



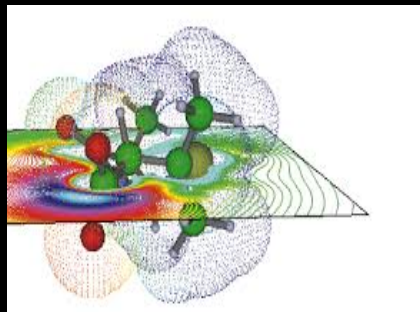
# IBM Spectrum Scale

**IBM Life Sciences**

**Madhav Ponamgi, Ph.D.**  
**mzp@us.ibm.com**

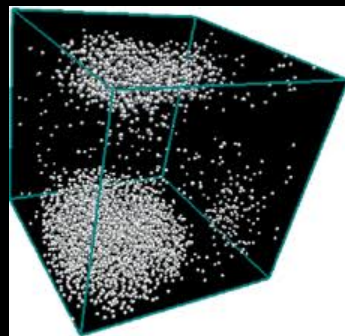


# Life Sciences



## § Computational / Quantum Chemistry

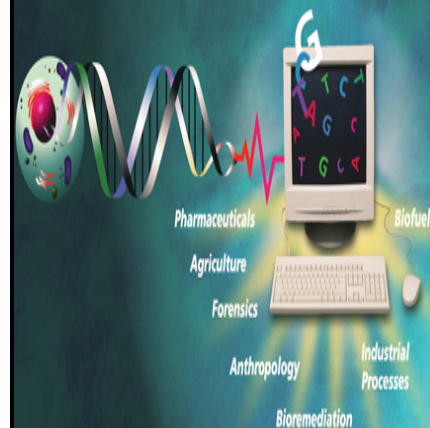
- Ab Initio / Hartree-Fock
- Gaussian
- GAMESS
- MOPAC



## § Molecular Dynamics

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

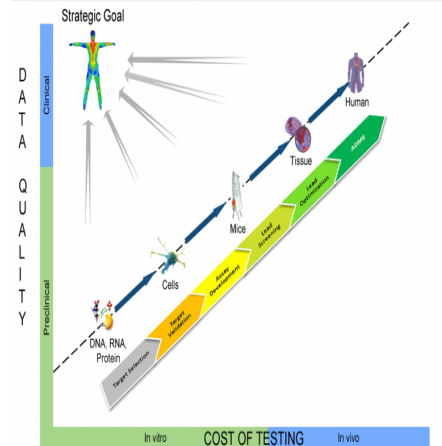
## Human Genome Project



## § Genomics / Proteomics

- Mass spectrometry
- De novo sequencing
- Sequest
- Scaffold
- Illumina / Accelrys / CLC

## IIH Bridge Yields Translational Medicine



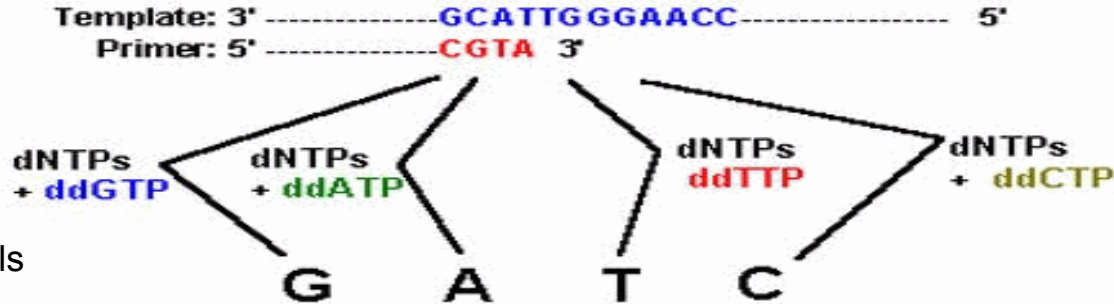
41 Improve predictive power of each phase of experimentation

## § Translational Medicine / Imaging

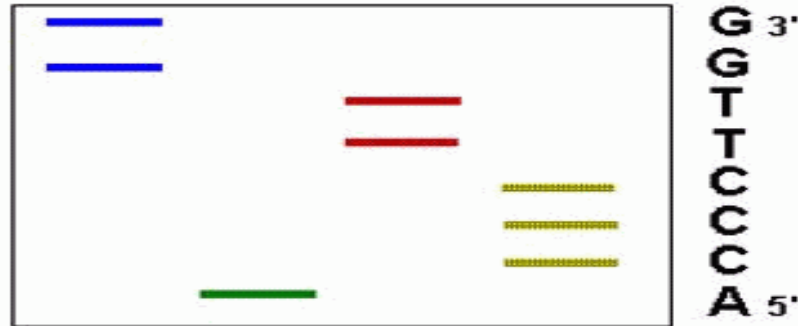
- Personalized genomics
- Health records / HIPPA
- Functional MRI
- Data mining / Security

# Traditional Sequencing Method – SANGER Method

## Sanger ddNTP Chain Termination Sequencing



- § 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 electrophoresis
  - Read back top to bottom



copyright 1996 M.W. King

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

# Next Generation Sequencers

**illumina**<sup>®</sup>



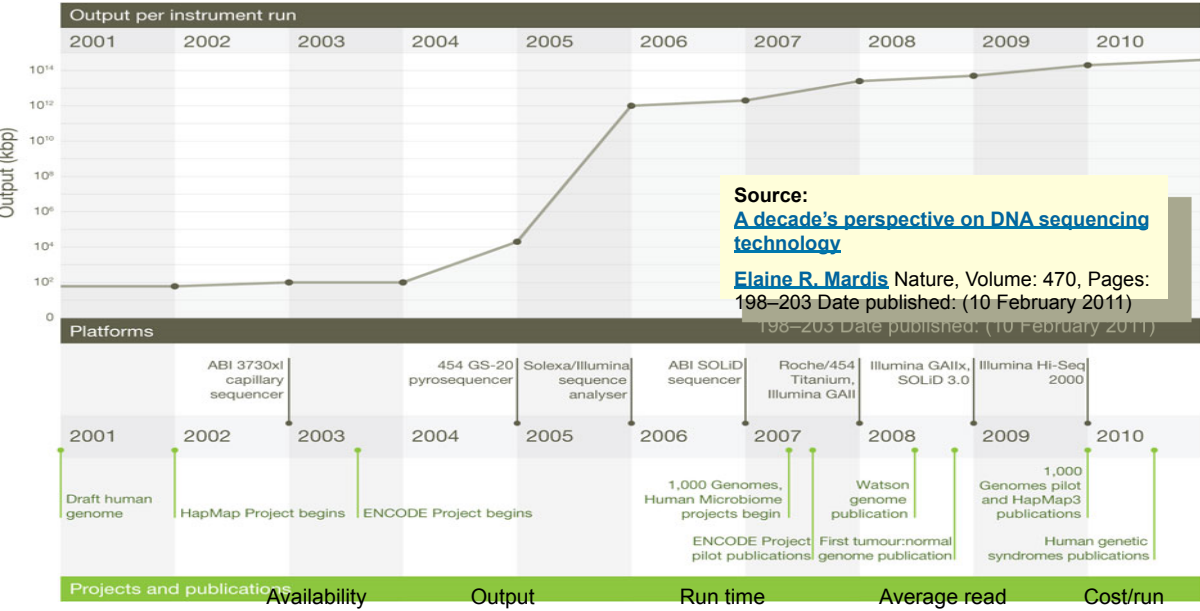
**PACIFIC  
BIOSCIENCES**<sup>®</sup>

**Oxford  
NANOPORE**<sup>™</sup>  
Technologies

**ion torrent**  
⬇ ⬆ ⬇ ⬇ ⬆ ⬆ ⬆ ⬆ ⬆ ⬆

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

# Evolution of DNA Sequencers



**Source:**  
[A decade's perspective on DNA sequencing technology](http://www.nature.com/news/A-decade's-perspective-on-DNA-sequencing-technology)  
 Elaine R. Mardis Nature, Volume: 470, Pages: 198–203 Date published: (10 February 2011)  
 198–203 Date published: (10 February 2011)



\$149,000

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



\$740,000

Illumina HiSeq 2500

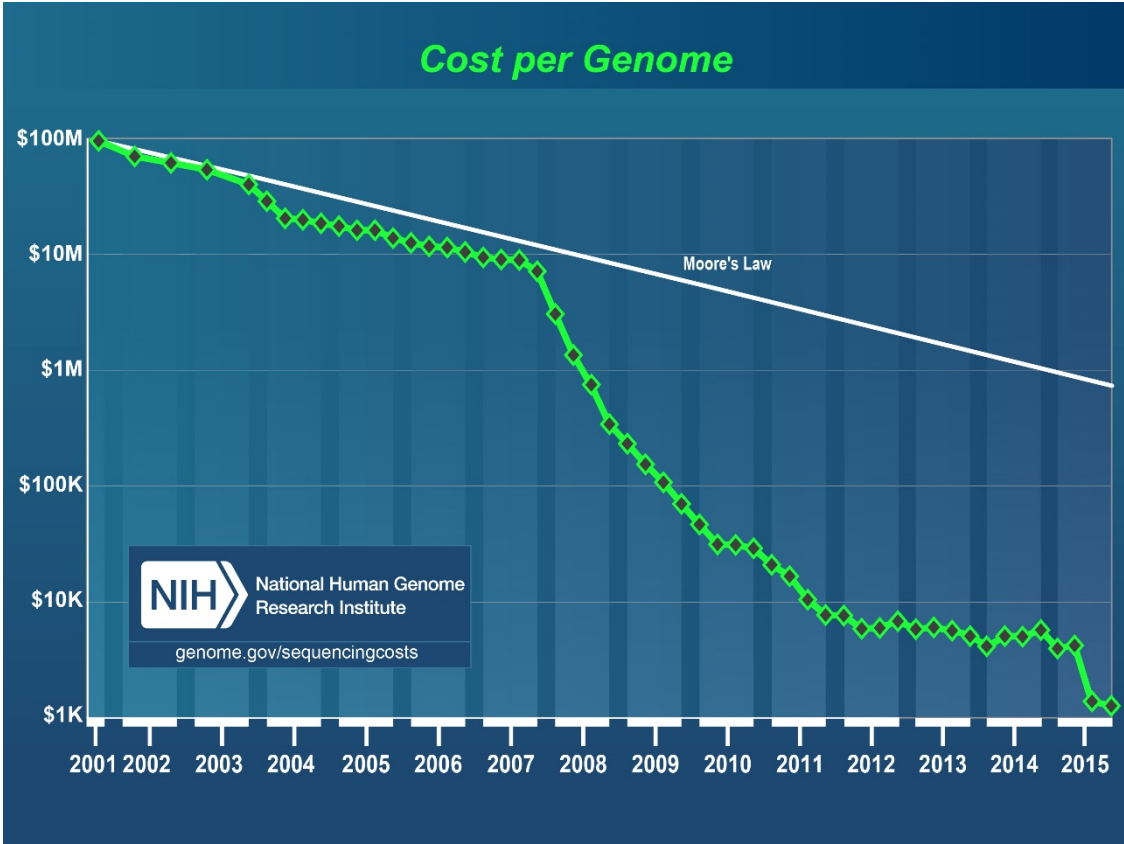
	Availability	Output	Run time	Average read	Cost/run
Proton 1	Sept 2012	10Gb	2-4 hrs	200 bp	\$999
Proton 2	Mar 2013	100Gb	2-4 hrs	>200bp	<\$999
HiSeq 2500	Mid 2012	600Gb	11 days	150bp	\$11,000

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



Oxford Nanopore MinION

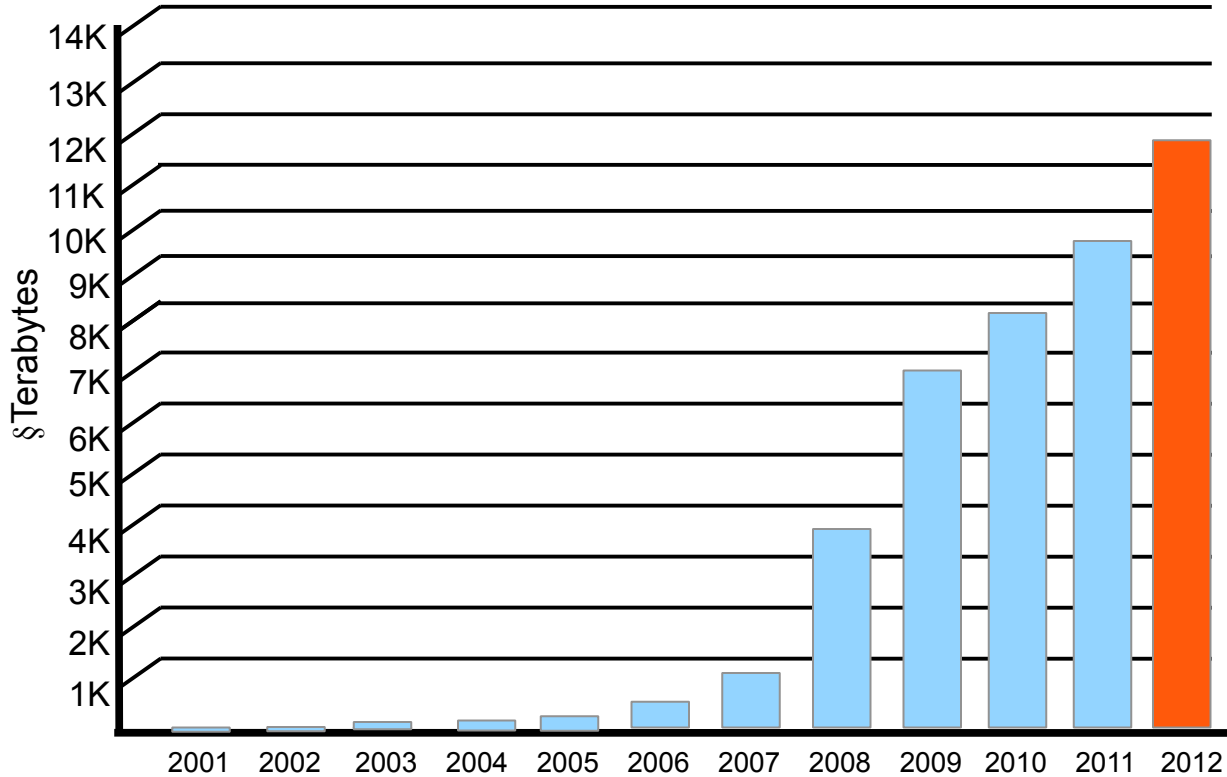
# Cost per Genome



<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" 4TB disks are a guide.. then we'd be looking at **5,000** disk farm
- Stacked horizontally, this would be a tower 147 meters in height
  - Statue of Liberty, 93 M
  - Eiffel Tower, 324 M



PHASE TWO : INTERPRETATION

SEIDMAN The Star Ledger



- 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

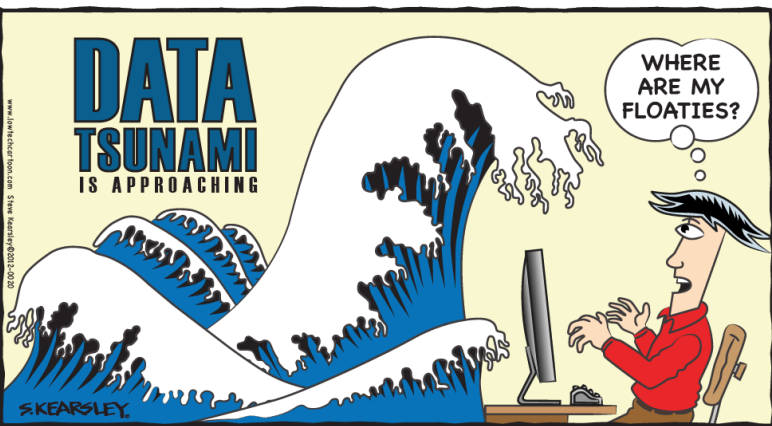
▪2 TB

§VCF file  
§Genomic Variant

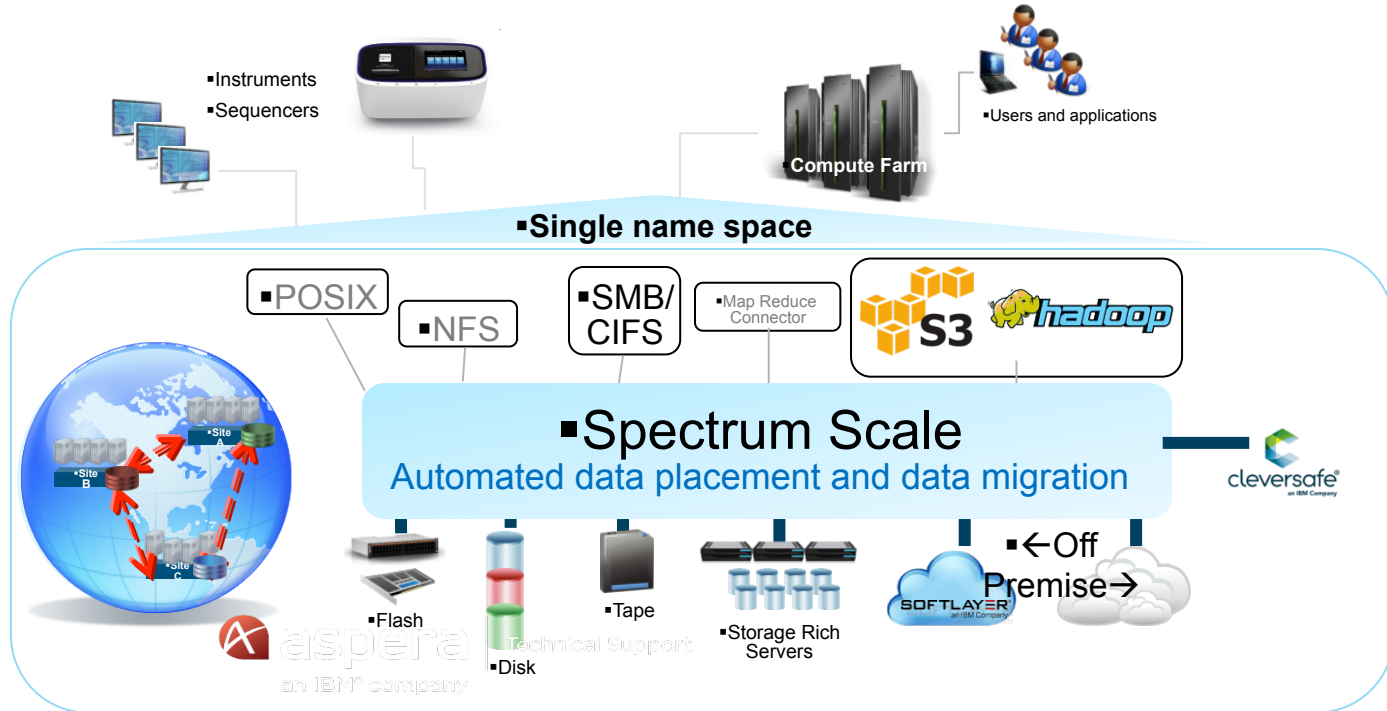
▪200 GB

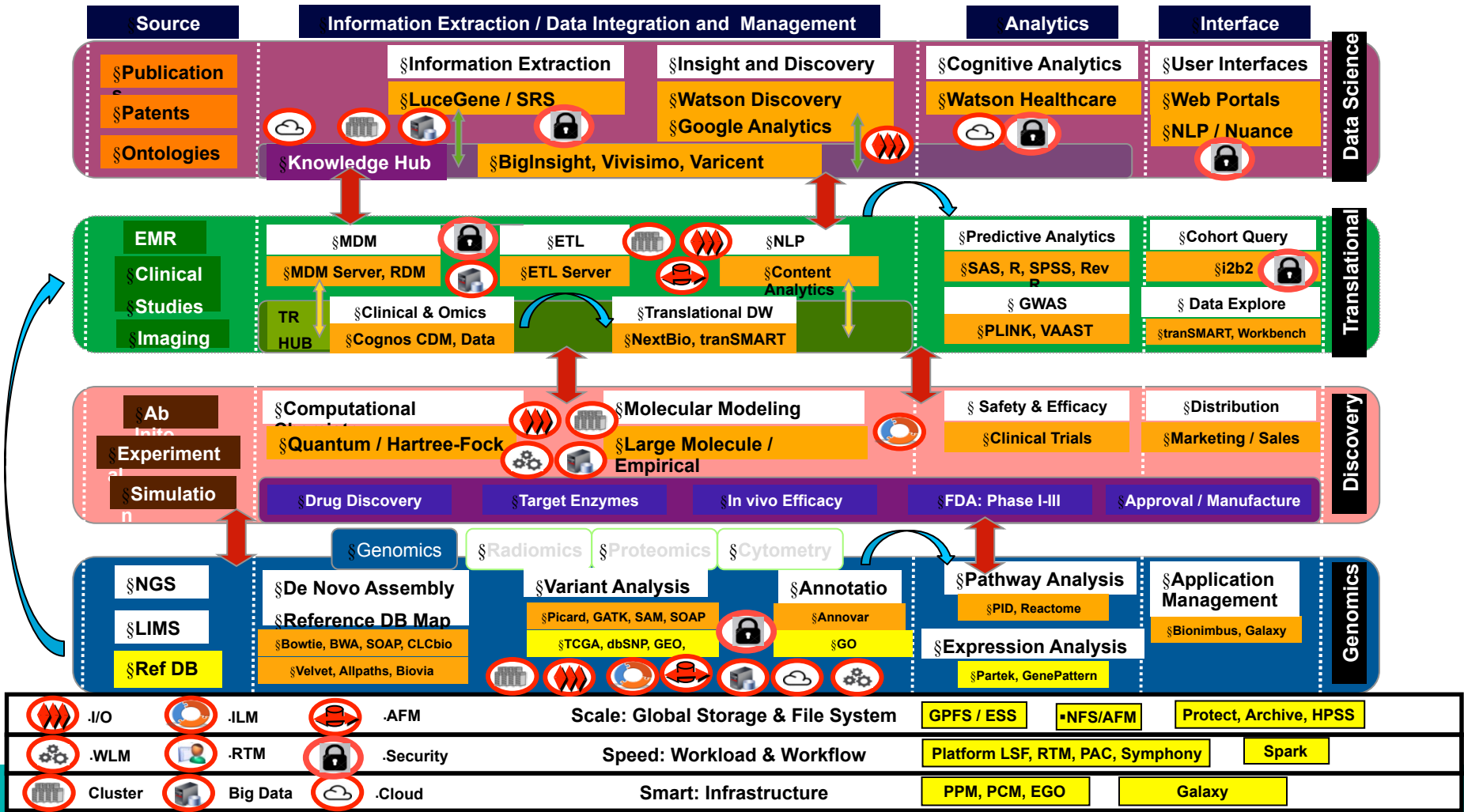


§Downstream  
Analysis

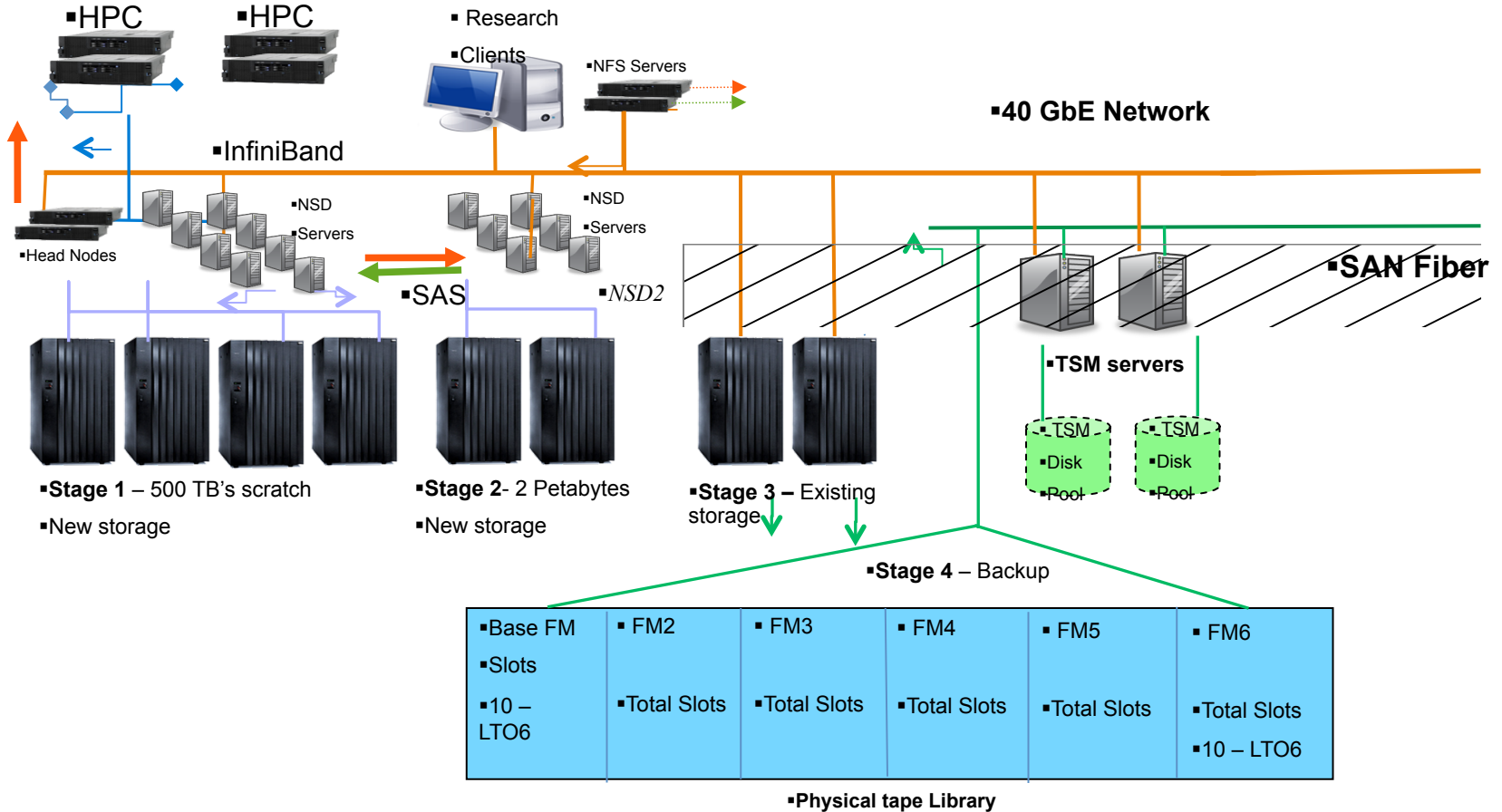


# Spectrum Scale: A File System for Genomic Pipelines

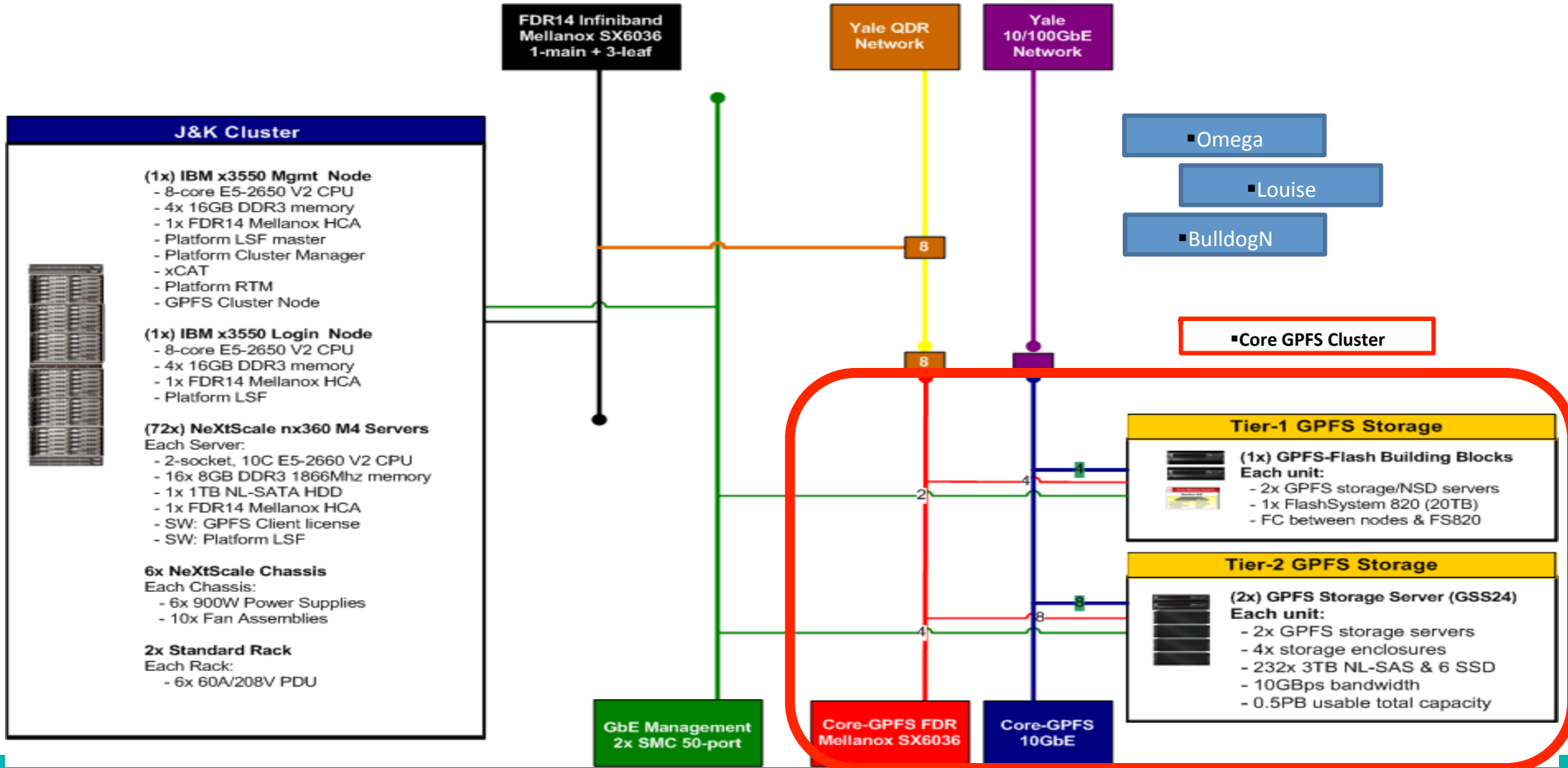




# Major Cancer Center



# Yale System Architecture



## Typical Genomics File Size Histogram

<b>scratch</b>	<b># files</b>	<b>% files</b>	<b>total size</b>	<b>% size</b>
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%
<i>Total</i>	2836692		114.47 TB	



# 3 weeks

## Diagnosis and Cure



Dr. Wartman, WashU  
Cancer researcher and  
leukemia survivor



3 weeks

Diagnosis and Cure



Dr. Wartman, WashU  
Cancer researcher and  
leukemia survivor



\$32 Billion

Healthcare analytics market by 2022

\$88 Billion

Precision medicine market by 2022



# Established & proven with blue-chip organizations

## ENTERPRISE

intuit. Capital One  
 NATIONAL INSTRUMENTS SUPERVALU  
 BARNES & NOBLE BOOKSELLERS CVS/pharmacy  
 CHOICE HOTELS INTERNATIONAL SaskTel  
 esure.com Grange Insurance  
 Walgreens  
 Kroger FedEx Express  
 WELLS FARGO Adobe  
 MITSUBISHI MOTORS CIGNA

## HEALTHCARE

WELLPOINT UnitedHealthcare  
 BlueCross BlueShield of Minnesota UPMC  
 HUMANA Guidance when you need it most Ochsner Health System  
 Sutter Health With You. For Life. BAYLOR Health Care System  
 PAML The Elliot at River's Edge  
 Banner Health MidMichigan Health  
 MEMORIALCARE Excellence in Health Care  
 sureScripts  
 Cleveland Clinic CATHOLIC MEDICAL CENTER  
 BayCare Health System Health First  
 SharedHealth Transforming Care  
 HealthFirst Corporation

## PUBLIC SECTOR

Ontario FLORIDA DEPARTMENT OF HEALTH  
 Newfoundland Labrador Manitoba  
 Government of Saskatchewan In-Q-Tel  
 New Brunswick Canada Department of Veterans Affairs  
 Capital Health COUNCIL  
 NCIS CONTACT POINT  
 because every child matters

## PARTNERS

accenture Capgemini CONSULTING. TECHNOLOGY. OUTSOURCING  
 Allscripts edico genome  
 dbMotion MEDSEEK  
 CSC EXPERIENCE. RESULTS. Infosys  
 MAXIMUS RELIABLE GOVERNMENT SERVICE FOR THE PEOPLE covisint AGFA  
 SAIC From Science to Solutions Cognizant Passion for building stronger businesses  
 Wellogic northgate HEALTHVISION Software and Solutions powered by Cloverleaf  
 transSMART FOUNDATION

# Molecular Biology Dogma



§DNA → § *Transcription*



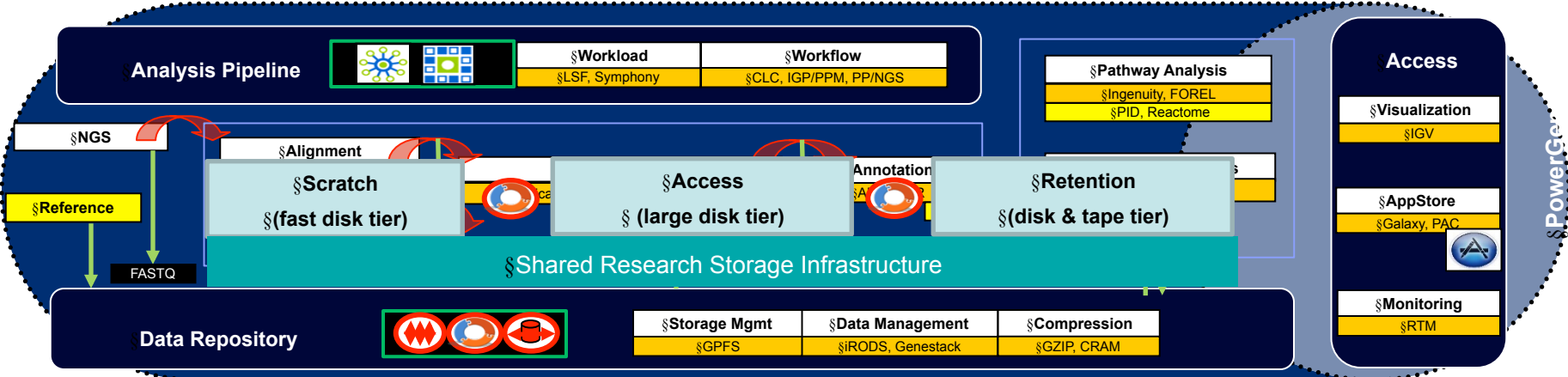
§RNA → § *Translation*



§Proteins



# § Genomics Platform Architecture



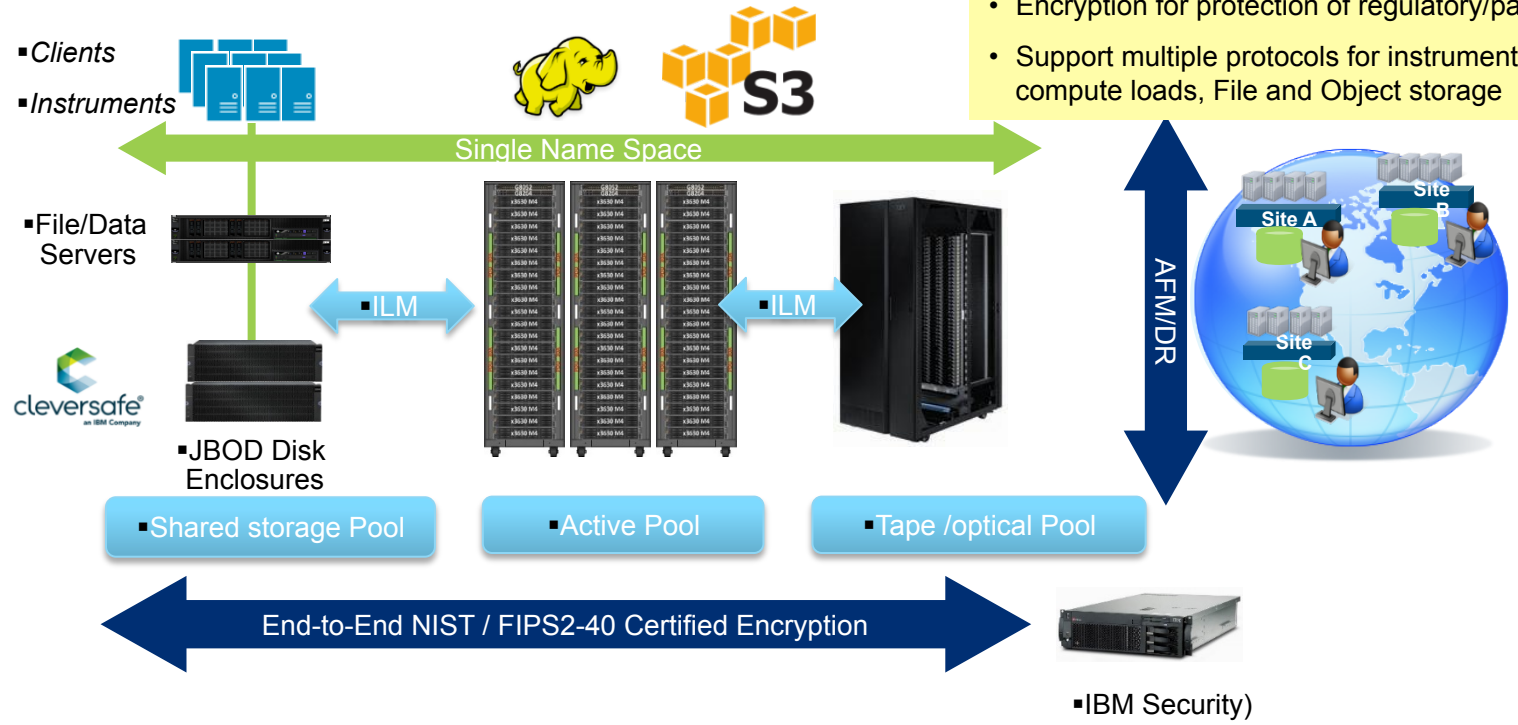
§ **Deletion:** scratch FS will prune files at certain intervals to remove aged/unused files

§ **Placement:** files can be placed automatically and optimally based on size into different pools

§ **Migration:** data in permanent FS can move from disk pool to tape/HSM pool

# Spectrum Scale is the backbone of the data hub

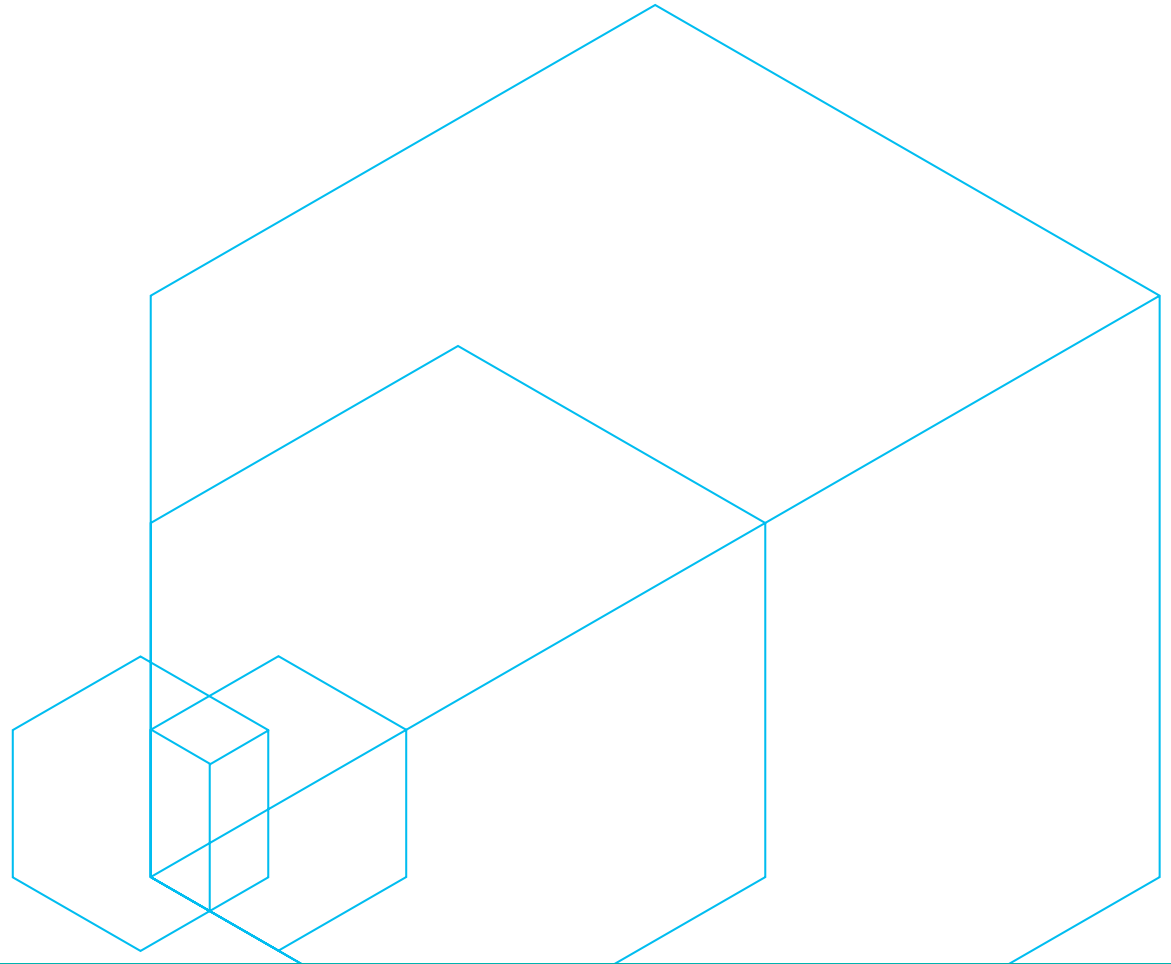
- Single Name Space
- Supports multiple tiers of storage: flash, spinning disk, tape and archive
- Geographically dispersed management of data including disaster recovery
- Encryption for protection of regulatory/patient data
- Support multiple protocols for instruments, compute loads, File and Object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



**Thank you.**



[ibm.com/systems](https://ibm.com/systems)





# Legal notices

Copyright © 2015 by International Business Machines Corporation. All rights reserved.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

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 intellectual 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

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at <http://www.ibm.com/legal/copytrade.shtml>

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

ITIL is a Registered Trade Mark of AXELOS Limited.

UNIX is a registered trademark of The Open Group in the United States and other countries.

\* All other products may be trademarks or registered trademarks of their respective companies.

## 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.