Powering Cutting Edge Research in Life Sciences with High Performance Computing

High performance computing (HPC) is the foundation of pioneering research in life sciences. HPC plays a vital role in handling large volumes of data and carrying out complex analysis during discovery research as it provides a platform with immense computational power and storage capabilities. However, not all researchers are able to harness the capabilities of HPC effectively. This can be because computational analysis in different areas of discovery research requires specific tools, each having their own compute, input-output (I/O), and throughput requirements. Furthermore, research data needs to be accessed, processed, analyzed, and visualized. Challenges with respect to data movement and storage, performance bottlenecks, and limited expertise compound the computational analysis challenges that companies face today.

Tata Consultancy Services (TCS) recently surveyed several leading life sciences companies to understand the main barriers to HPC adoption and usage. This paper presents the findings of the survey and also discusses HPC's potential in accelerating the outcomes from research and the steps that can be taken to optimize its usage.

Introduction: Transforming research with high performance computing

Emerging disciplines such as bio-informatics, systems biology, and next generation sequencing are changing the way research is carried out in the life sciences domain. These disciplines drive data-driven research, assist in developing more potent regimes for treating disorders, and offer alternate strategies to the standard blockbuster drug approach.

All of these disciplines intrinsically depend on a very strong technology backbone. HPC and complex data analytics platforms and high throughput storage systems are some of the key technologies driving these disciplines forward. Hardware components, middleware, and application software also play an important role in enabling these disciplines.

Let’s take a look at some examples that illustrate the nature and role of these underlying platforms. Consider the large volumes of data generated during Next Generation Sequencing (NGS) analysis. Deriving insights from this data requires a number of computational and data reorganization steps. This necessitates the use of sophisticated data analysis tools and simulation platforms such as an HPC cluster coupled with high throughput parallel file system based storage. Using these tools and platforms, simulations can be carried out to analyze several ‘what if’ scenarios, and thereby, avoid experimental dead ends.

Similarly, there is a pressing need in the pharmaceutical sector for a larger number of prospect drug molecules to improve the drug discovery pipeline. Highly efficient, fast, and cost-effective predictive research is now possible through the use of in-silico or virtual studies. Robust HPC infrastructure, coupled with the right analysis and processing tools, supports fast access, processing, analysis, and visualization of data.

TCS surveyed leading life sciences companies to assess the adoption of HPC in the pharma and medical devices sectors and understand the challenges in using it effectively. This paper presents the findings and our analysis of the survey results.

How organizations currently use HPC

In terms of adoption, 78 per cent of the respondents stated that scientists in their teams use HPC at least once a week. In fact, as shown in figure 1, 56 percent of scientists use HPC every day. This finding underscores the importance of HPC in carrying out research. These organizations use HPC in areas such as NGS, molecular modeling, molecular docking, and systems biology.
Despite the gradual adoption of HPC, certain focus areas in drug discovery and research continue to present a challenge to organizations.

**Key challenges to effective computing**

**Data movement:** A majority of problems or incidents arise during data movement. This could be due to the fact that most of the analysis in life sciences involves large input data sets and output files. At times, data may also need to be moved from the storage device to cluster nodes for further analysis. Ineffective data movement and management strategies are a key challenge to leveraging HPC.

**Compute capacity:** As shown in Figure 2, survey respondents identified compute capacity, or the lack of it, as the second most important challenge. This necessitates the use of appropriate methods to handle large data sets, both during computation and data transfer to ensure the effective use of compute capacity.

**Why current organizational systems are unable to support HPC**

To improve the usage of HPC, it is essential to understand the underlying details. The performance of HPC systems is dependent on a complex interplay of various factors including, but not limited to, the configuration of both the hardware platform and tuning of application software.
As depicted in Figure 3, companies cited poor performance of current systems used to perform HPC as the biggest challenge, followed by the lack of support to address incidents or problems.

**Barriers to realizing the potential of HPC**

Lack of expertise: As shown in Figure 4, 56 percent of the survey respondents stated that the lack of expertise in running bio apps in the current environment was a significant barrier to making the most of HPC.

Deployment of multiple tools: 44 percent of respondents also identified difficulties in handling multiple tools as a challenge. With many of the tools deployed in different areas performing similar or identical functions, managing the overlap in functionality further added to the challenge of working with disparate tools.

Inefficient analysis: The data and results generated during the drug discovery cycle need to be analyzed by various groups across the organization. Work distribution across multiple users often results in variable workloads that in turn lead to suboptimal performance and resource utilization.

Unavailability and prohibitive costs: 33 percent of respondents identified the unavailability of the HPC environment and high cost of maintenance as hindrances to its effective usage and adoption.

![Figure 4: Barriers to using HPC](image.png)

**Recommendations for unlocking HPC’s true potential**

HPC adoption is still in its nascent stage in several life sciences organizations. Despite early challenges and issues including accessibility to HPC infrastructure, availability of HPC enabled tools, data movement across location and data management; it is evident that organizations can realize their research and innovation vision by harnessing the power of HPC. We present key recommendations and strategies that life sciences companies can use to overcome their current HPC adoption challenges:

**Performance enhancement:** Performance can be improved at various levels of increasing complexity through strategies such as black box optimization (by analyzing dependencies on numerical libraries and the environment), system level optimization (through optimal compiler and build settings), code level optimization (with code parallelization, efficient data structures, and appropriate programming paradigms) and algorithm level optimization (using alternative numerical methods and approximations).

The suboptimal use of resources such as computational capacity, memory, or I/O is the main cause of application failure. Life sciences companies can partner with domain and technology experts who can provide end-to-end consultancy and support to run HPC related tools efficiently in a cluster environment.

**Efficient handling of large data sets:** Methods such as data partitioning during computation, and technologies for data transfer acceleration and remote visualization will help drive efficiencies in high data volume research.

**Internal training:** Overcoming the challenge of limited technology expertise requires companies to identify the gaps in their teams’ abilities. By combining internal training with staff augmentation strategies, they can ensure access to the right skills and technical knowledge.
Simplification of infrastructure: Companies can partner with a HPC management services provider and leverage their partner’s expertise to parallelize, optimize, and scale serial code. This can help create a cost efficient environment and reduce the overheads of managing complex infrastructure. There is a diverse resource requirement for various classes of Life Sciences tools and applications. Hybrid HPC platforms help enable optimal performance for complex simulation as well as data analytics. These platforms help address issues of increasing data volumes and data movement.

Conclusion: Drive research success with HPC

Trends indicate greater use of new technologies to virtualise research and use of bio-simulations and models for better understanding of biological processes. Statistics indicate that NGS analysis will be on the rise as the cost for sequencing drops. This is likely to lead to an exponential rise in corresponding data volumes. Moreover, the analysis of biological data generated by various disciplines poses several challenges - given its diverse nature as well as the complexity of data movement within and across collaborating institutes.

A powerful HPC platform can be effectively leveraged to process, analyze, manage, and store this diverse data to enable pioneering research. With the right support, both in terms of expertise and methodologies, HPC has the potential to dramatically improve time-to-insight and research success across the life sciences value chain.

About the Authors

Sumit Misra
Lead – North America, Global Consulting Practice, Research
Sumit has been with Tata Consultancy Services (TCS) for over seven years. As the GCP Research Lead for the North America geography, he handles research assignments and also authors market intelligence reports for the region. He has an MBA from the Xavier Institute of Management, Bhubaneswar.

Arundhati Saraph
Lead - HPC, Life Sciences
At TCS, Arundhati works on leveraging HPC to address scientific problems in niche areas of life sciences. These include discovery research, genomics, metagenomics, predictive toxicology, and bioinformatics - that find application in the pharmaceutical and healthcare sectors. Arundhati has a Ph.D. from the National Chemical Laboratory, Pune, and has previously worked with several national institutes in India.

Sanket Sinha
Head, Delivery and Pre-Sales – HPC
Sanket heads the solution and pre-sales function of the HPC practice at TCS. As a member of the Tata Administrative Services (TAS), Sanket has been associated with the Tata Group for close to a decade. He has a B.Tech degree in mechanical engineering from IIT, Kharagpur and a management degree from IIM, Ahmedabad.

Srihari Narasimhan
Technology Consultant, HPC CoE
Srihari is a technologist with the HPC practice at TCS. His work experience includes leading the HPC activities for medical imaging, healthcare services, financial risk analytics, energy resource optimization, and materials modeling business units at the GE Research Labs. Previously, Srihari worked as a full-time research and teaching faculty at the University of Stuttgart, Germany, from where he completed a Ph.D. in Computer Science.
About TCS’ Life Sciences Business Unit

With over two decades of experience in the life sciences domain, TCS offers a comprehensive portfolio in IT, Consulting, KPO, Infrastructure and Engineering services as well as new-age business solutions including mobility and big data catering to companies in the pharma, biotech, medical devices, and diagnostics industries. Our offerings help clients accelerate drug discovery, advance clinical trial efficiencies, maximize manufacturing productivity, and improve sales and marketing effectiveness.

We draw on our experience of having worked with 7 of the top 10 global pharmaceutical companies and 8 of the top 10 medical device manufacturers. Our commitment towards developing next-generation innovative solutions and facilitating cutting-edge research - through our Life Sciences Innovation Lab, research collaborations, multiple centers of excellence and Co-Innovation Network (COIN™) - have made us a preferred partner for the world’s leading life sciences companies.

Contact

For more information about TCS’ Life Sciences Business Unit, visit:
http://www.tcs.com/industries/life-sciences/Pages/default.aspx
Email: lshcip.pmo@tcs.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India’s largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

IT Services
Business Solutions
Consulting