of major life science applications. This is a robust and scalable platform for industrial strength *in silico* analysis. For the first time, many smaller biotech companies with limited in-house skills will have an affordable computational solution that can give them a competitive edge.

Cabot Partners is a collaborative consultancy and an independent IT analyst firm. We specialize in advising technology companies and their clients on how to build and grow a customer base, how to achieve desired revenue and profitability results, and how to make effective use of emerging technologies. To find out more, please go to www.cabotpartners.com.