IBM eX5 Improves Next Generation Sequencing for Sanger Institute

Sponsored by IBM
Srini Chari, Ph.D., MBA
January 2011
mailto:chari@cabotpartners.com

Executive Summary

Life sciences and High Performance Computing(HPC) have a symbiotic relationship -- life sciences research today relies heavily on HPC infrastructure even as HPC itself evolves rapidly to meet newer, more demanding data analysis requirements in areas such as the Next Generation Sequencing (NGS). IBM and Intel collaborate worldwide with HPC application providers and users, helping them optimize their applications on the IBM eX5 Enterprise Server portfolio to solve challenging problems in industry, academia and research. One such example is the Sanger Institute that deploys the IBMs System x3950 M2 for NGS research. IBMs new eX5 Enterprise Systems, featuring the Intel® Xeon® processor 7500 series, perform much better for NGS applications such as Velvet and deftly address its extreme memory requirements. The new eX5 series offerings - System x3850 X5 and System 3950 X5, not only provide increased HPC density but also enables users to add a full drawer of extra memory, a feature available exclusively from IBM in an x86 processor based HPC server. Many large memory intensive HPC applications can now run with just a handful of faster, multi-core, scalable eX5 series IBM systems, resulting in minimized costs, energy requirements and great simplicity of deployment.

This paper presents an overview of NGS followed by key technological and computational issues related to life sciences research, with specific reference to Sanger Institute, that are addressed through IBM's Intel based eX5 systems. Based on real scientific experiment data sourced from IBM and Sanger Institute, we highlight how IBM eX5 servers scale up and are best suited for compute and memory intensive applications such as Velvet, ABySS and SOAP. Life sciences applications benefit not only from the flexibility, dense compute capacity, 512GB-1TB memory exclusively engineered in eX5 enterprise systems for Velvet kind of applications, but also from superior I/O, memory performance and high speed interconnects coupled with IBM's acumen for simplified cluster administration, engineering innovation for unparalleled power and energy efficiencies.

Introduction to Next Generation Sequencing

The *Human Genome Project*[1] was certainly one of the most significant milestones in genomics research and gene sequencing. Building on that and driven by the need for fast, affordable and accurate genome information, the Sanger Institute is today at the forefront of an explosion of activity in next generation sequence through research such as '*Broken genomes behind breast cancer*'[2] and '*Novel method to reveal drug targets: Interactions between proteins studied on a global scale*'[3]. Advances in computer hardware and software and new computing paradigms such as cloud computing help scientists at Sanger and across the world to address the challenges of scale and efficiency required to compute, store, sift, visualize and analyze gigantic volumes of genomic research data.

Gene Sequencing: Genome DNA sequences record all the genetic information contained in a given organism. But creating a sequential list of the base pairs comprising the DNA of a particular plant and animal is extremely difficult – there is no mechanism to read a single strand of DNA like a punch tape. Instead, scientists use a crude technique that first breaks[4] *(Figure 1)* DNA into short pieces that they know how to identify. They then reassemble the original sequence based on how these short fragments overlap. This is a complex and error-prone process – much like shredding several copies of a critical document and then having to reconstruct the original by matching bits of shreds using overlapping text and other patterns[5]. Reliably sequencing a single stretch of DNA involves combining many dozens of duplicate data sets to arrive at an acceptable level of fidelity.

[1] Human Genome Project (HGP)
2 Sanger link - Broken Genomes behind breast cancer
3 Sanger link - Novel method to reveal drug targets: Interactions between proteins studied on a global scale
4 Thomas Keane – Next Gen Sequencing
5 A short primer on Bioinformatics

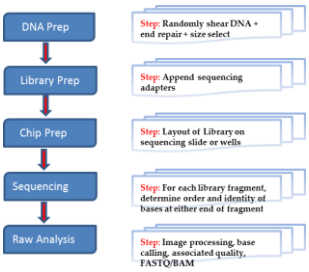

Figure 1: Next Generation Sequencing Process (Sanger –NGS[4])

Locating the genes: Once the genome DNA sequence is established, scientists locate and delimit the coding regions from the non-coding or unused ones. This is a part of determining the function of each coding region within the DNA. The search for coding regions utilizes systems and computer algorithms from areas such as signal processing, cryptography, and natural language processing techniques that are good at distinguishing information from random noise. The work is tedious and requires scientists to sift through sequencing data and apply various alignment algorithms to look for overlaps between short DNA fragments. In theory, this is similar to text analysis techniques for electronic documents that have been used for years now. In practice, because of the immense volumes of data involved in next generation sequencing (NGS), the work is complex, resource and labor intensive, and error prone.

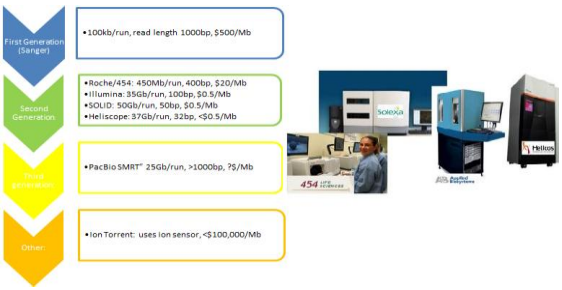

Figure 2: Next Generation Sequencing Technologies ***(Source: Sanger, NGS ecosystem[7])***

This is why research institutes like Sanger Institute use the IBM eX5 for running NGS applications such as Velvet, ABySS and SOAP on projects such as 'Human genome sequencing of an African male individual' and 'Gene sequencing of Plasmodium falciparum (Malaria causing parasite). IBM's Intel based eX5's architecture, and especially its unparalleled and huge memory capacity, helps applications such as Velvet to meet the challenges of extreme memory requirements and large volumes of output data. Besides the unique memory architecture, eX5 solutions provide outstanding compute speed, performance, scale, flexibility, reliability, energy, simplified management and overall efficiency benefits.

The key feature of eX5 is its MAX5 snap-on memory expansion unit ("memory drawer") that enables up to double the memory capacity of the standard Intel-supplied chip set for Nehalem-EX. Compared to other server designs with smaller memory ranges, eX5 can deliver significant economic benefits for running and deploying NGS applications in life sciences domain.

Advances in Next Generation Sequencing (NGS)

Gene sequencing data is rich in information. A typical 100 base pair (bp) piece of sequence contains 4^{100} i.e., 1.6e+60 combinations of the four bases A, C, G and T, and the complete haploid human genome is more than 3,000,000,000 bp[6]. Until mid-2000, the Sanger chain termination method (1977) was the gold standard in sequencing . It could typically generate a 1000 bp high-quality sequence called "read". Subsequent technology innovations have ushered in a new wave of NGS techniques aimed at providing inexpensive human genome sequences. This has led to huge gains in DNA sequence throughput and reductions in cost-per-base. Advances in Nanotechnology have enabled scientists to run hundreds of thousands to millions of sequencing reactions simultaneously. As a result, the DNA sequencing field is brimming with innovative ideas[7] that are path breaking in terms of application to healthcare and clinical

6 Paper: Techniques of Genome Mapping and Sequencing
7 Computational Analysis of Metagenomes: Daniel Huson – Chair of Algorithms in Bioinformatics

research (*Figure 2*). Today, there many new and many under development instruments that promise to offer orders of magnitude higher sequencing throughput per dollar spent than before. Despite the high cost of newer instruments, the volume of data generated (see *Figure 3*) and high error rates, NGS has opened up the field of genomics enabling researchers to pursue several new avenues than were impossible with the older techniques.

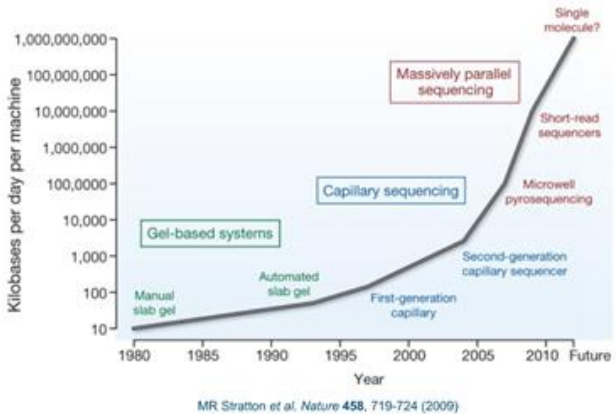

Figure 3: Perspectives - Kilobases generated per machine per day

Current state of affairs in Gene Sequencing: In spite of the intense research and technological advancement in genomics, there are surprisingly few organisms with fully sequenced genome today. The timeline and the amount of work that is yet to be done in terms of gene decoding beyond the simple bacteria[8] is indicated in *Figure 4* below. We are yet to overcome technical and computational hurdles before sequencing becomes an automatic process and scientists can sequence majority of the species or even learn the role and origin of the non-coding DNA sequences. Human Genome involves accessing and processing exabytes of data and analyzing terabases of sequences – this is a difficult task even with the latest technology. Some of the key challenges that NGS researchers face are data storage, access, faster analysis, tools for navigating genomics data, and dealing with metadata.

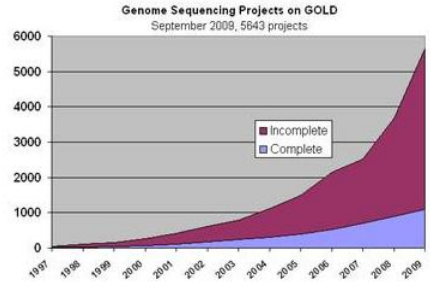

Figure 4: Sequencing of Genomes ***(Source – GOLD Genomes Online Database, www.genomesonline.org)***

Applications of Gene Sequencing

With the conclusion of the Human Genome Project, a number of radically new sequencing technologies have reached exciting phases of development. These technologies can dramatically increase the output of DNA sequence per machine and slash the cost by miniaturizing the process and performing millions of reactions in parallel. Although in most cases the individual read lengths of each DNA fragment are much shorter than traditional Sanger methods, the sheer quantity of sequence and sophistication of computational methods for assembling that sequence are such that these technologies are already finding many exciting applications in academia and industry. As newer sequencing platforms offer orders of magnitude improved throughput and cost efficiencies, routine personalized genome sequencing is suddenly looking to be within the reach of consumers and several companies are looking at commercializing these advances. The most common application areas[9] of gene sequencing are 'Re-sequencing'

[8] Lecture: Metagenomics and current state of Genomics

[9] Inside Pharma Reports – NGS

followed by *de novo* sequencing, genotyping, comparative genomics, systems biology, bioinformatics, diagnostics and protein functions (*Figure 5*).

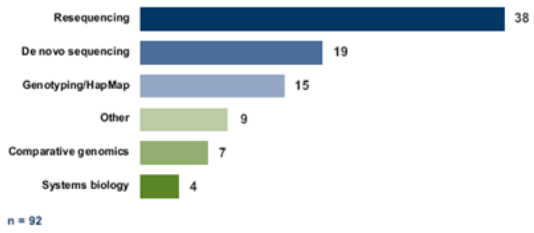

Figure 5: Applications of Gene Sequencing ***(Source – Insight Pharma Reports, NGS Survey, 2007)***

The trend towards faster sequencing has not only pushed the cost of sequencing lower but also closer to the goal of a $1000 for a human genome that many companies and researchers hope to achieve in the near future. At this price point, experts believe that human genome sequencing will reach the global middle class that approximately spends similar amounts for newer routine medical procedures. A genetic readout of coding regions of an individual's DNA could potentially reveal their predisposition to common and rare diseases. This information can help individuals tune their diet, lifestyle, and medication for a full, healthy and better life.

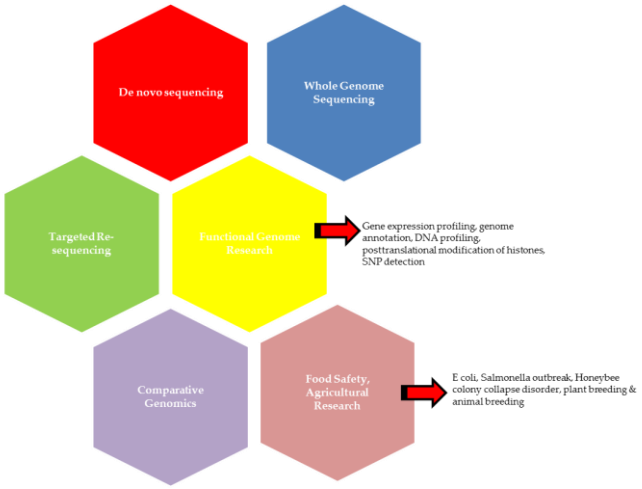

Figure 6: Applications of NGS (*Source IBM*)

Already, disciplines such as meta-genomics, epigenetics, discovery of non-coding RNAs, and protein-binding sites use some of these newer sequencing technologies today. There are essentially two types of problems: alignment problems (for which a previously sequenced reference sequence is available) and *de novo* assembly problems (for which no reference sequence is available). The problem lies in dealing with volumes of experimental data during the analysis. Some of the NGS applications such as Velvet require enormous amounts of system memory to be able to load and analyze the requisite data sets to produce meaningful inferences from the genomic experiments.

Each of the NGS technologies strikes a different balance between cost, read length, data volume and rate of data generation. The focus of first wave of NGS technologies was to re-sequence genomes in a shorter time and at a lower cost as compared to traditional Sanger sequencing. The Solexa GA platform and 454 GS20 pyrosequencing that were developed by Illumina and Roche respectively, generated reads of around 36 and 100 nucleotides respectively. These short reads were adequate for re-sequencing applications but it was widely assumed that they would be too short for *de novo* assembly. Since the introduction of these technologies, the ambition and scope of applications have increased enormously culminating in large-scale meta-genomic and evolutionary analysis of tens of species simultaneously.

Gene Sequencing Ecosystem

Over the past five years, NGS has revolutionized large-scale sequencing resulting in a drastic increase in the number of bases obtained per sequencing run (*Figure 3, Figure 7*) while at the same time decreasing the costs per base. Compared to Sanger sequencing, NGS technologies yield shorter read lengths. Despite this drawback, they have

greatly facilitated genome sequencing, first for prokaryotic genomes and within the last year for eukaryotic genomes. This advance was possible due to the development of software that allows the '*de novo'* assembly of draft genomes from large numbers of short reads. In addition, NGS is now used in case of meta-genomics studies, detection of sequence variations within individual genomes, e.g., single-nucleotide polymorphisms (SNPs), insertions/deletions (indels), or structural variants. Earlier, many of the high-throughput studies utilized hybridization-based methods such as microarrays. They have quickly adopted NGS technologies now. This includes the use of NGS for transcriptomics (RNA-seq) or the genome-wide analysis of DNA/protein interactions (ChIP-seq).

Genomic projects underpin almost all aspects of modern biology. This includes molecular biology, biodiversity studies and medical research including drug development and vaccines. As part of their NGS efforts, many research institutions have purchased newer instruments. Consequently, they now need to store, curate, and access and analyze immense amount of generated sequence data. The new DNA sequencing equipment generates billions of base pairs worth of sequence data per day and this will only rise. This new equipment shifts the bottleneck away from the generation of DNA data onto the ongoing data management, data access and data processing to ensure that the information is readily available to support the research. Some of the established and emerging players who have brought NGS applications to the market include Life Technologies, Applied Biosystems, Illumina, Roche 454, Pacific Biosciences, Ion Torrent, Helicos and several others. Many facilities around the globe today deal with high-throughput sequencing[10].

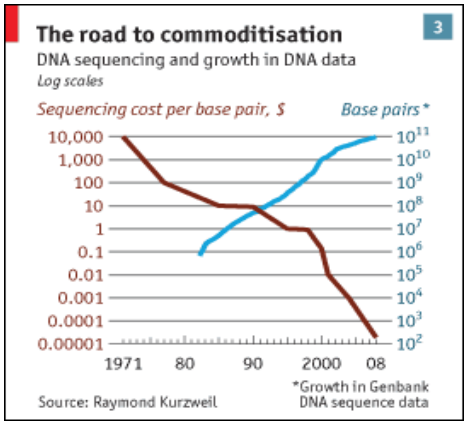

Figure 7: DNA sequencing growth (*Source: The Economist*)

Disease and Drug Research - are we there yet? The final genome sequencing was supposed to lead genomics into cures for many diseases including cancer, diabetes and depression. Scientists believe that the real cures are still years away. There are still many barriers between knowledge of the genome and final solutions to diseases despite the significant lowering of costs through newer sequencing techniques available today.

Importance of Gene Sequencing in an era of global competition: Healthcare is receiving increasing importance and priority from governments and economists around the globe[11]. The Beijing Genomics Institute (BGI) is purchasing[12] 128 new HiSeq 2000 sequencing systems to aid $1.5 billion of research in the areas of sustainable development, healthcare, agriculture and bio-energy. This purchase represents the largest single order of such systems to date. As many of the drug patents reach maturity, and with the increase in number of diseases that can be cured with targeted drugs, it is becoming increasingly important to invest and lead such research initiatives.

Healthcare and Gene Sequencing

Once the computational challenges are addressed, experts predict that the cost and efficiency gains in sequencing could usher in an era of personal genomics with personalized, predictive, preventive, and participatory medicine within a decade. They see an urgent need to develop semantic ontologies that span genomics, molecular systems biology, and medical data. Although the development of such ontologies would be costly and difficult, the benefits will far outweigh the costs. Availability of such ontologies would allow a revolution in web-services for personal

10 Genomics: High-throughput sequencing facilities http://ngsbuzz.blogspot.com/
11 The Economist – article on healthcare
12 BGI recent purchases

genomics and medicine. For more than a decade now, the cost per nucleotide of DNA sequencing has been reducing exponentially.

The availability of cheap, diploid, full-genome sequences may still be several years away. However, there are many low-cost tests for large numbers of SNPs and other sequence variations that are already being used by companies such as 23andMe, NaviGenetics and deCODE Genetics to provide personalized disease-susceptibility profiles. KNOME is offering full personal genome sequencing for those who can afford the current costs. The extent to which the availability of such data will lead to improvements in health-care depends on *four criteria*, as set out by the US Centers for Disease Control:

- Accuracy of genotyping,
- Predictive value of genotypes,
- Clinical utility of knowing genotypes,
- Ethical, legal and social issues involved.

Technical Barriers for NGS: System resources (CPU, Memory)

Steven Salzberg, director, Center for Bioinformatics and Computational Biology, University of Maryland, has some interesting observations regarding technological barriers. He believes that managing and analyzing NGS data and the software that is making it happen is critical to the evolution of genomic research. It is not only important to keep the processed data but also to distinguish the raw data from the sequencer's images. Files of imaging data from the sequencing plates, gels, or slides are gigantic. Image-processing software figures out the nucleotides from the images and generates files that are large, but not nearly as large as the images. Therefore, it is important to save all the sequence reads in order to recall them during analysis, but not the raw images. Those images can generate about a terabyte of data for one experiment. Once that terabyte of data is compressed down to DNA sequence, it converts to tens of gigabases.

Another issue is that software designed for Sanger sequencing or other traditional techniques may not be adequate for NGS or the short-read re-sequencing. Instrument manufactures, academics and third party software companies are all working together to address this issue. For the task of assembly, that is, reconstructing a genome from the reads, there are new assemblers, which are geared for short reads. According to Salzberg, Velvet has the best and most popular assemblers available today and it works quite well for assembling very short reads. However, one of the major limitations of new assemblers is that for very short reads they do not yet seem able to handle large mammalian-sized genome. The problem during assembling a mammalian or animal genome from short reads is the sheer amount of data. That requires researchers to be very careful while using these algorithms in order to manage the memory of the HPC systems. It is not just CPU time that is an issue but also memory. Analysis requires reading all data in at some point, and the computational systems and setup do not have sufficient memory. The machines just cannot handle it and the system can potentially crash causing loss of data and effort.

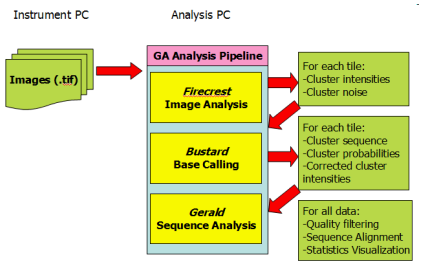

Figure 8: Workflow for NGS data analysis

Sanger's Role in Gene Sequencing

Sanger Institute pursues research that builds understanding of gene function in health and disease and creates resources of lasting value to biomedical research. Genomes are the archival instructions upon which an organism is built. The sequence data provided by the Human Genome Project is a rich source of information that drives improved understanding of human health and variation. Studying human sequences, comparing model organism genomes and investigating the effects of pathogens on humans builds knowledge of the diversity of our genomes and how this affects our susceptibility to disease.

NGS at Sanger: A fundamental role of DNA sequencing is to generate large continuous regions of DNA sequence. The 'Whole-Genome' shotgun method has proven to be the most cost-effective and least labor intensive method of sequencing that was applied for human genome and completed by a BAC-by-BAC strategy. The Capillary

sequencing reads are ~600-800bp in length and these methods involve computing all overlaps of reads and then resolving the overlaps to generate the assembly. However, the volume and the short read length of data from the next generation sequencing machines cannot utilize the read-centric overlap approaches. Pevzner et al. introduced an alternative assembly framework based on '*de Bruijn*' graph[13] that was based on an idea of a graph with fixed-length subsequences (k-mers). The key in this method is not storing read sequences but k-mer abundance information in a graph structure. *Figure 8* shows a typical workflow for NGS data analysis.

HPC infrastructure at Sanger: Typical storage and compute required at Sanger for genomic research analysis project are depicted in *Figure 6.* There is prodigious amount of data generated by high-throughput genomic research equipment. The rate of data production far exceeds the capability to analyze it. The process and workflows required to analyze the data are extremely complex and resource intensive. The information storage equipment required for 'Vertebrate Re-sequencing' activity at Sanger is shown in *Figure 10*.

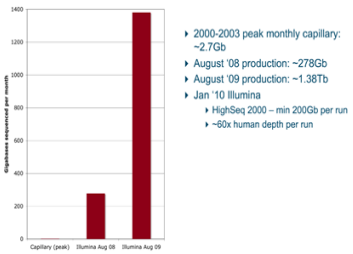

Figure 9: Typical Storage and Compute requirements at Sanger

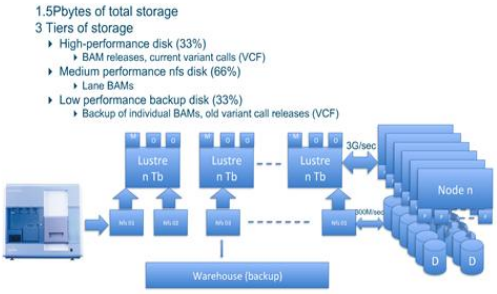

Figure 10: Vertebrate Re-sequencing Information Storage ***(Source: Sanger)***

A look at Sanger's future HPC needs: The Sanger Institute aims to be at the leading edge of genome scale scientific research, and high throughput sequencing is the bedrock of their work. At a generation rate of over a terabase of sequences a week, the key requirement at Sanger from HPC is downstream meta-analysis of this sequencing data. The rate of increase in sequencing technologies puts a lot of pressure on the HPC infrastructure. Every six months, Sanger has a *twofold* increase in their compute and storage requirements as the rate of data output from their sequencers doubles. Some of the primary NGS software in use is listed in *Figure 11*.

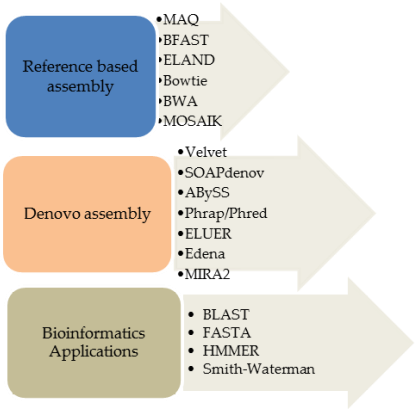

Figure 11: Primary NGS Software

How Sanger benefits from IBM solutions

As part of several gene sequencing studies that utilize huge datasets such as the Human Genome sequencing of an African Male individual and Plasmodium Falciparum 3D7 clone – Malarial parasite, the Sanger Institute uses latest techniques, analysis and NGS algorithms including Velvet, SOAP, and AbySS. IBMs eX5 based systems have been successfully deployed at Sanger for conducting experiments related to many such studies where genome dataset sizes range from 3.5 GB (Velvet), 170GB (ABySS) and 3.3GB (SOAP).

At Sanger, a typical ABySS run using a dataset belonging to African Male individual sequencing study and "read"

[13] Slide #75 – NGS Tutorial by Thomas Keane

size of 68bp consumed 180GB of system memory on an IBM x3950 M2 [72334MG] system with 8 Intel(R) Xeon(R) quad core CPU E7400 @ 2.40GHz and 512 GB of memory. This was a Debian/Lenny OS and Lustre 1.8 File System based setup. In another experiment that used Plasmodium Falciparum – the Malarial parasite sequencing dataset, the peak memory consumed by Velvet (kmer =21) running on the same IBM system was 334GB. For the same dataset – SOAP used only 7G of peak memory. Some of these runs especially for Velvet with kmer=21 would not even complete earlier on systems that were not equipped with large memory as in IBM's eX5 series systems.

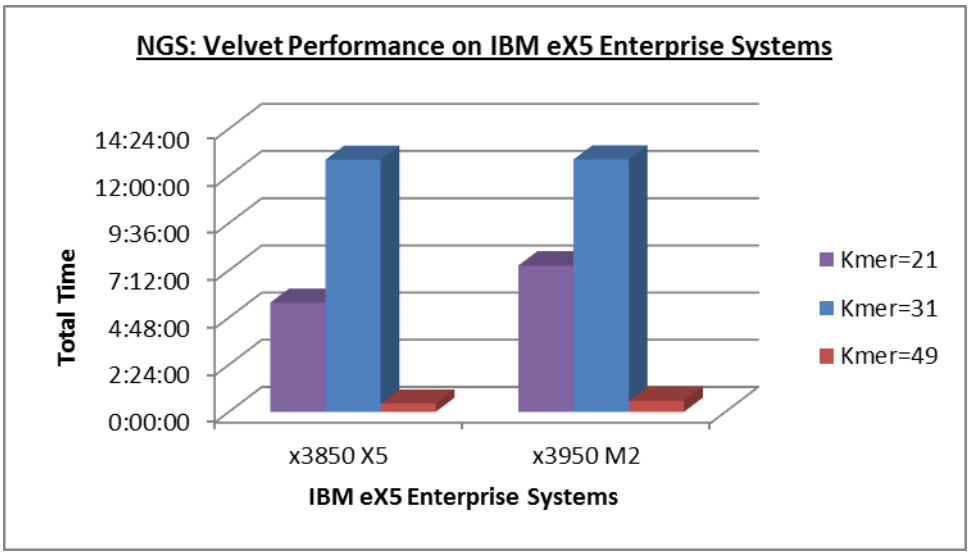

Figure 12: NGS: Velvet Performance Improvements with IBM eX5 Enterprise Systems (*lower time is better*)

For a perspective on how Velvet operates, consider some of the basics in terms of peak memory requirements and the volume of data generated through Velvet. Velvet is a set of algorithms manipulating '*de Bruijn*' graphs for genomic Sequence assembly. It was designed for short read sequencing technologies, such as Solexa or 454 Sequencing and was developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute. The tool takes in short read sequences, removes errors, and produces high quality unique contigs. It then uses paired-end read and long read information, when available, to retrieve the repeated areas between contigs. Among other assembly methods that use paired-end reads, laboratory tests on Solexa 36 platform indicate that Velvet performs best in terms of resulting assembly length and accuracy of each contig. Velvet assembled 96% of the genome with an error rate of 0.33% per nucleotide. The sequencing field is evolving rapidly and the assembly software packages such as those for Velvet are under active development and are likely to continually improve. For example, the algorithm for exploiting paired-read information was substantially revised for Velvet version 0.7 compared with version 0.6, yielding a tenfold increase in N50 contig length[14].

In theory, Velvet's algorithms work for any size of reads. However, the engineering aspects of Velvet, in particular the memory consumption, make it incapable of dealing with read sets of a particular size. This of course depends on how big a real memory machine is used for Velvet. There are cases where Velvet has been "routinely" used for multiple strains such as Drosophila sized genomes that require ~120MB on a 125GB machine. It is common for Velvet to be used into the 200-300MB region, but rarely further[15]. Although sheer read size is important, the memory size is not just about the size of the genome but also about how error prone the reads themselves are.

A key feature of IBM's Intel based eX5 architecture is its MAX5 snap-on memory expansion unit ("memory drawer") that enables up to double the memory capacity of the standard Intel-supplied chip set for Nehalem-EX which is also used in other servers based on the same chip. However, as compared to server designs with smaller memory ranges, eX5 can deliver significant economic benefits and satisfy the much needed peak memory requirements of NGS applications such as Velvet. *Figure 12* compares Velvet's performance on an Intel 7500-based IBM x3850 system vs. the one running the same Velvet experiment at Sanger using Intel 7400 processor based IBMx3950 system. It shows how the new memory architecture of IBM x3850 (Intel 7500) boosts Velvet performance for k-mer=21. For an IBM x3850 based on Intel 7500 processor, Velvet (k-mer=21) takes much less

[14] Paper - Application of next generation sequencing technologies to microbial genetics by Daniel MacLean, Jonathan D. G. Jones and David J. Studholme
[15] Source: Ewan Birney's post on Velvet user list

time than the IBM x3950 system based on Intel 7400 processor with same amount of RAM – 512 GB.

Memory Drawer for NGS extreme memory requirements: Today, x86 clusters are the mainstay for running bioinformatics and sequencing code at the Sanger Institute. IBM's Intel x86 based eX5 architecture and large memory capability provides the best fit in terms of price, performance, energy and space requirements at Sanger. IBMs eX5 system benchmark results for Velvet, SOAP and ABySS runs using datasets from the Human Genome Sequencing African individual and Malaria Parasite – Plasmodium Falciparum clearly indicate IBM's offering as the most suitable and cost effective solution with a 512 GB memory footprint.

Compact Architecture for Simplified Management: Sanger Institute is a multi-vendor shop when it comes to their HPC infrastructure and hence standards such as IPMI are critical in terms of systems management. Being able effectively manage large numbers of machines with as little human intervention as possible is vital to Sanger. IBM eX5 compact architecture helps Sanger deal with their HPC infrastructure for NGS research with a relatively small number of system administrators.

Green, optimal floor space & energy efficient: At the Sanger Institute, power, space and cooling are becoming increasingly important in the context of NGS research. Space and power constraints drive them to blade servers, but they do move away from blade form factor when requirements dictate otherwise – for example, when faced with the memory requirements of NGS research. IBM's eX5 has 512 GB of memory that helps memory-intensive NGS research such as ABySS, see *Figure 13* below.

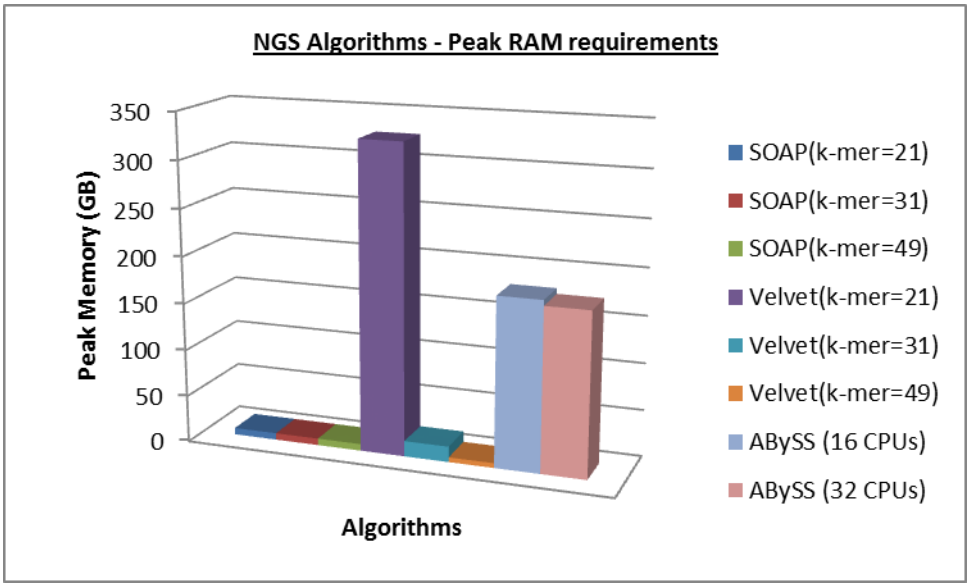

Figure 13: IBM eX5 Enterprise Systems address huge memory demands of NGS Algorithms

How IBM eX5 Architecture meets the challenges of NGS

IBM eX5 is the fifth generation of the Enterprise X-Architecture that takes full advantage of Intel's next-generation Nehalem-EX processor. It is an x86 server design optimized for hosting workloads in demanding enterprise environment with two new rack-mounted servers - the x3690 X5 with 2 sockets and the x3850 X5 with four sockets. This is the first time IBM is extending the Enterprise X-Architecture to a blade form factor with the BladeCenter HX5.

IBM's unique advantages Scalability & Reliability: While most other x86 server suppliers use Intel's standard chip sets as the building blocks for complete server systems, IBM has differentiated its x86 servers with unique technology since acquiring Sequent in the late 1990s. The first few generations of this technology, which IBM brands as Enterprise X-Architecture, offered moderate differentiation, at least enough for IBM to separate itself from the commodity orientation of most other x86 suppliers. The architecture hit its stride with eX4 in 2007, which introduced unique reliability and scalability features, including the ability to protect against memory failure at three different levels – a feature that can dramatically boost the uptime of servers. While eX5 still has still reliability features than lower-end Intel Xeon platforms, many of its earlier reliability features have been incorporated into the base Intel architecture. Moreover, scalability and reliability is now commonly achieved with clustering software rather than with hardware features.

Breaking the Gene Sequencing Memory Barrier: In today's highly virtualized enterprise environments, and for the memory hungry next generation sequencing technologies, users consider the memory capacity of a server to be more important than almost any other feature, according to IBM. Many servers utilize only a fraction of their CPU power. Most modern servers have to be rebalanced with far larger memory ranges per processor than previous designs. A key feature of eX5 is therefore its MAX5 snap-on memory expansion unit ("memory drawer") that enables up to double the memory capacity of the standard Intel-supplied chip set for Nehalem-EX that most other x86 systems suppliers will use in their servers. Compared to server designs with smaller memory ranges, eX5 could deliver significant economic benefits in terms of requirements of the Next Generation Sequencing processes. Many of the NGS deployments are typically memory bound, rather than processor bound -- larger memory ranges mean that more reads can be deployed per socket. The superior memory ranges in eX5 will also be valuable for customers who want to deploy other classes of workloads on x86 servers such as very large databases.

IBM's Scale and flexibility: IBM originally started working on scalable and reliable memory management in the X-Architecture with the intention of supporting large scale-up SMP servers for monolithic workloads. These investments are now ready to pay off with IBM's Nehalem-EX systems, which may allow as many as 128 cores (the maximum supported by the eX5 architecture) to be harnessed in a single server footprint. This can tremendously boost the gene sequencing yield times given the scale and speed of processing power.

IBM's unique interconnect: Even on individual 2-socket servers, the larger memory range granted by MAX5 drawers promises to deliver significantly better value relative to other Nehalem-EX –based platforms. IBM's implementation of MAX5 relies on the standard scalability port built into Nehalem-EX, but it also uses some IBM eX5 technology for linking computing modules. While competitors can duplicate functions that use Intel's standard scalability ports, the other functions are unique to IBM.

Conclusion

The NGS research at the Sanger Institute is a typical example of how HPC is used in leading edge research in life sciences.IBM estimates that IT spending for NGS is conservatively $200M/year. Multiple industry studies indicate that revenue from HPC servers will continue to grow much faster than overall server revenues with clusters being the dominant platform –almost 70% of HPC servers[16]. However, these studies also suggest that the costs of server management, power and cooling, and facilities management in HPC data centers will outpace the costs of buying new servers. This has caused a severe crisis in HPC data centers. And IT solution providers such as IBM have responded by designing innovative solutions such as IBM's Intel based eX5 with huge memory capacity that address these issues in running life sciences applications and NGS algorithms such as Velvet while retaining all the attractive attributes of industry standard cluster architectures for HPC. HPC users have always demanded computing solutions that have the best performance, price/performance, and now are increasingly demanding energy efficient platforms.

Over the last decade, with the widespread penetration of industry standard clusters, HPC capital expenses as a percentage of IT spend have decreased. However, the associated operational expenses to manage these higher computing density HPC data centers have escalated largely because of increased costs in systems administration, energy, and facilities. IBM eX5 series of servers perform excellently on a wide range of life sciences problems such as high memory requirements of Velvet and volume of output data generated through its super large memory footprint (512 GB). For a given price/performance, IBM provides flexible, reliable, simplified management solutions that are energy and power efficient and aimed at assisting scientists deal effectively with the compute and memory needs of resource hungry, complex and continuously evolving life sciences applications. For more details on IBM's Intel based eX5 servers visit IBM eX5 product website (http://www-03.ibm.com/systems/info/x86servers/ex5/).

[16] IDC's Top 10 HPC Predictions For 2008, http://www.hpcwire.com/hpc/1987614.html