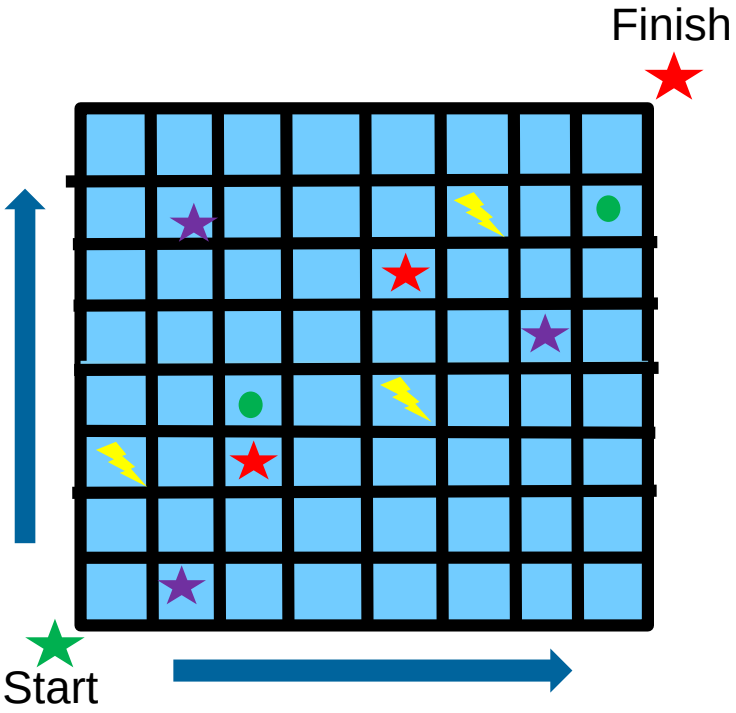
IBM Spectrum Scale

IBM Life Sciences

Madhav Ponamgi
mzp@us.ibm.com

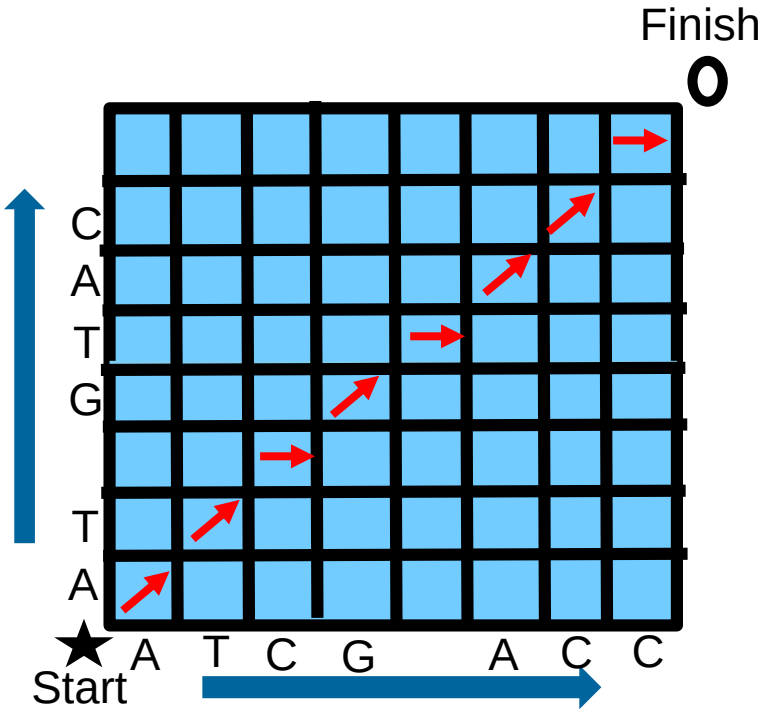
Alignment Paths



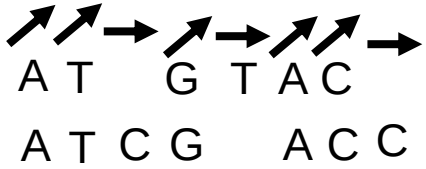
City block traversal problem. Beginning from the start travel North / East to the finish.

- How many possible paths in a M x N grid?
- How long is a path?
- Is there an optimal path?
- What if we weight the edges, can there be an optimal path?
- How do we compute an optimal path?
- What are all the optimal paths?

Genomic Alignment Paths



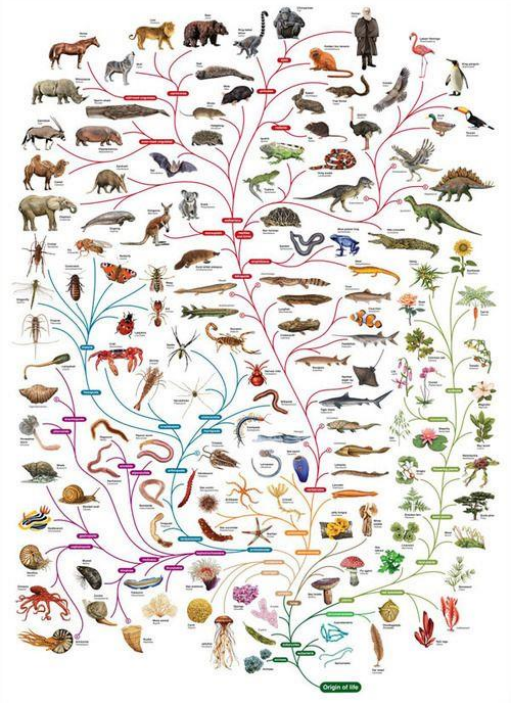
Determining the optimal alignment of short sequence of base pairs.



- Non-trivial to find “best” alignment
- Scoring matrix for multi-dimensional alignment
- Using weighted edges, a Dynamic Programming method can be used to construct matchings
- V. Levenshtein introduced edit distance – shortest sequence of single line edits match two sequences

Phylogenetics

- Sort by reversal: 612345 -> 162345 -> 126345-> 123645-> 123465->123456
- Used for comparing genome to genome sequences
- Across species similar genes are laid in different groups – called Synteny. Mouse and human genomes are similar with about 300 blocks re-arranged.
- How to get from one arrangement to another in the most “parsimonious” sequence of steps
- Depending on length of genomes can be I/O intensive
- Faster method: : 612345 -> 543216->123456



Motif finding methods

CGGGGCTGGTCGTCACATTCCCCCTTTTGATATTTGAGGGGTGCTGCCAATAA
CCAAAAGCGGACAAGGGGATCCGTTGACGACCTAAATCAACGGCCAAGGCC
AAGGCCAGGAGGCGCCTTTGCTGGTTCTACCTGAATTTCTAAAAAAGATTATA
ATGTAATGTCGGTCCTCCTGCTGTACAACTGAGATCATGCTGCTGCTTTCAAC

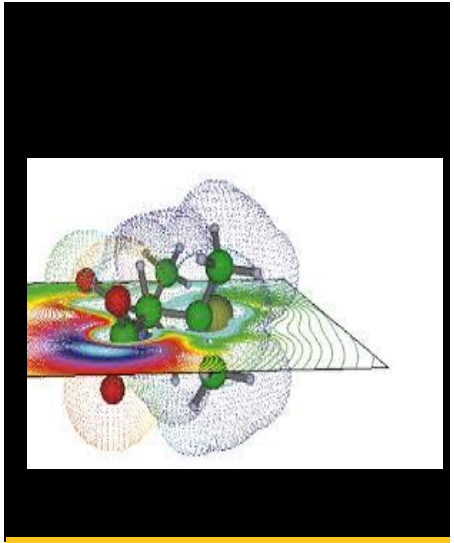
Genomic sequences often contain motifs, short sequences of base pairs, that frame gene encoding sequences

- The motifs may not be known in advance
- The length or size of the motif may not be known (i.e. how many base pairs)
- It's not unusual for the motifs to vary slightly within the genomic sequence (i.e. the motif may have slight mutations as it recurs)
- Goal is to identify and locate these sequences so gene can be identified

Possible solution methods:

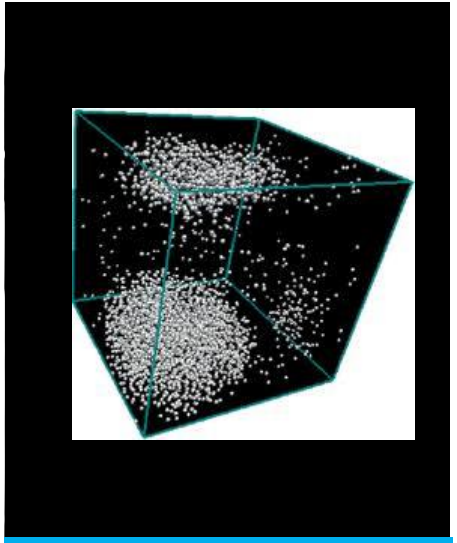
- Brute-force method
  - Examine all size sequences of size k contained within a sequence of size n
  - This results in an exponential algorithm of order O(n^k)
- Branch-and-Bound
  - Using branch-and-bound techniques and search trees the brute force method can be improved upon
  - The I/O pattern for this type search is not restricted serial apart from the initialization because the parallel methods evaluate the genome independently (k << n)

Life Sciences



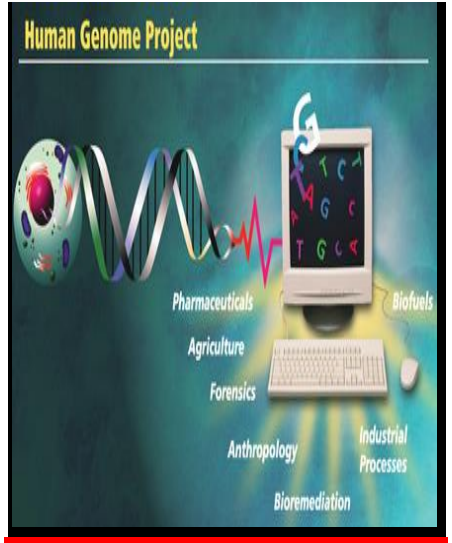
§ **Computational / Quantum Chemistry**

- Ab Inito / Hartree-Fock
- Gaussian
- GAMESS
- MOPAC



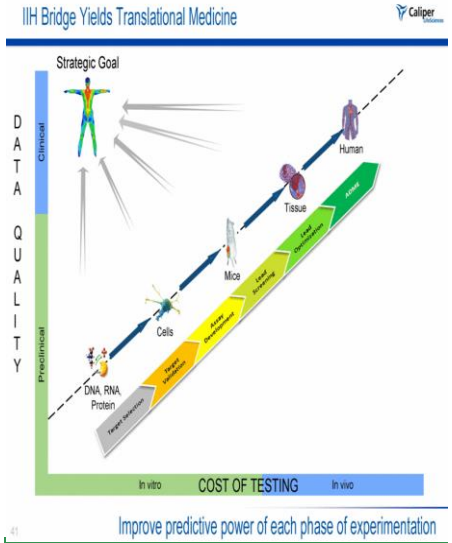
§ **Molecular Dynamics**

- Large molecules / enzymes
- Classical physics / empirical
- NWChem
- NAMD
- Charmm



§ **Genomics / Proteomics**

- Mass spectometry
- De novo sequencing
- Sequest
- Scaffold
- Illumina / Accelerys / CLC

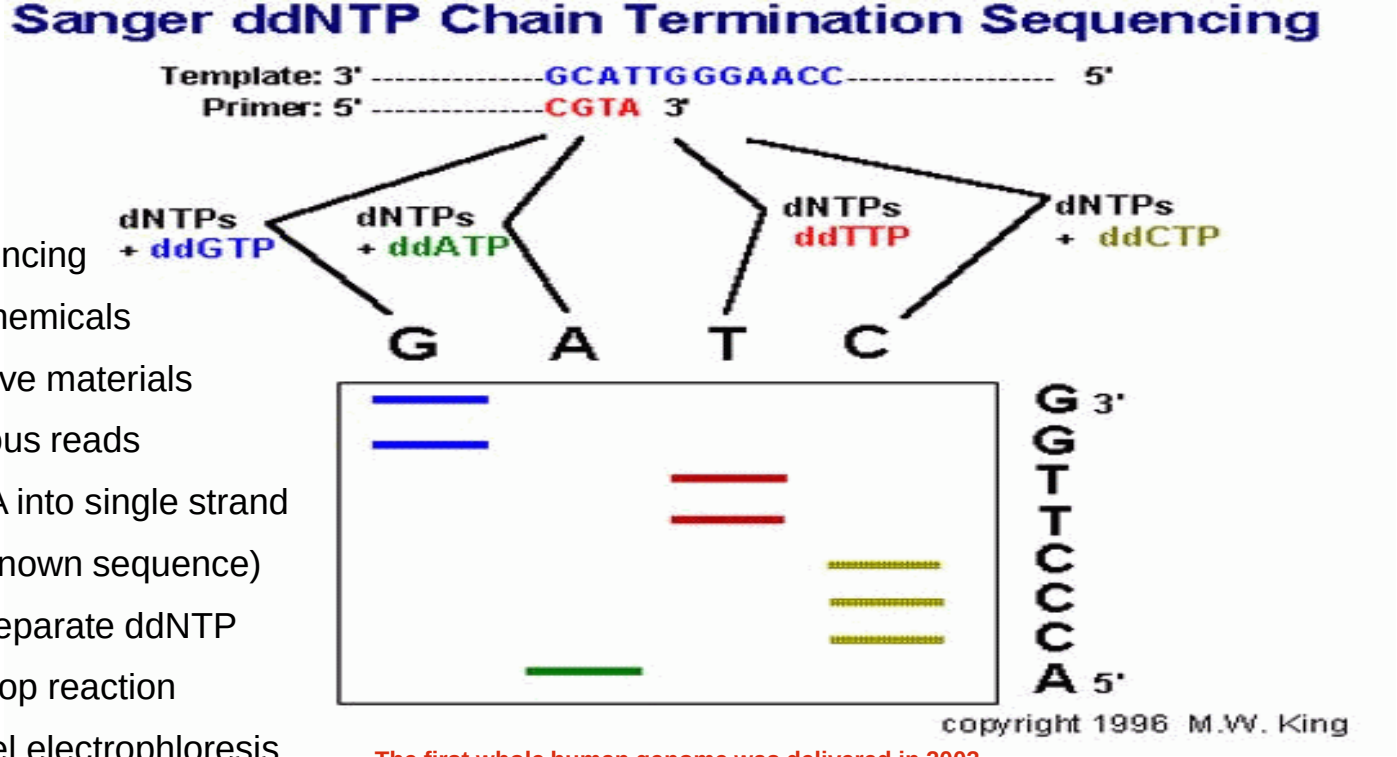


§ **Translational Medicine / Imaging**

- Personalized genomics
- Health records / HIPPA
- Functional MRI
- Data mining / Security

Traditional Sequencing Method – SANGER Method

- Sanger sequencing
- Fewer toxic chemicals
- Less radioactive materials
- Long continuous reads
- Separate DNA into single strand
- Add Primer (known sequence)
- Place into 4 separate ddNTP
- The ddNTP stop reaction
- Sort across gel electrophloresis
- Read back top to bottom

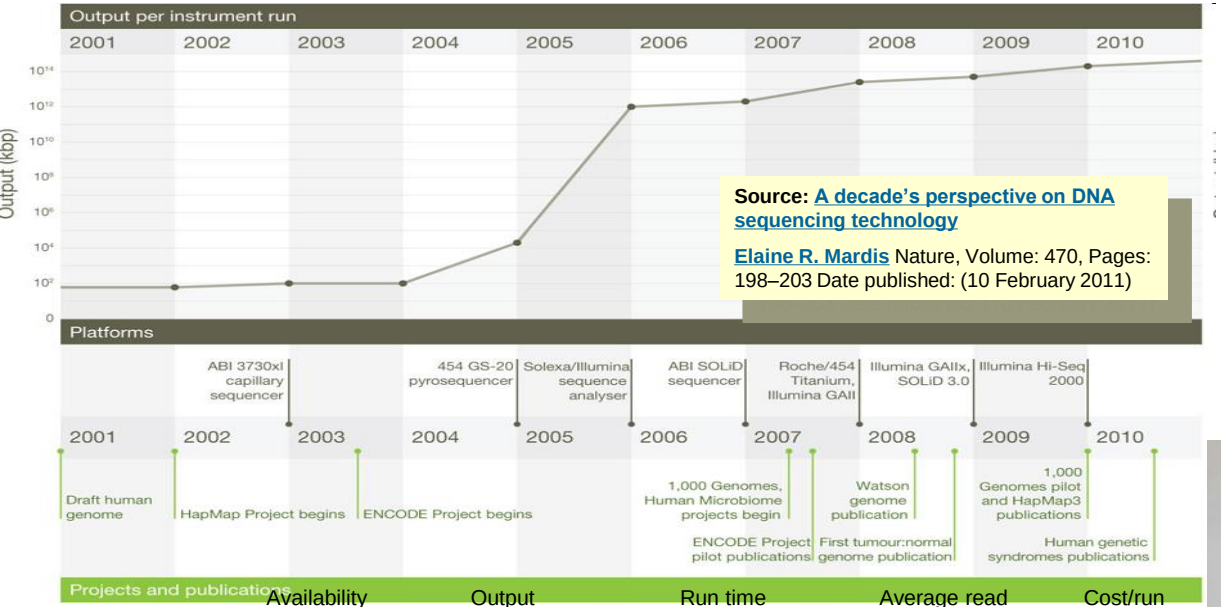


The first whole human genome was delivered in 2003
It cost $3B and took 13 years to complete

Next Generation Sequencers

- Data needs to be assembled and analyzed for gene identification, gene variations and gene functionality

Evolution of DNA Sequencers



	Availability	Output	Run time	Average read	Cost/run
Proton 1	Sept 2012	10Gb	2-4 hrs	200 bp	$999
Proton 2	Mar 2013	100Gb	2-4 hrs	>200bp	<$999
HiSeq 2500	Mid 2012	600Gb	11 days	150bp	$11,000

Source: http://blueseq.com/knowledgebank/sequencing-platforms/

$149,000

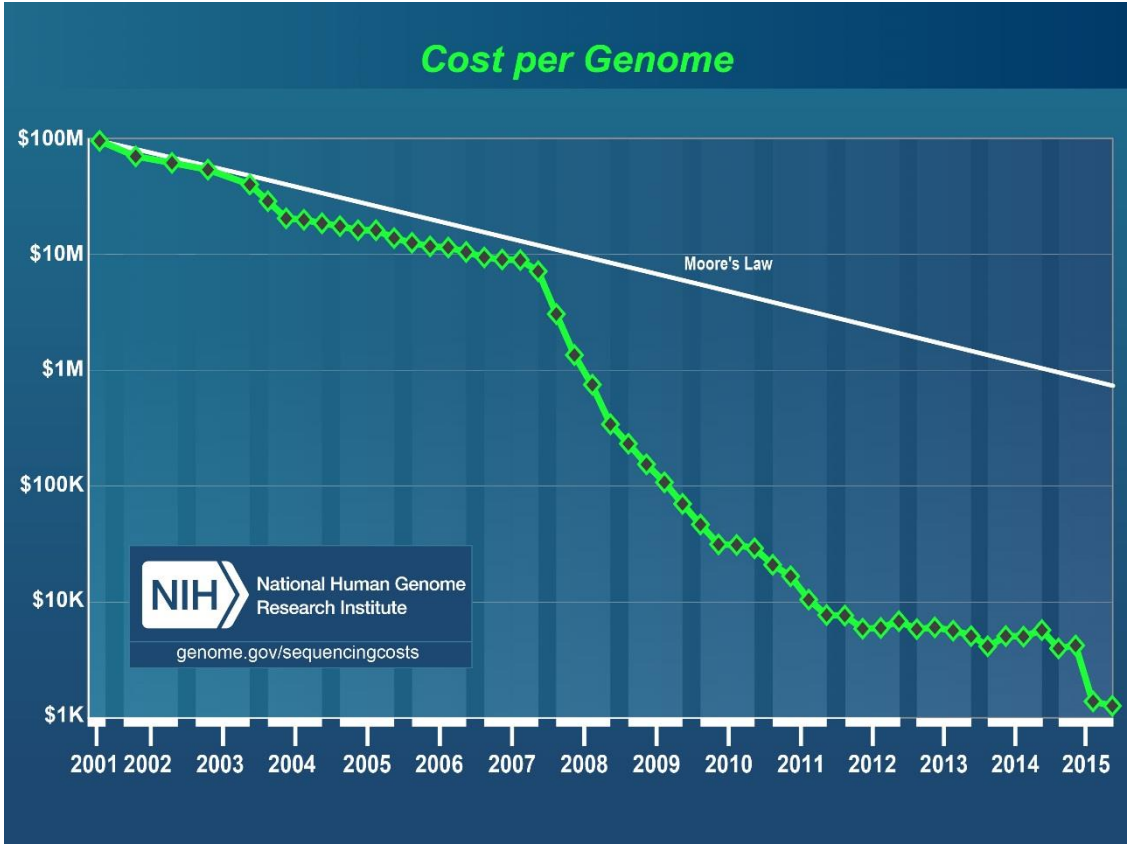
Life Technology Ion Proton Sequencer (late 2012)
http://www.youtube.com/watch?v=OKhxoGcr4Rk

$740,000

Illumina HiSeq 2500

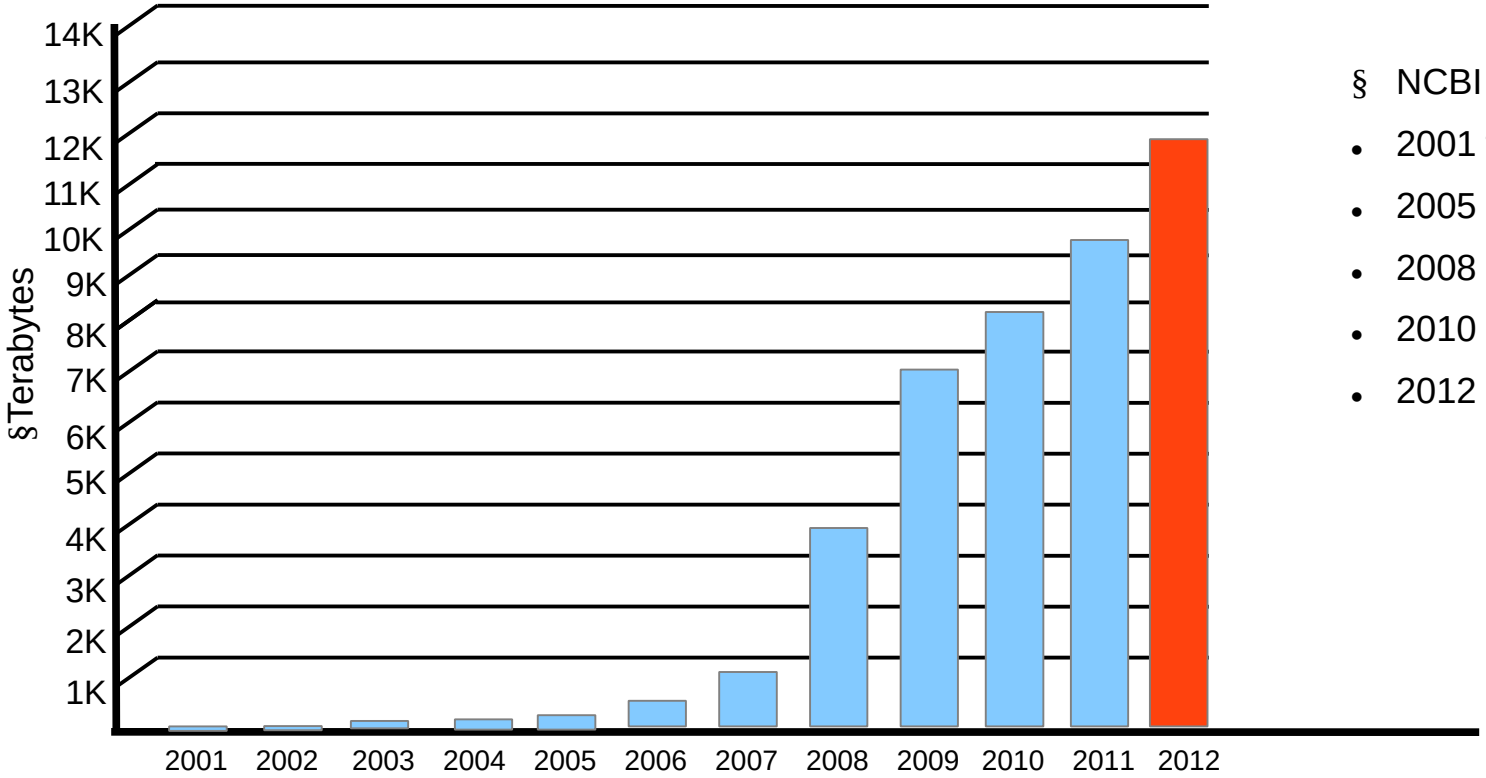
Oxford Nanopore MinION

Cost per Genome



https://www.genome.gov/27541954/dna-sequencing-costs/

Explosive Growth of Data



§ NCBI GenBank

- 2001 first human genome
- 2005 NGS publication
- 2008 Solexa sequencer
- 2010 BGI center opens
- 2012 78TB analysis 1 week

Disk Challenge

- **20 Petabytes by 2018**
- If current 3.5” 8TB disks are a guide.. then we'd be looking at 2,500 disk farm
- Stacked horizontally, this would be a tower 127 meters in height
  - Statue of Liberty, 93 M
  - Eiffel Tower, 324 M

- NEED Analytics
- NEED High Performance Computing
- NEED High Performance Storage

Genomics Data High-Level Pipeline

Base Calling

(Vendor Tool)

▪13 TB

§FastQ/FastA

§Raw NGS Reads

Alignment or

Assembly

▪8TB

§Sam / Bam

§Aligned NGS Reads

Variant

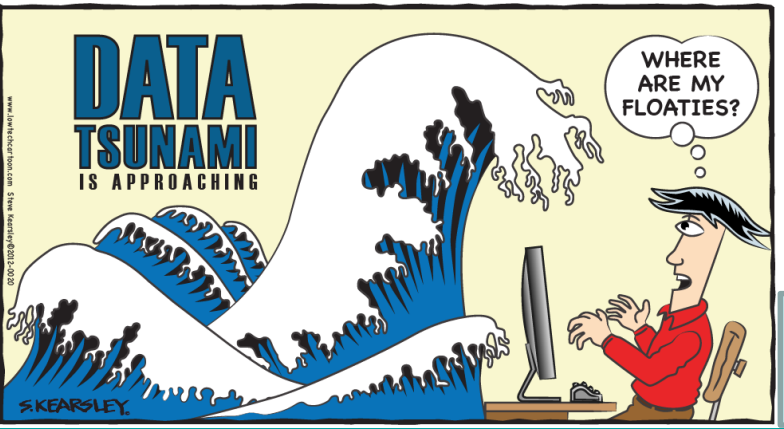
Calling

▪2 TB

§VCF file

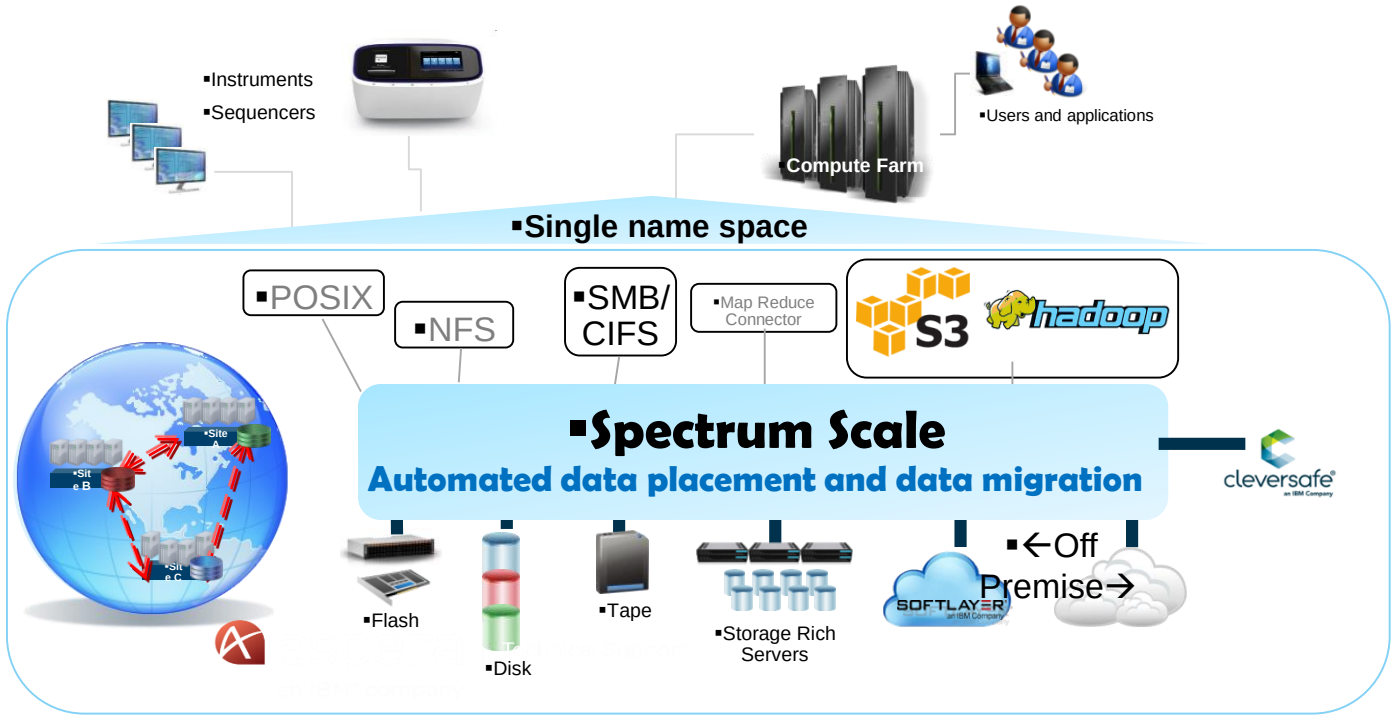
§Genomic Variant

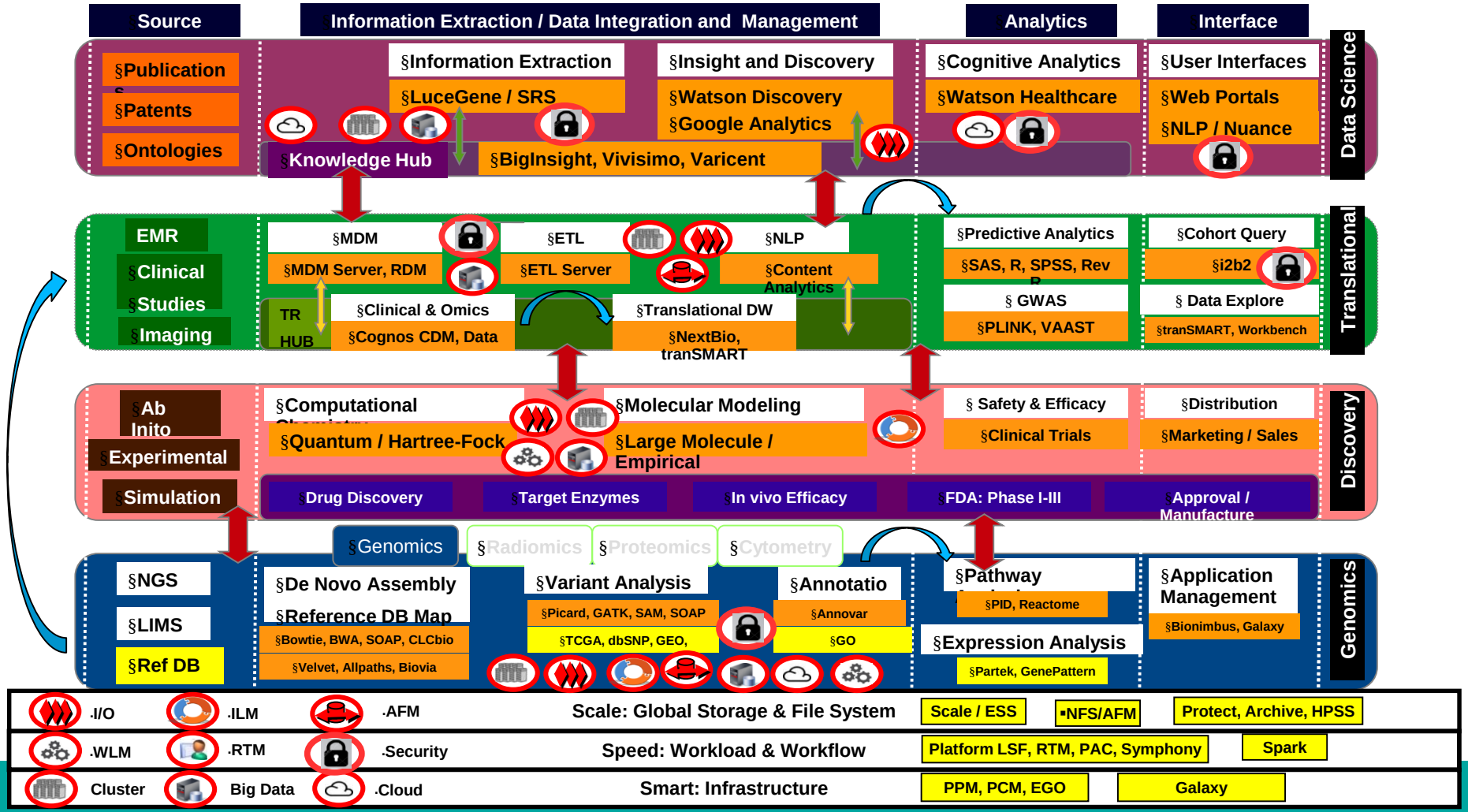
▪200 GB



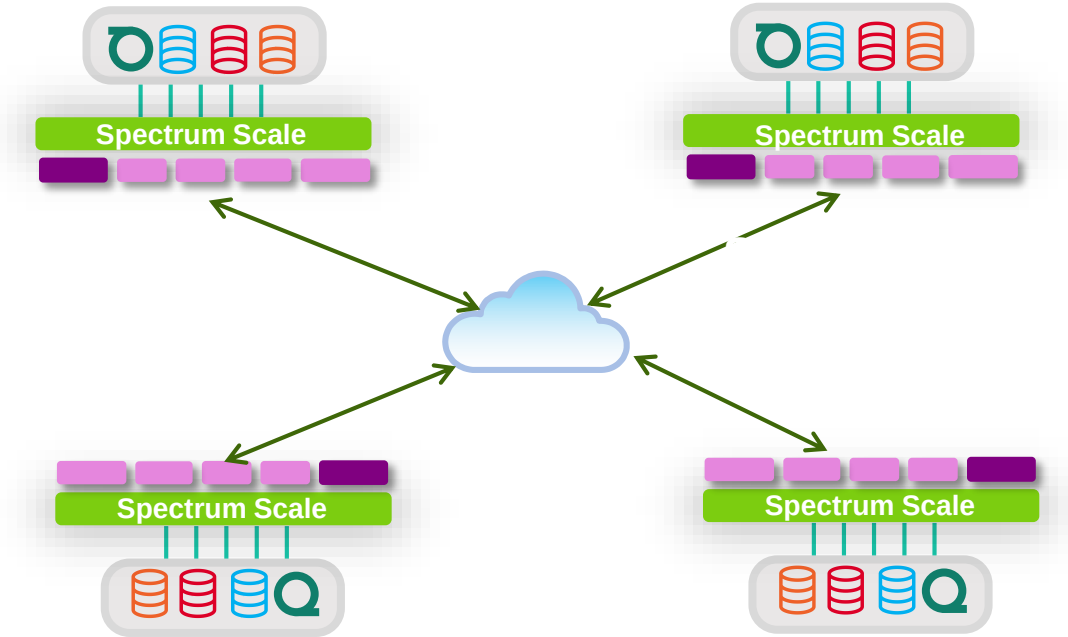
§Downstream Analysis

Spectrum Scale: A File System for Genomic Pipelines



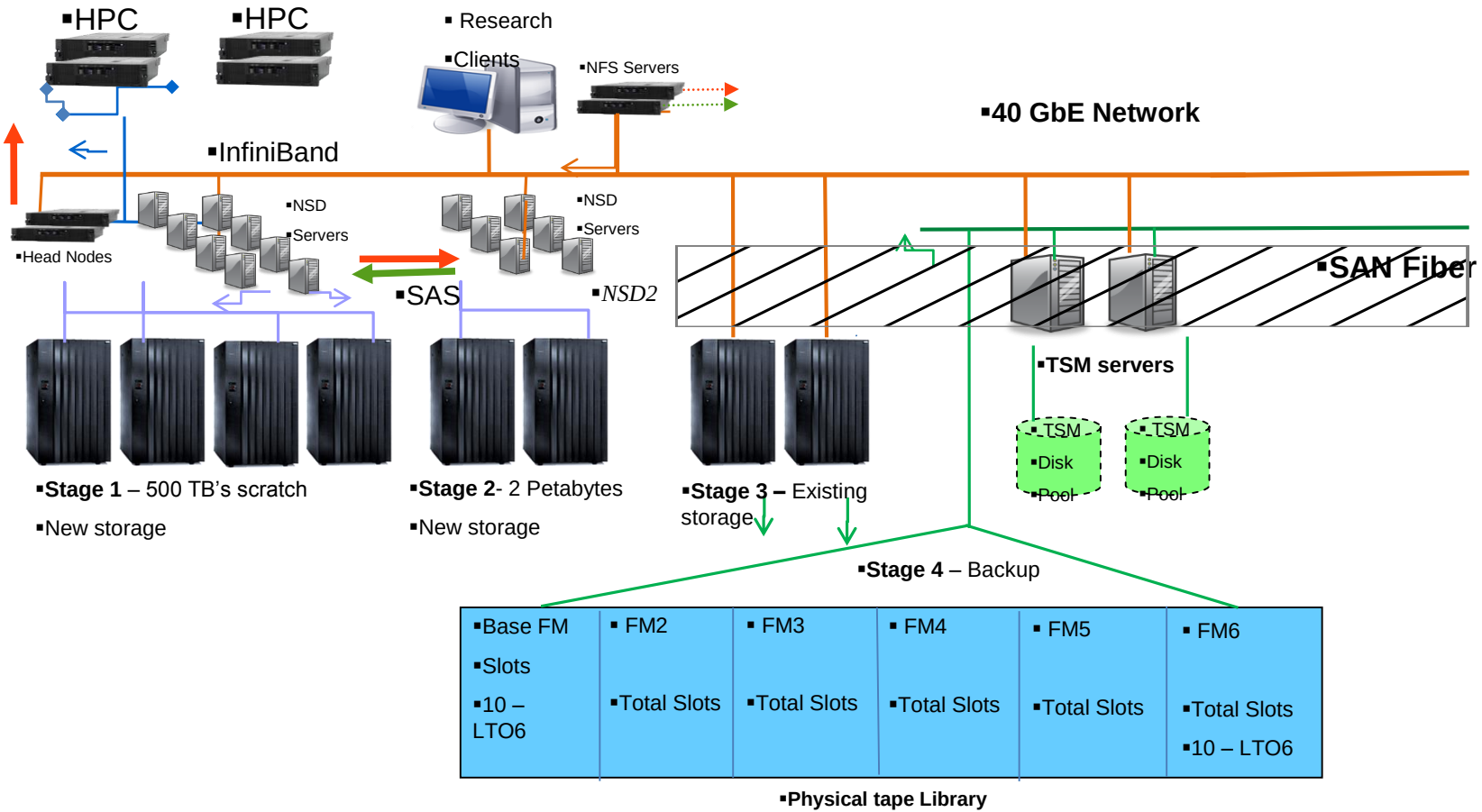

§Source
Information Extraction / Data Integration and Management
Analytics
§Interface
§Publication
§Patents
§Ontologies
§Information Extraction
§LuceGene / SRS
§Insight and Discovery
§Watson Discovery
§Google Analytics
§Knowledge Hub
§BigInsight, Vivisimo, Varicent
§Cognitive Analytics
§Watson Healthcare
§User Interfaces
§Web Portals
§NLP / Nuance
Data Science
EMR
§Clinical
§Studies
§Imaging
§MDM
§MDM Server, RDM
§ETL
§ETL Server
§NLP
§Content Analytics
TR HUB
§Clinical & Omics
§Cognos CDM, Data
§Translational DW
§NextBio, tranSMART
§Predictive Analytics
§SAS, R, SPSS, Rev
§ GWAS
§PLINK, VAAST
§Cohort Query
§i2b2
§ Data Explore
§tranSMART, Workbench
Translational
§Ab Inito
§Experimental
§Simulation
§Computational
§Quantum / Hartree-Fock
§Molecular Modeling
§Large Molecule / Empirical
§ Safety & Efficacy
§Clinical Trials
§Distribution
§Marketing / Sales
§Drug Discovery
§Target Enzymes
§In vivo Efficacy
§FDA: Phase I-III
§Approval / Manufacture
Discovery
§Genomics
§Radiomics
§Proteomics
§Cytometry
§NGS
§LIMS
§Ref DB
§De Novo Assembly
§Reference DB Map
§Bowtie, BWA, SOAP, CLCbio
§Velvet, Allpaths, Biovia
§Variant Analysis
§Picard, GATK, SAM, SOAP
§TCGA, dbSNP, GEO,
§Annotatio
§Annovar
§GO
§Pathway
§PID, Reactome
§Expression Analysis
§Partek, GenePattern
§Application Management
§Bionimbus, Galaxy
Genomics
.I/O
.ILM
.AFM
Scale: Global Storage & File System
Scale / ESS
▪NFS/AFM
Protect, Archive, HPSS
.WLM
.RTM
.Security
Speed: Workload & Workflow
Platform LSF, RTM, PAC, Symphony
Spark
Cluster
Big Data
.Cloud
Smart: Infrastructure
PPM, PCM, EGO
Galaxy

Multi-Site Pharmaceutical AFM Deployment

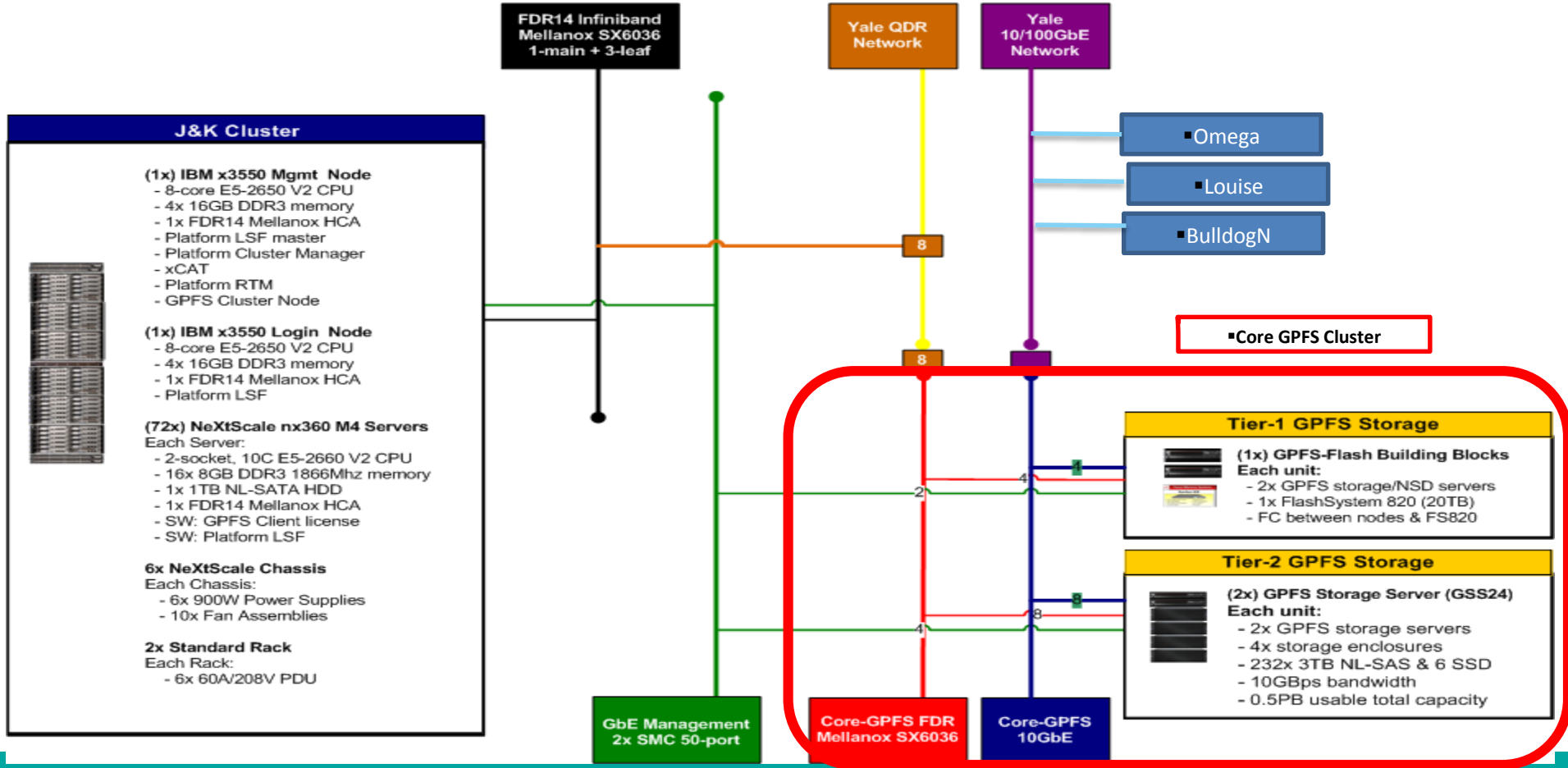


- 4 Sites with roving users
- IW Mode with no Pre-Fetch
- Each site consists of ESS storage with two CES nodes
- Users are in separate directories without common shared files
- Campus inter-connect is 1GigE

▪Major Cancer Center



University System Architecture



Typical Genomics File Size Histogram

scratch	# files	% files	total size	% size
0 bytes	57985	2.04%	0 bytes	0.00%
< 512 bytes	178958	6.31%	40.11 MB	0.00%
< 64 KB	1315112	46.36%	17.72 GB	0.02%
< 2 MB	582402	20.53%	323.16 GB	0.28%
< 100 MB	567624	20.01%	10.58 TB	9.24%
< 1 GB	117519	4.14%	34.00 TB	29.70%
> 1 GB	17092	0.60 %	69.56 TB	60.77%
Total	2836692		114.47 TB	

Personalized Medicine

COLUMBIA UNIVERSITY

Acute Lymphoblastic Leukemia (ALL)

- Relapse 2014 (variation)
- Full Exome sequenced
- Connection to Philadelphia Chromosome and 1-nucleotide mutant NT5C2 + NUP214-ABL1 (resistance to chemo drug)
- Treatment with BMS Sprycel

3 weeks

Diagnosis and Cure

Dr. Wartman, WashU

Cancer researcher and

leukemia survivor

Washington University in St. Louis
School of Medicine

3 weeks

Diagnosis and Cure

Dr. Wartman, WashU

Cancer researcher and

leukemia survivor

Washington University in St. Louis
School of Medicine

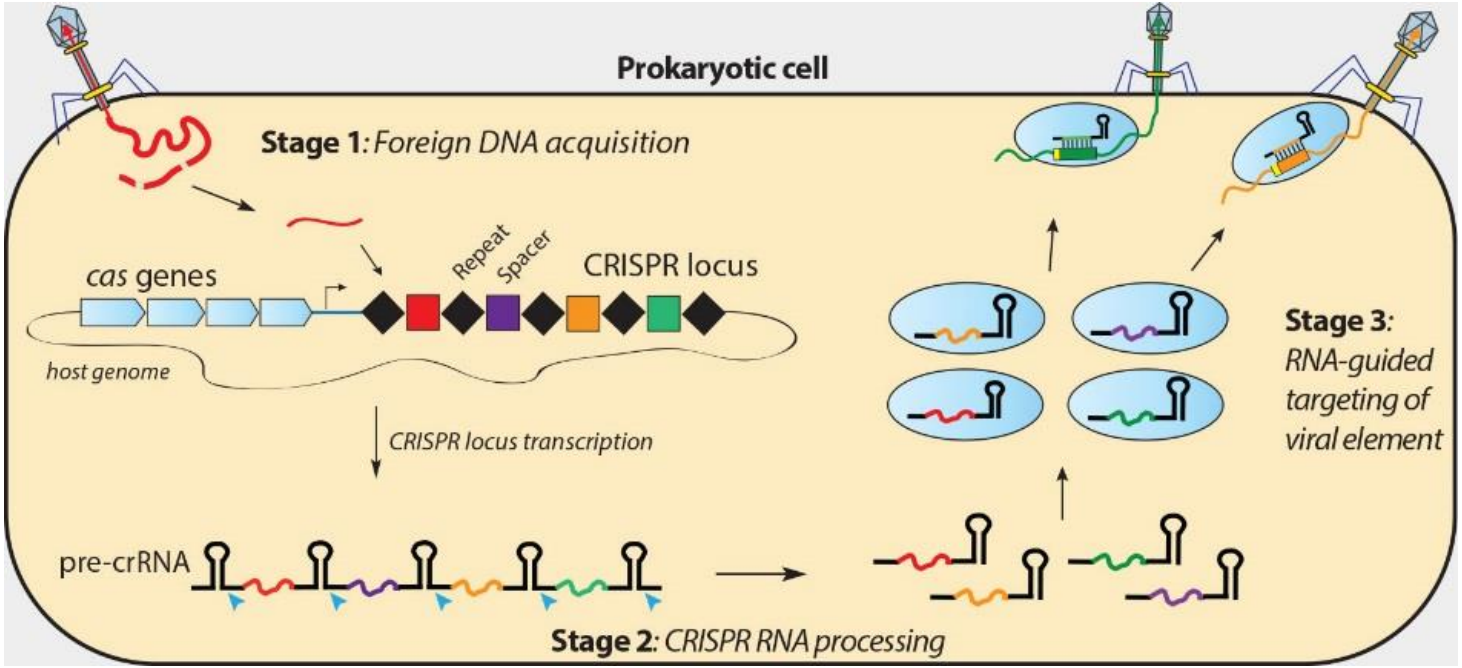
$32 Billion

Healthcare analytics market by 2022

$88 Billion

Precision medicine market by 2022

Crispr CAS9 Genome Editing



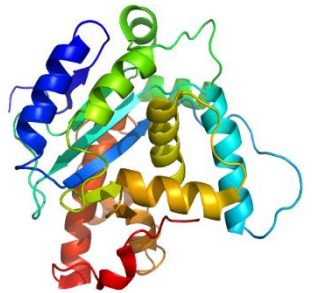
Molecular Biology Dogma

Big Data

§DNA → §*Transcription*

§RNA → §*Translation*

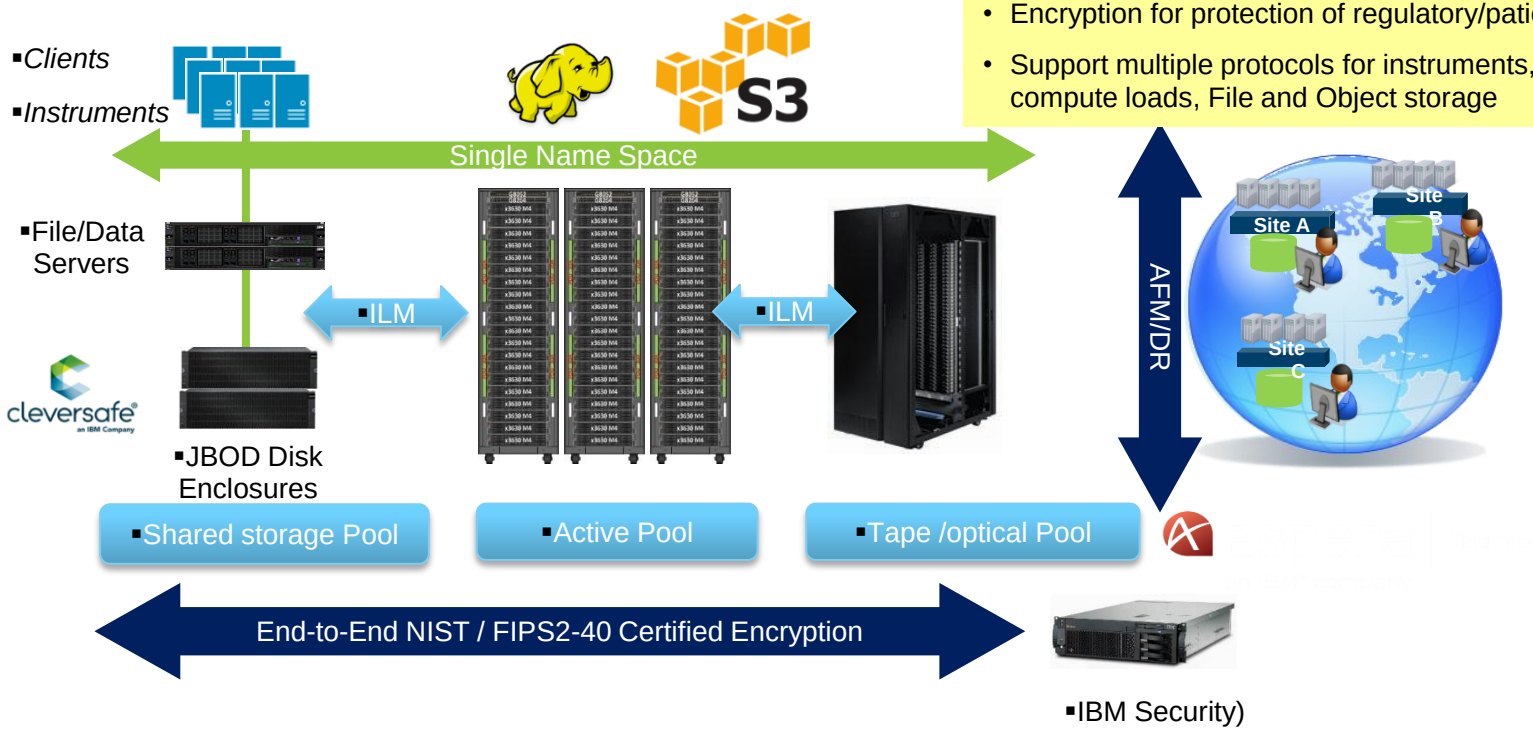
§Proteins



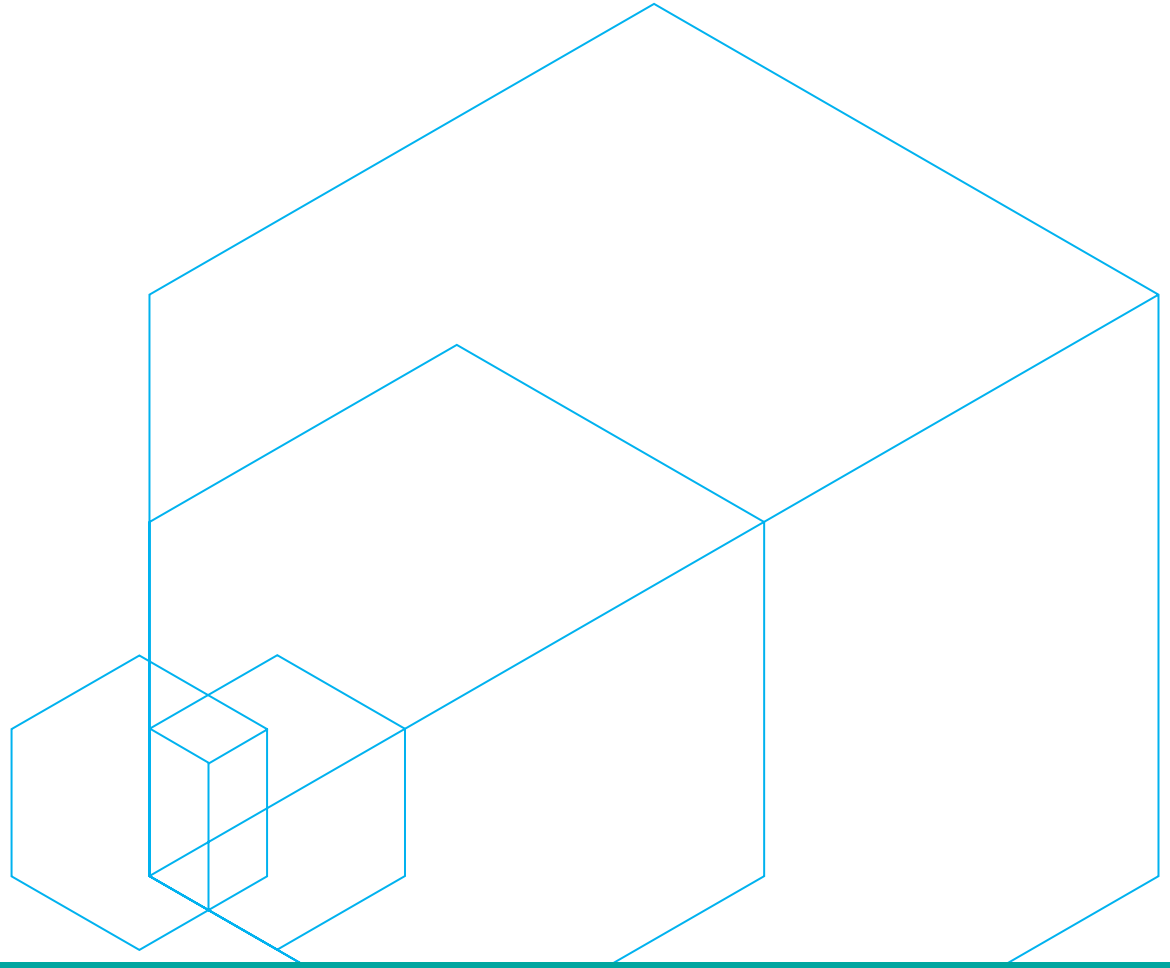
Established & proven with blue-chip organizations

Spectrum Scale is the backbone of the data hub

- Single Name Space
- Supports multiple tiers of storage: flash, spinning disk, tape and archive
- Geographically dispersed management of data including disaster recovery
- Encryption for protection of regulatory/patient data
- Support multiple protocols for instruments, compute loads, File and Object storage



Thank you.



ibm.com/systems

Legal notices

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 1 0504- 785
U.S.A.

Information and trademarks

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

DCP03213USEN-06