Cost-Benefit Analysis: Comparing the Cray® Urika®-GX System with Public Cloud Implementations for Life Sciences

Sponsored by Cray

Ajay Asthana Ph.D., Srini Chari, Ph. D., MBA and Rama Gullapalli, MD, PhD info@cabotpartners.com

May, 2016

Cabot Partners Group, Inc. 100 Woodcrest Lane, Danbury CT 06810, www.cabotpartners.com

Executive Summary

The economic benefit of life sciences research is huge across multiple sectors including Pharmaceutical R&D, Healthcare and Agriculture. With increasingly affordable gene sequencing and imaging technologies, it is now much faster and cheaper to generate raw data. But analyzing and integrating this growing volume of life sciences data to glean valuable insights is challenging and is holding back innovation.

Life sciences IT departments need performance-optimized systems for life sciences computations to overcome the rising complexity and costs associated with developing and managing applications on commodity on-premise infrastructures across numerous silos.

The Cray® Urika®-GX system has several novel features to reduce complexity, bring high-performance and scalability to run multiple complex analytics for better outcomes. The integration of processors, networks, software and storage leads to shorter application development cycles and faster time to value. It requires minimal ongoing administration or tuning, allowing customers to optimize their Total Cost of Ownership (TCO) for life sciences.

Many researchers are also using cloud computing as an alternative. This is driven by the promise of better flexibility, greater collaboration and scale, lower usage costs (for compute/storage) and almost no capital, facilities and systems administration costs.

However, there are many challenges with running all life sciences workloads entirely on public clouds. These include hidden costs: moving, managing and securing data throughout its lifecycle; and additional costs including compliance and productivity of scientists.

The comprehensive TCO analysis presented in this paper compares the Cray Urika-GX system with a public cloud alternative from Amazon Web Services (AWS) for three configurations – small, medium and large. Very favorable assumptions are used for AWS. This cost-benefit analysis framework considers tangible as well as hidden costs (data transfer, compliance, security and productivity loss of scientists).

*Compared to a public cloud such as AWS, life sciences clients **can lower the three-year TCO for all configurations** with the **Cray Urika-GX** System**.** A breakeven analysis demonstrates that the breakeven point for small and medium configurations is between 2-3 years. For large configurations, this breakeven point is between 1-2 years. For workloads with greater data transfer to and from the cloud, this breakeven point occurs even earlier.*

Clients who may be concerned solely with short-duration analytics and are willing to discard this data may choose a public cloud solution. For the vast majority of clients, hybrid cloud approaches that combine or augment an on-premise Cray Urika-GX System with the cloud have the potential to offer a better solution for the broad spectrum of life sciences workloads.

Cabot Partners
Optimizing Business Value

Economic Impact of Life Sciences is Huge Across Many Sectors

Human genome can be sequenced in a few hours for about $1000

The rate of progress in genomics and high resolution imaging is astounding. Rapidly declining gene sequencing costs, advances in recording technology and affordable High Performance Computing (HPC) solutions to process ever larger datasets is transforming life sciences. Today, a human genome can be sequenced in a few hours and for about $1000, a task that took 13 years and $2.7 billion to accomplish during the Human Genome Project.[1] Similarly, analyses that relate neuronal responses to sensory input and behavior run in minutes on clusters, turning brain activity mapping efforts into biological insights.[2]

By 2025, Economic impact of Next-Generation Sequencing (NGS) and HPC between $700 billion to $1.6 trillion a year

By 2025, the economic impact of next-generation sequencing (NGS) and related HPC technologies (Figure 1) could be between $700 billion to $1.6 trillion a year.[3] Bulk of this value is estimated to result from delivering better healthcare by prolonging and improving lives. NGS enables earlier disease detection, better diagnoses, discovery of new drugs and more personalized therapies.

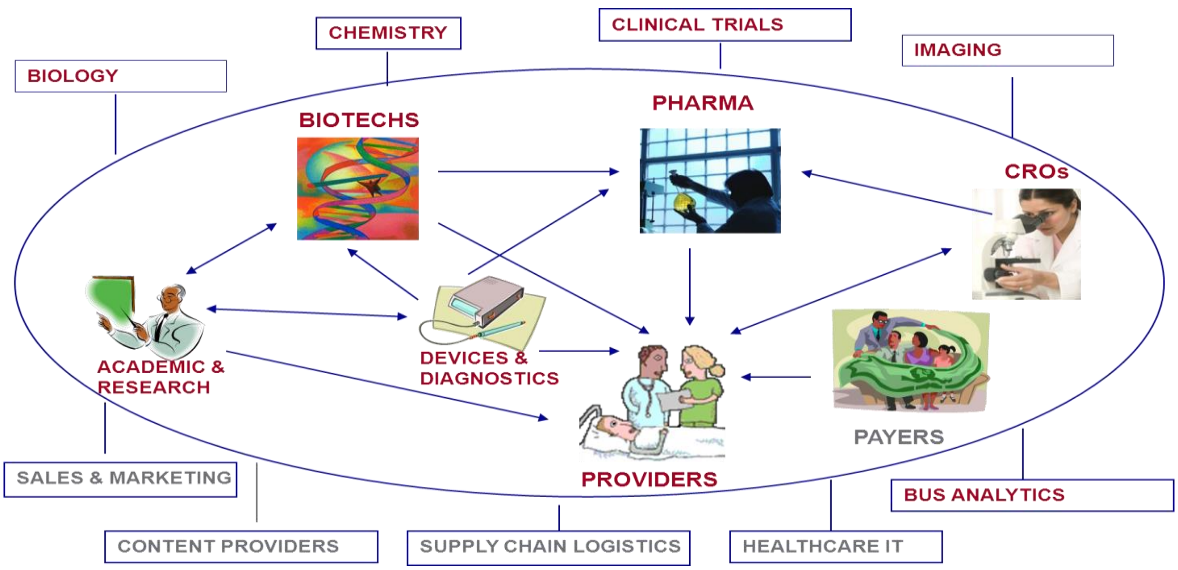

Figure 1: Healthcare/Life Sciences Disciplines/Industries (Red) Benefit from HPC

Life Sciences Analytics can help bring billion dollar drugs to market faster

Computational chemistry, bioinformatics and statistical analyses are used to accelerate drug discovery and development; providing bio-pharmaceutical firms with a first-mover advantage to bring drugs more efficiently to market. The economic impact and stakes are huge. It typically costs $2.8 billion[4] to bring a new drug to market, and blockbuster drugs can bring in billions of dollars of new revenue annually.

In addition, agricultural genomics can better help feed the world's growing population by raising sustainable productivity in places with food shortages while conserving water. Using NGS, plant and animal breeders and researchers can identify desirable traits, leading to healthier and more productive crops and livestock. Longer-term, NGS could help genetically engineer less expensive biofuels that consume less energy compared to plant-based biofuels.

[1] http://www.veritasgenetics.com/documents/VG-PGP-Announcement-Final.pdf
[2] Jeremy Freeman, et. al., "Mapping brain activity at scale with cluster computing", Nature Methods, July 2014.
[3] McKinsey Global Institute, "Disruptive technologies: Advances that will transform life, business, and the global economy", May 2013.
[4] http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study

Clinical / diagnostic analysis needs faster, more reliable and secure infrastructure

Getting Insights from the Growing Data Volumes is Challenging

Data volumes and access patterns intensify and vary widely as life sciences data becomes more prevalent in time-critical clinical/diagnostic analyses. What took days to analyze in a pure research context must now be done reliably in hours, even as larger number of projects and files must be tracked. To generate actionable insights, the HPC infrastructure must:

- Store and rapidly analyze the growing volumes of complex life science data and be in-synch with the output speed of sequencers and/or imaging instruments
- Enable collaboration and data/workload sharing across organizations and geographic boundaries in a secure environment that ensures data privacy, and
- Provide a common analytics platform (Figure 2) to seamlessly integrate complex and diverse life science and healthcare data/applications throughout the pipeline: Data Acquisition/Preparation, Base Analytics, Contextualization and Advanced Analytics
- Minimize the Total Cost of Ownership (TCO) for the entire life sciences workflow.

A common analytics platform needed for entire Life Sciences Analytics Pipeline

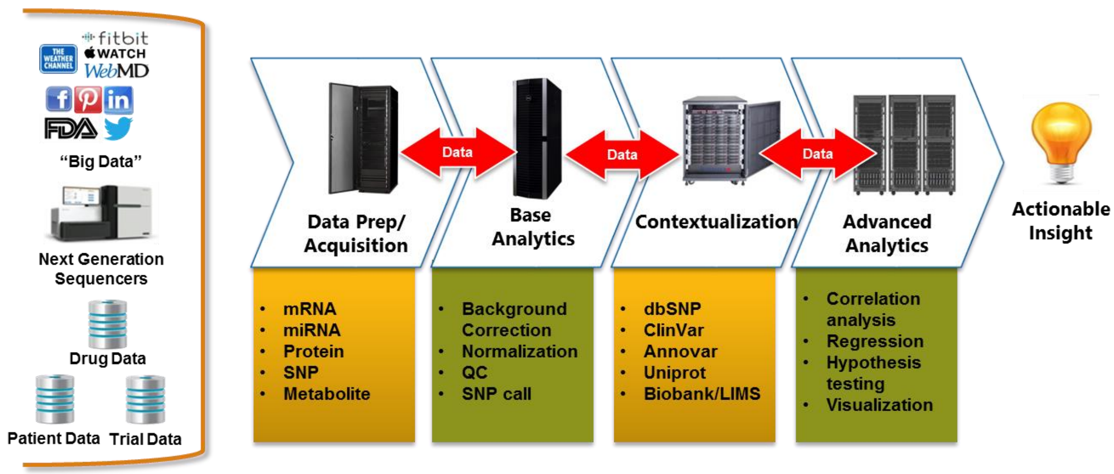

Figure 2: A Typical Life Sciences Analytics Pipeline

Considerations to Minimize the TCO for Life Sciences Workflows

The information technology (IT) organization must collaborate closely with end users to agree on investment decisions, data management policies and infrastructure deployment choices. This helps optimize the total **Costs** and **Benefits** for the entire workflow.

Direct and hidden costs must be considered across data lifecycle

Costs. A range of key direct and hidden costs should be carefully considered:

- *IT*: Capital for servers, storage, networks, software purchase, etc. or corresponding cloud service provider charges. Today, many IT organizations' budgets are narrowly focused on these direct costs; ignoring hidden costs that may hamper innovation and productivity.
- *Operational:* Labor (salaries for end-users and IT staff), energy, IT hardware and software maintenance, software license, etc.
- *Other:* deployment and training, downtime, bandwidth, productivity loss, lack of adequate compliance, security and data protection, etc.

IT investments yield tangible and intangible benefits

Benefits. The benefits from an IT investment for life sciences come in several categories:

- *Strategic:* better patient outcomes, improved reputation for the institution, better stakeholder partnerships, leadership IP portfolio, ability to attract/retain top talent, etc.
- *Research:* more innovation, better collaboration, greater insights, improved quality, etc.
- *Operational:* faster time to results, greater throughput and more users supported, improved user productivity, better capacity planning, etc.
- *Infrastructure:* improved system management, administration, and provisioning, enhanced security, higher utilization, scalability, reduced downtime, access to robust proven technology and infrastructure management expertise, etc.

It is critical to evaluate the costs and benefits of various deployment alternatives.

Infrastructure Deployment Options: Cloud or On-Premise?

Easy access, flexibility and collaboration key public cloud benefits

To minimize TCO, many life sciences organizations have deployed on-premise, performance-optimized systems such as the Cray Urika-GX. Increasingly, driven by easy access and lower infrastructure usage costs per user and the need to collaborate, there is growing interest in using public clouds such as the Amazon Web Services (AWS) on commodity clusters for life sciences. Key costs/benefits with each alternative include:

Public Cloud Benefits. Key advantages for life sciences clients and workloads include:

- Unified, location-independent platform for data and computation
- Pay-per-use and available even to small labs
- Affordable infrastructure costs for individual end-users
- Linear scaling with parallel execution for some workloads
- Ability to quickly scale to large number of compute resources when needed
- Basic infrastructure deployment and management complexities handled by cloud provider
- Up-to-date base technology platform for processing and storage
- Better collaboration between scientists with a centralized computing environment.

Security, data transfer, usability and compliance are key challenges

Public Cloud Costs/Limitations: Challenges running life sciences workloads entirely on public clouds typically magnify with the number of cloud users and problem size include:

- *Data Security and Protection:* How secure and private is the Data? Is Data stored in a redundant manner to ensure recoverability? How much control does the user have over remote storage?
- *Data Transfer Speed:* How to move large data into the cloud efficiently? Especially frequently changing data that must be ingested and analyzed from a range of sources? With slow upload speed because of limited Wide Area Network (WAN) bandwidth, many cloud providers suggest that users ship data on hard drives through overnight mail, making this manual, cumbersome and expensive.
- *Scalability:* Many life sciences applications are hard to parallelize on commodity clusters. Some workloads may run for extended periods of time; sometimes creating 'runaway" jobs that may burn up, waste and even exhaust cloud resources. Many

Runaway and assured deletion costs are additional risks

On-premise still dominant deployment architecture delivering predictable high-performance

Cray Urika-GX System is pre-integrated and optimized for Life Sciences Analytics

analytics algorithms need fast data movement within the compute cloud from hierarchical storage systems. These algorithms require highly custom and data-optimized configurations that may not be supported.

- *Usability:* Bioinformatics involves a pipeline of scripts and a command line interface. Setting up nodes in the cloud requires users to possess a deeper understanding of the intricacies of the system to manually setup these command lines and scripts. Setting up and sourcing specific applications needed for the entire life sciences workflow is cumbersome and sometimes not feasible. All this is expensive and time-consuming and may have to be repeated for each user type in the life sciences organization.

  Life sciences R&D and innovation require close collaborations between scientists to visualize and interpret analysis results. When working with large data sets on a cloud, effective remote visualization is required to mitigate large data movement challenges from the cloud to remote users. Many public clouds do not support these capabilities effectively thus hampering user productivity and innovation.
- *Regulatory Compliance:* Most life sciences IT solutions must be validated for compliance with FDA and other agency regulations. On the cloud, the applications and the installed systems environment must be validated. With localized IT infrastructure, this is assured by the internal IT group, but in the cloud, this must also be done by the cloud provider. This typically adds additional costs and complexity especially during some compliance audits.

On-Premise. The entire storage and compute infrastructure is all on-premise. This is a well understood deployment architecture. Existing institutions often optimize their workflows and typically get better predictable performance. It is also easy to recover data when the data owner is no longer at the institution. Unlike on many cloud providers, costs associated with data transfer, assured deletion of data and runaway jobs are not a concern. But costs for facilities, labor, energy should be considered.

Hybrid clouds, that combine or augment an on-premise HPC system such as the Cray Urika-GX system with a public cloud, have the potential to offer a better solution for a broad spectrum of life sciences workloads. The Cray Urika-GX System is proven for life sciences.

The Cray Urika-GX System is Optimized for Life Sciences Analytics

The Cray Urika-GX System is designed to run complex analytics on Terabytes (TB) and Petabytes (TB) of data, much faster than commodity clustered systems. It supports several memory-storage hierarchies and includes a fast, low-latency interconnect (Cray Aries). A pre-integrated, comprehensive analytics software stack delivers actionable insights faster.

This software stack includes an optimized set of tools for capturing and organizing a wide variety of data types from different sources, and executing analytic jobs. Key stack components are: Cray Graph Engine, Cluster/Workload Management (YARN, SLURM and Apache Mesos), Hadoop, Spark, File Systems (Lustre, HDFS) and CentOS. This system also requires minimal ongoing administration or tuning which lowers the TCO for customers.

Here are some real-life examples:

De Novo assembly computational and memory-intensive

Cray Urika-GX enables faster and higher-quality assembly

Drug repurposing is compute intensive; slowing quick decision making

De Novo Assembly

Description/ Challenges	• Method to determine the nucleotide sequence of a contiguous strand of DNA without using of reference genomes • Computationally challenging and memory-intensive • Critical to explore novel genomes and highly varied portions of the human genome • Many agricultural research projects use De Novo since a good reference is often absent • Downstream genomic interpretation requires large scale, big data integration with a wide variety of structured and unstructured sources.
Solution/Results	• Urika-GX provides an ideal big data platform for data preparation and downstream interpretation / integration • Urika-GX's large memory and low-latency interconnects support the extreme parallelism (up to 15000 cores) required for high speed assembly • Spark optimized platform for downstream genomic interpretation • CGE provides a highly differentiated analytics capability • Human genome was assembled in under 9 minutes and the wheat genome was assembled in under 40 minutes.
Benefits	• Higher throughput provides scientists a practical method to leverage De Novo assembly more broadly and translates directly to lower cost for many organizations • Allows researchers to scale to higher coverage depths, leading to a higher quality assembly • Being able prepare the data, perform the assembly and the advanced analytics required to interpret the results all on the same platform simplifies the compute infrastructure and eliminates expensive and time consuming data movement.

Drug Repurposing

Description/ Challenges	• Researchers want a way to quickly get to "yes" or "no", in order to prioritize drug repurposing opportunities • A wealth of data both proprietary and public and from many sources, both structured and unstructured • Time consuming process as all needed data has to be remodeled to fit each hypothesis • Researchers can't see the relationships between the data which would help identify promising candidates for new therapies.
Solution/Results	• Scalable solution allows the data set to expand over time • Handles all types of Big Data workloads by combining highest performance Graph Engine with optimized Hadoop/Spark • Graph representation of connections and associations between

Urika-GX's large memory and scale speeds decision making

	drugs and targets at scale enable thousands of hypotheses to be validated or rejected in parallel • Open solution, so customers can deploy any analytic tools now or in the future • Urika-GX decreases analysis time by eliminating the need for a new data model to test each hypothesis • Data assembled in a single graph in a vast shared memory so unknown relationships between data can be discovered • Rapid integration of multiple source data due to performance efficiencies of Aries interconnect and RDF/SPARQL interface.[5]
Benefits	• Significant increase in the number of identified drug opportunities that have a higher probability of success • Validated a thousand hypotheses in the time it previously took to validate one • Quickly eliminate drug candidates that are unable to deliver desired results.

More drug opportunities with higher probability of success

To illustrate the TCO advantages of the Cray Urika-GX System over AWS, a NGS workflow is analyzed. Direct and Hidden costs are considered for the entire data lifecycle.

Building the TCO Model: Cray Urika-GX System and Public Cloud

The comprehensive Cost-Benefit Analysis presented here compares the *Total Cost (Direct + Hidden)* of the Cray Urika-GX System with the AWS public cloud for three configurations – small (16 Urika-GX nodes – 64 TB of data), medium (32 Urika-GX nodes – 128 TB of data) and large (48 Urika-GX nodes – 192 TB of data). The Urika-GX systems were sized to yield over 80% utilization levels consistently for all scenarios.

The following assumptions/data used in the TCO analysis were obtained from recently published articles and validated through a rigorous process of interviewing subject matter experts. In all cases, these assumptions favored AWS over the Cray Urika-GX System.

Comprehensive TCO model includes Direct and Hidden costs for several scenarios

Direct Costs: Drivers, Sources and Assumptions

- *Compute (Cray Urika-GX):* Cray provided the average price for the System (servers + network) for each configuration; of which 80% was assumed to be the server cost.
- *Compute (AWS):* Averaged from the AWS website for 25% more servers (a conservative assumption since Urika-GX is typically faster) for each configuration.
- *Storage (Cray Urika-GX):* Storage for the Cray Urika-GX must be purchased separately. So, the current cost of storage systems corresponding to each configuration was added.
- *Storage (AWS):* Obtained from the AWS website and scaled for each configuration.
- *Network (Cray Urika-GX):* Assumed to be remaining 20% of Urika-GX System price.

[5] http://www.w3.org/TR/rdf-sparql-query/

Direct costs: Compute, Storage, Network/Data Transfer, Facilities and Sys. Admin.

- *Network (Data Transfer for AWS):* Data must be moved to and from the cloud. AWS charges explicitly for transfer from the cloud. In this line item, only these costs are considered. While AWS doesn't charge explicitly for data transfer into the cloud, there are several costs associated with data transfer delay (to and from the cloud) that impact productivity of the scientists and the organization.
- *Facilities includes Power/Cooling (only for Cray Urika-GX):* Facilities costs[6] include cost of power, cooling and floor-space. These costs increase as configurations get larger. Electricity price was assumed to be $0.09 KW/hour.
- *System Administration (only for Cray Urika-GX):* Full Time Equivalent (FTE) people costs for systems operations increases as configurations get larger. The estimated number of administration resources were based[7] on many recent analytics deployments. The cost of one FTE (Administrator) was assumed to be $100/hr.

Hidden Costs: Drivers, Sources and Assumptions

Hidden costs: Compliance, Runaway, Security and Productivity Loss

- *Compliance (Urika-GX and AWS):* The cost of achieving regulatory security compliance is on average $3.5 million each year.[8] Only 2% of this was applied for the medium Cray Urika-GX configuration, and prorated for the small and large Cray configurations. For AWS, these costs were proportionately increased to account for less infrastructure visibility, and greater complexity and more advanced skills needed for cloud-based audits based on the configuration sizes.
- *Runaway (AWS only):* Costs associated with the small percentage of runaway[9] cloud jobs (invisible to the user) that continue to consume IT resources. Also includes assured deletion costs on remote storage when user IDs are deleted. Assumed to be 15% of storage and compute costs for the medium configuration, and prorated for others.
- *Security (Urika-GX and AWS):* The bill[10] for security for a 50-person organization is $57,600 annually for on-premise IT. This cost is ascribed to the medium Cray Urika-GX configuration, and prorated for the other configurations. For AWS, costs were proportionally increased to account for less infrastructure visibility and greater vulnerability of public cloud.
- *Productivity Loss (Both):* Lost productivity of scientists related to delays caused by slow execution speed and data transfers – grows with larger configurations that typically support more scientists. The cost of one FTE (Scientist) was assumed to be $146.91/hr.[11] These productivity loss costs are lower for the Cray Urika-GX System because of no significant data transfer delays.

On-premise includes Maintenance costs

In addition, for the on-premise scenario, the acquisition costs for the Cray Urika-GX System and associated storage are applied in Year 1 with an additional 20% annual maintenance cost for the remaining years.

The comprehensive TCO model is run for various scenarios for each configuration to objectively assess impacts of Data Transfer and to estimate the Breakeven Points.

[6] Cray Urika-GX System – Technical Specifications, 2016.
[7] ITG Paper "Cost/Benefit case for IBM Puredata system for Analytics" Comparing costs and time to value with Teradata Data Warehouse Appliance, May 2014.
[8] Poneman Institute, "The True Cost of Compliance", http://www.tripwire.com/tripwire/assets/File/ponemon/True_Cost_of_Compliance_Report.pdf
[9] http://blog.iland.com/cloud/blog/runaway-cloud-went-hill-blewthe-budget-window/
[10] http://www.bloomberg.com/bw/articles/2014-10-31/cybersecurity-how-much-should-it-cost-your-small-business
[11] O'Reilly, "2015 Data Scientists Salary Survey", 2015.

Lower TCO for the Cray Urika-GX compared with Public Cloud

16.7% lower TCO over 3 years with the large Cray Urika-GX System

The Total Costs (in today's dollars with a 9% discount rate) for a three and five-year horizon are presented in Figures 3 and 4 for a base case of 25% Data Transfer from the Cloud.

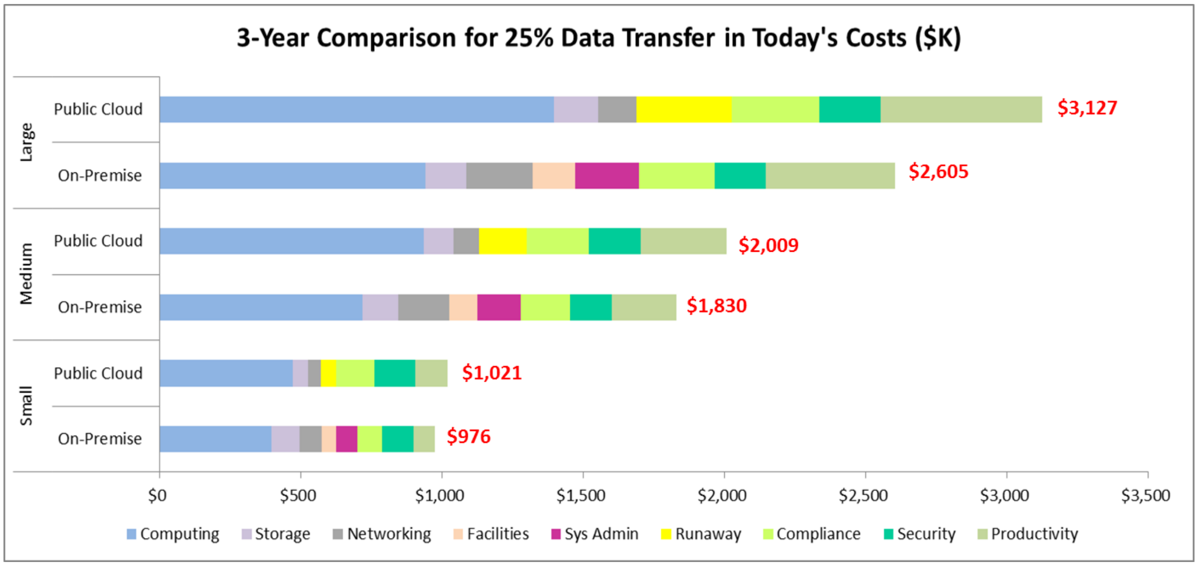

Figure 3: Three-Year TCO Results Comparing Cray Urika-GX with AWS

25.5% lower TCO over 5 years with the large Cray Urika-GX System

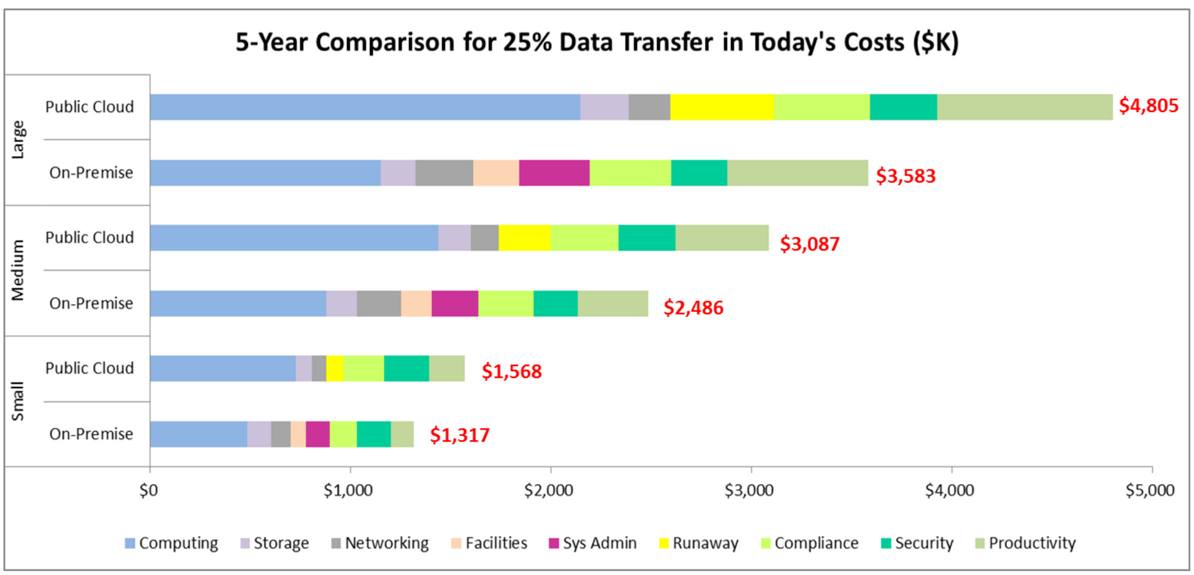

Figure 4: Five-Year TCO Results Comparing Cray Urika-GX with AWS

Cray Urika-GX lowers TCO for all scenarios considered – much lower for challenging large cases

The TCO advantages of the Cray Urika-GX over AWS are clear in all cases – even under very favorable assumptions for AWS. These cost savings increase as the configuration and duration get larger – from 4.4% for a small configuration over a three-year duration, to 25.4% for a five-year duration. Most TCO analyses only consider Direct Costs. These are slightly lower for AWS for the small case, and cross over for the medium case with a five-year duration. However, for the large case (more realistic with the explosion of data volumes), the Direct Costs for the Cray Urika-GX are lower than AWS for both durations.

Breakeven Analysis reinforces the business case for Cray Urika-GX for big data clinical/ commercial settings

Figure 5 illustrates a Breakeven Analysis by examining the Cumulative Costs for every year for the three configurations with 25% Data Transfer. As depicted by the three ellipses, the Breakeven Points for the small and medium configurations fall between two and three years. For the large case, the Breakeven Point occurs earlier – between one and two years. This reinforces the business case for the Cray Urika-GX System over AWS, in many clinical/commercial settings, as data volumes grow and need to be stored for longer duration. For short-duration pure Research studies (data is often discarded), AWS could suffice.

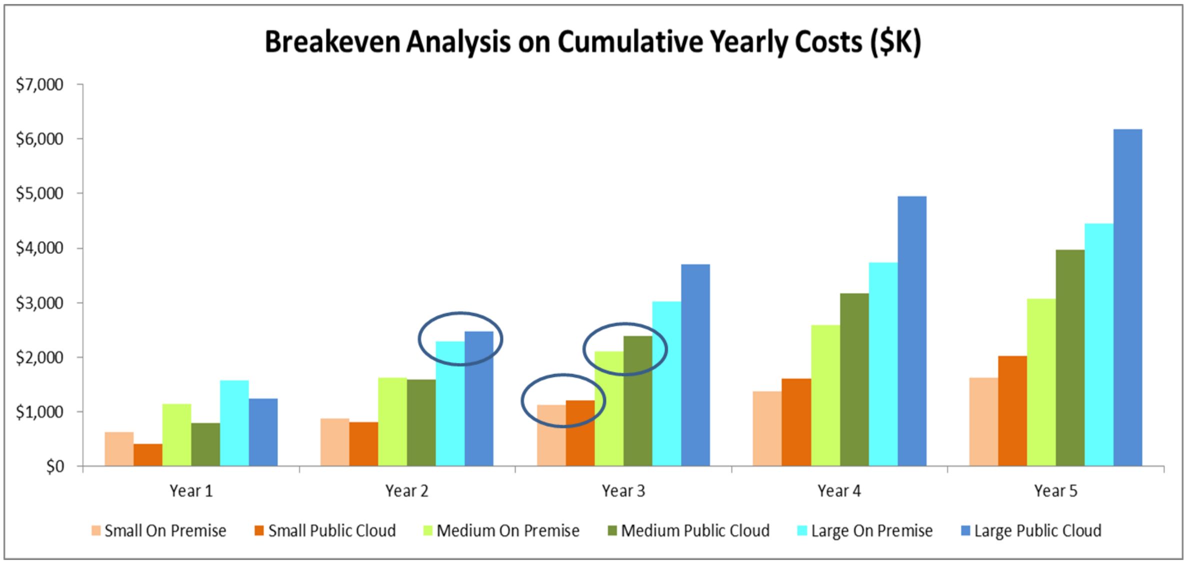

Figure 5: Breakeven Analysis for Cumulative Costs with 25% Data Transfer

Cloud may suffice for short duration pure Research

Next, Data Transfer from the cloud is varied from 0% to 100% to study the impact of moving large data sets over WANs for visualization or integration with other on-premise applications. The Total Cost increases as Data Transfer from the cloud grows (Figure 6); emphasizing the need to consider these hidden costs especially as data volumes continue to explode throughout the Life Sciences Analytics pipeline.

Greater the data transfer from the cloud; larger the Total Costs

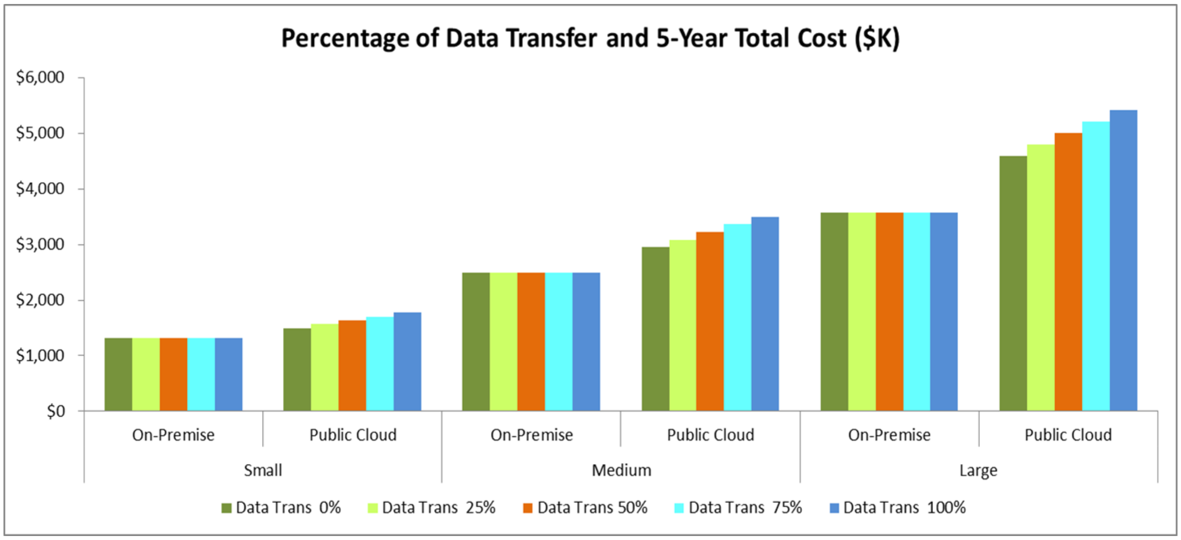

Figure 6: Total Cost for Cloud Grows with Greater Data Transfer

TCO for Cray Urika-GX could be lower if other drivers such as time to value are considered

The Total Costs associated with the Cray Urika-GX System included in this analysis are significantly lower because of the unique appliance-style packaging with bundled high-value Analytics software that improves performance and scientists' productivity. Other key advantages of the Cray Urika-GX not considered here include lower downtime and faster time to value. These advantages further improve the productivity of the organization.

Conclusions and Recommendations

High Performance Analytics in Life Sciences has a huge economic impact across many sectors from healthcare to pharmaceutical to agriculture. Life Sciences data volumes are exploding as higher-resolution instrument costs become more affordable. But analyzing all this data quickly, reliably, and cost-effectively to produce insights is challenging – especially as life sciences applications are more integrated in clinical and drug-discovery processes.

The Cray Urika-GX System provides a common, cost-effective, high-performance integrated Analytics platform. It is optimized for Life Sciences and proven to address the industry's most compute and data-intensive problems from De Novo assembly to Drug Repurposing.

Cray Urika-GX System provides a cost-effective and integrated Analytics platform with consistently lower TCO

Increasingly, public clouds such as Amazon Web Services (AWS) are also being used as an alternative platform for Life Sciences Analytics. Key advantages with public clouds include better flexibility, greater collaboration and scale, lower usage costs (for Compute/Storage) and almost no Facilities and Systems Administration costs. But there are several Hidden Costs – Data Transfer, Security, Compliance and Productivity Loss of scientists – that must be included in any objective TCO analysis.

The comprehensive TCO analysis for a typical Life Sciences Analytics workflow presented here, includes Hidden Costs. This analysis compares the Cray Urika-GX System with AWS for three configurations – small, medium and large. In all cases, the Cray Urika-GX System delivered lower TCO for three and five years.

A breakeven analysis demonstrates that the Breakeven Point for small and medium configurations is between 2-3 years. For large configurations, this Breakeven Point is between 1-2 years. For workloads with greater data transfer from the cloud, this Breakeven Point occurs even earlier. This amplifies the business case for the Cray Urika-GX in many clinical/commercial settings where growing volumes of data must be stored longer-term.

Clients who may be concerned solely with short-duration analytics and are willing to discard data, may choose a public cloud solution. But for the vast majority of clients, hybrid cloud approaches that combine or augment an on-premise Cray Urika-GX System with a public cloud have the potential to offer a better cloud solution for the broad spectrum of life sciences workloads.