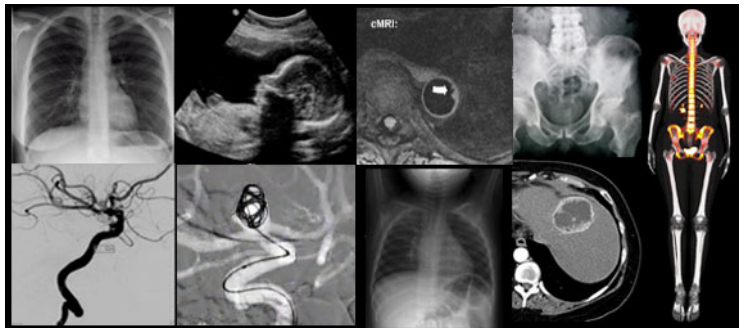
Data management challenges in todays Healthcare and Life Sciences ecosystems

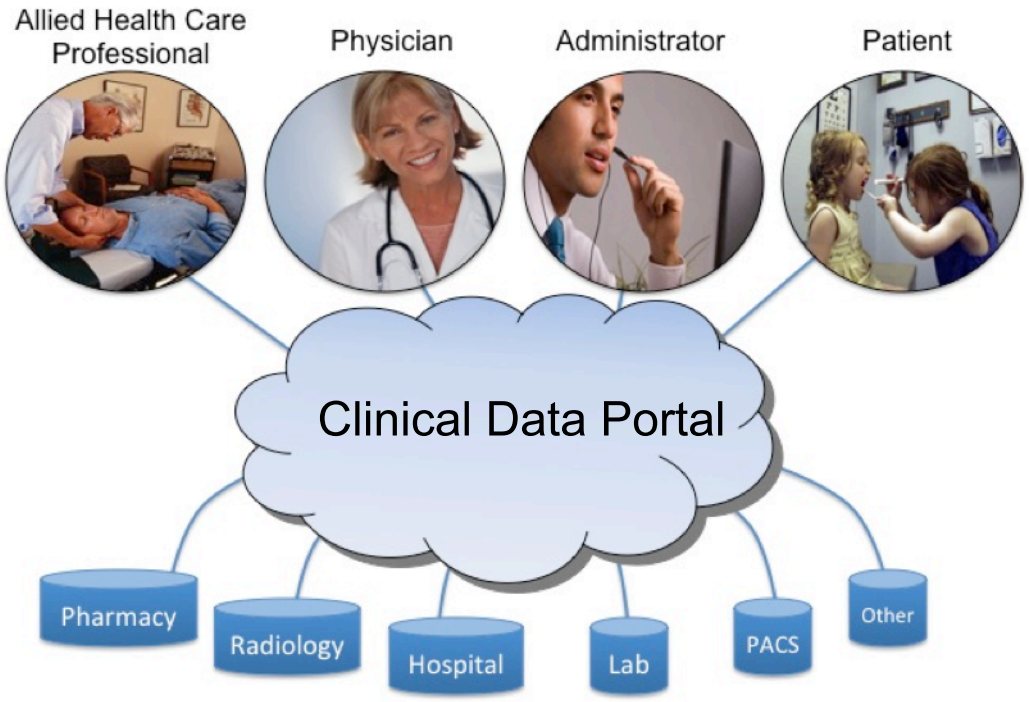
Jose L. Alvarez
Principal Engineer, WW Director Life Sciences
jose.alvarez@seagate.com

Evolution of Data Sets in Healthcare – RIS Example (1995!)

- Specialized Medicine data silos
- Multiple modalities yield separate DICOM images repositories
- Minimal Data distribution
- HL7 transfer for Demographics from HIS System
- No single data repository - VNA



Data Aggregation from multiple clinical disciplines



- Heavy HL7 interfaces to pass data between systems.
- Separate metadata sources that eventually yield an aggregate data set.
- Beginnings of Clinical portals with some decision support.

There are integration engines but data management issues are still present…

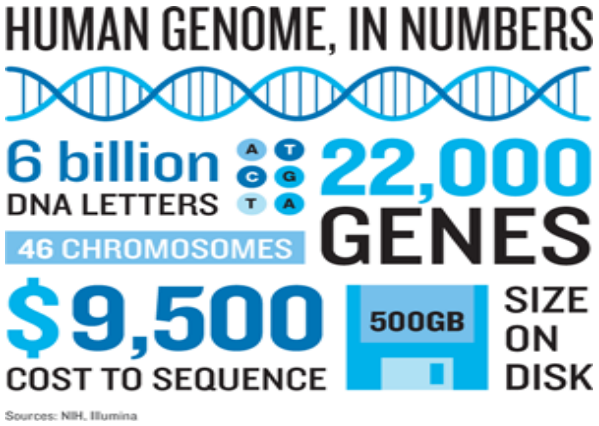
- Data integrity across all sources
- Chain of trust
- Security
- Data encryption on transit and at rest
- Data durability and longevity

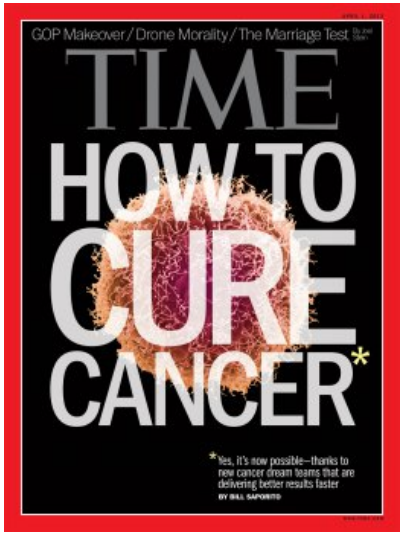
The New paradigm of the connected individual

- Distributed data sources
- Always on mode – instant gratification
- Challenge for data storage and curating
- Effective Biometrics for meaningful patient history
- Chain of trust
- Security
- Long term data preservation

Why Genomics data sets are important and disruptive

$4.27b by 2017!





MIT Technology Review

50 SMARTEST COMPANIES

#1 illumina

Seagate Cloud Systems and Solutions Portfolio

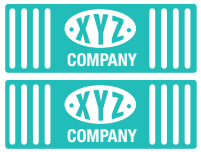
HPC

- Engineered to optimize capacity and performance
- 40% fewer racks require
- 1TB/s + file system performance

Scale-Out Systems

- Engineered solutions for object storage
- Validated architectures for open source and software-defined storage
- Private cloud appliances for backup and recovery
- Modular, scalable components for DIY customers

OEM



- 2+ million enclosures
- 17+ Exabytes shipped
- Drive Variety (HDD, SAS, SATA, SSD, hybrid)
- Enclosures, controllers
- Customer-driven partnership
- Services: Logistics, fulfillment, warranty, design, supply chain

Cloud Services

- Backup as a service
- Disaster recovery as a Service
- Archive as a service
- Endpoint backup
- Managed services

Seagate

Our Technology Investment Roadmap

Lowest Cost/
TB Storage

Increasing Total
Performance

Flash
Controllers

Lowering System
Power and Cooling Costs

Lowering System
Volume

Self-Healing
Components and Systems

Lowest Total System Lifecycle Cost
(Deployment, Operation, Disposal)

Seagate HCLS Mission/Vision

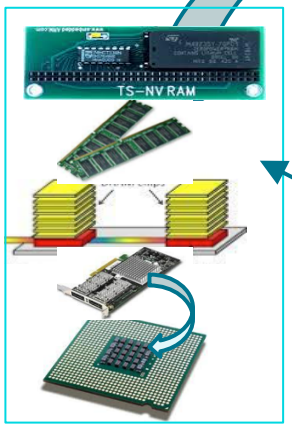
Seagate® brings an open approach to Intelligent Information Infrastructure™ accelerating bioinformatics pipelines and helping manage next-generation workloads—with scale, performance, security and cost aligned to today's challenges.

Our multi-tiered data storage solutions enable high-throughput, scalable geo-distributed storage, while meeting the complex compliance and data management challenges of high performance computing in bioinformatics.

Simplifying Scientific Workflows

Using new memory and storage tiers

Memory Tiers:
3D stacked, NVRAM,PCM, etc.
Close if not tied to the chip

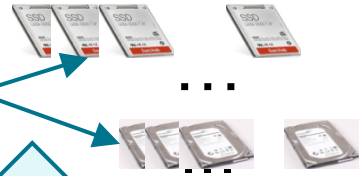


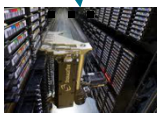
Interconnect on chip

In-Transit Analysis

In-Situ analysis

Data moves back and forth between tiers





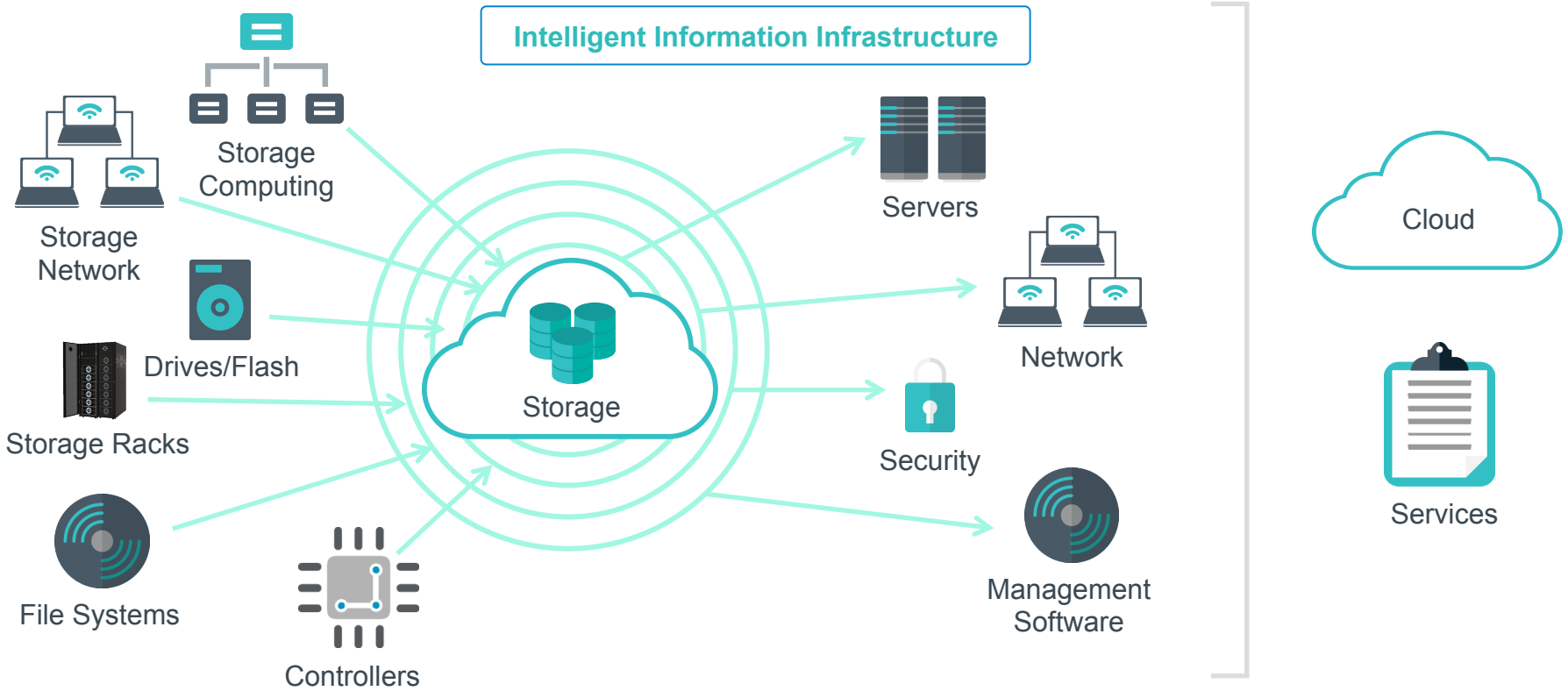
Cold Storage:
Tape or Disk

- Science is not completed in totally separate stages anymore.
- But is rather a flowing cycle.
- Work and reorganization is done **between** physical media.
- Data on physical media is analyzed, reduced, and manipulated.

Seagate

Intelligent Information Infrastructure - Powered by Seagate

Delivering with PARTNERS' components through complete systems and into the cloud



Building a Secure High Performance Data Fabric

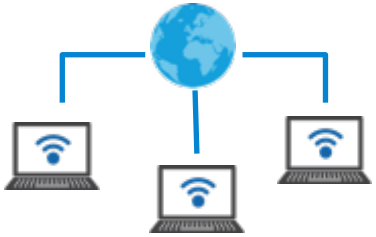
Acquire

Process / Analyze

Intelligent Information Infrastructure

Store

Archive



Distribute

Accelerating Customer Workflows

Acquire, Analyze, Store, Distribute and Archive.

The Big Data Transition in Genomics

	2010	2015
Objective:	Research	Clinical
Time to sequence	1 week	10 human genomes per day
Cost to sequence	$100K	$1000
Applications:		
POSIX	Casava	Casava, GATK (Broad), SOAPdenovo (BGI)
Analytics	None	SAP Hana – Drug discovery Hadoop, Others.
Cloud	None	Instrument to Illumina BaseSpace, Amazon, Google
Leaders – Instruments	Life Tech, illumina, 454	illumina, Life Tech, PacBio
Leaders – Compute	Dell, HP and IBM	BGI (1PF), Google (600K cores), Amazon
Leaders – Storage	CIFS, NFS, GPFS, Lustre	NFS,CIFS, GPFS, Lustre, Object Stores

DIAGNOMICS

illumina®

ign Illumina Genome Network

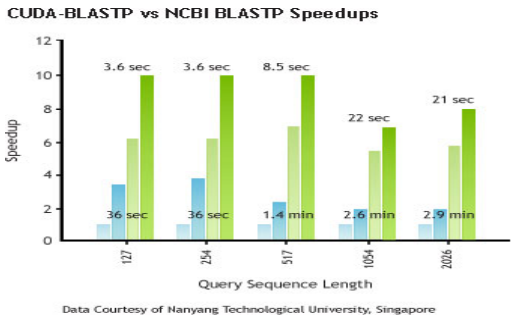
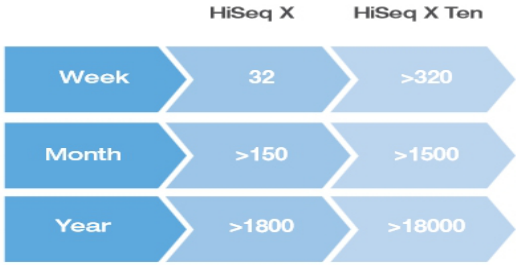
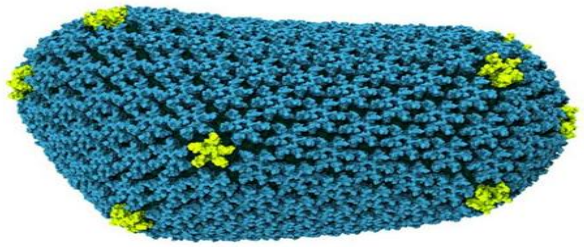


Figure 2: The HiSeq X Ten Enables Sequencing of Tens of Thousands of Human Genomes per Year



Life Science Solutions : At Scale!

Until the arrival of petascale supercomputers, no one could piece together the entire HIV capsid – an assemblage of more than 1,300 identical proteins – in atomic-level detail. The simulations that added the missing pieces to the puzzle were conducted during testing of Blue Waters, a new supercomputer at the National Center for Supercomputing Applications at the University of Illinois."



Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics
Nature 497, 643–646 (30 May 2013) doi:10.1038/nature12162
Received 02 November 2012 Accepted 05 April 2013 Published online 29 May 2013

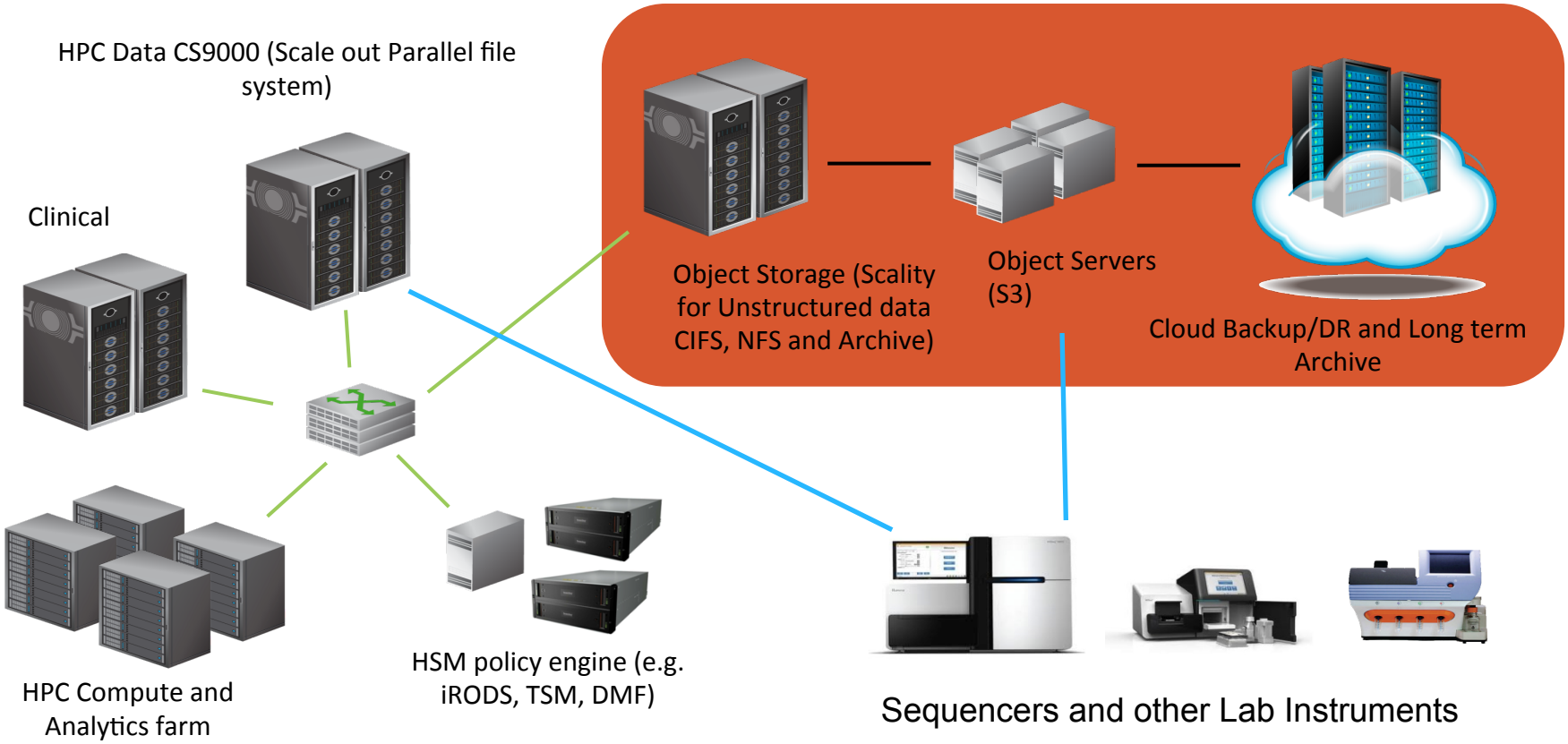
NCSA BLUE WATERS SUSTAINED PETASCALE COMPUTING

>1TB/second Aggregate Bandwidth	Competitive Storage System	Cray Sonexion Storage System	Delta
Number of Racks:	55	36	-34%
Square Footage:	985 ft²	644 ft²	-34%
Hard Drives:	46,200	17,280	-62%
Power:	~0.859MW	~0.534MW	-38%
Heat Dissipation (BTUs):	2,728,000	1,165,600	-53%

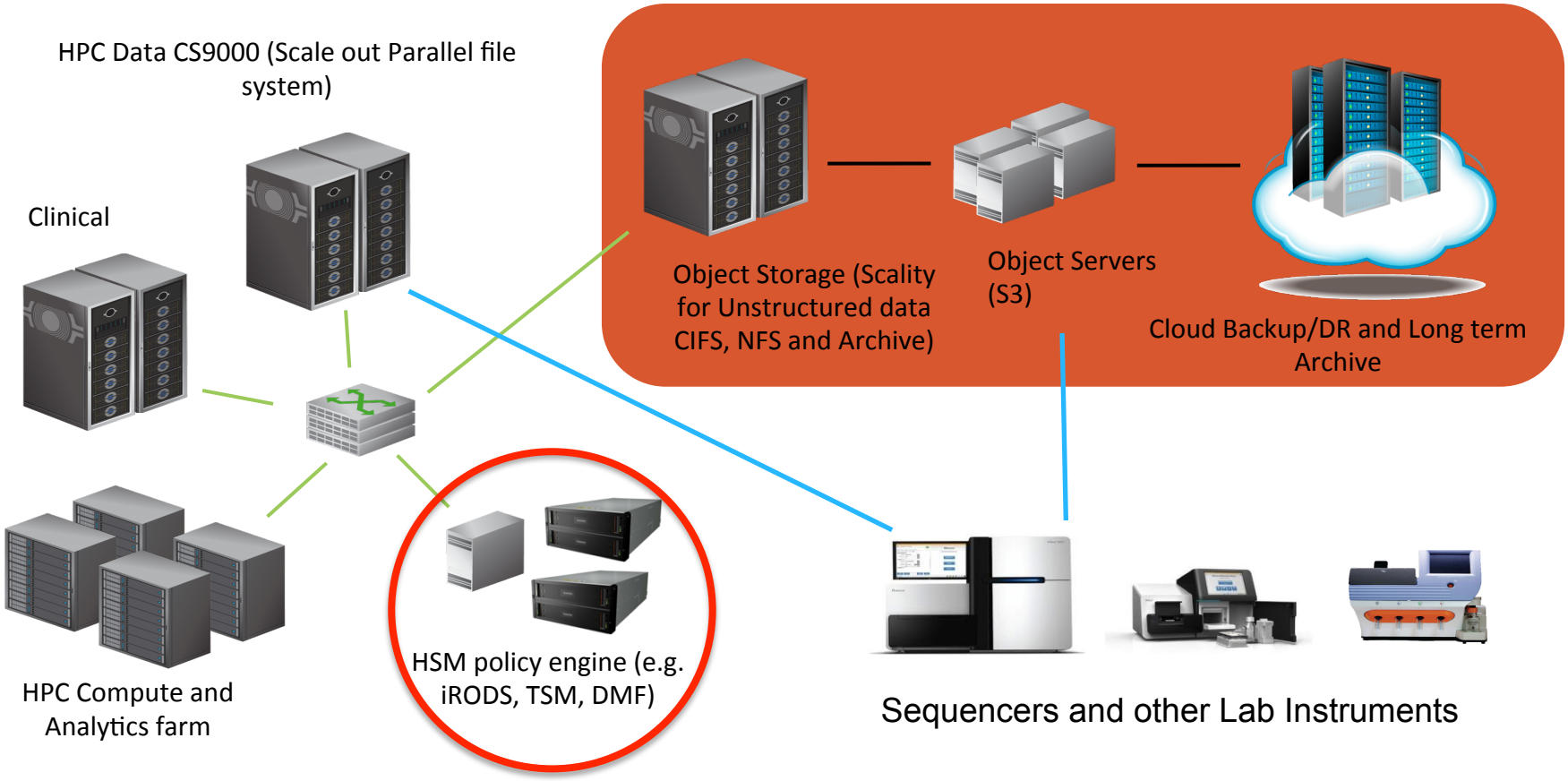
Exponentially less cost, space, cooling and power – less is more!

Seagate

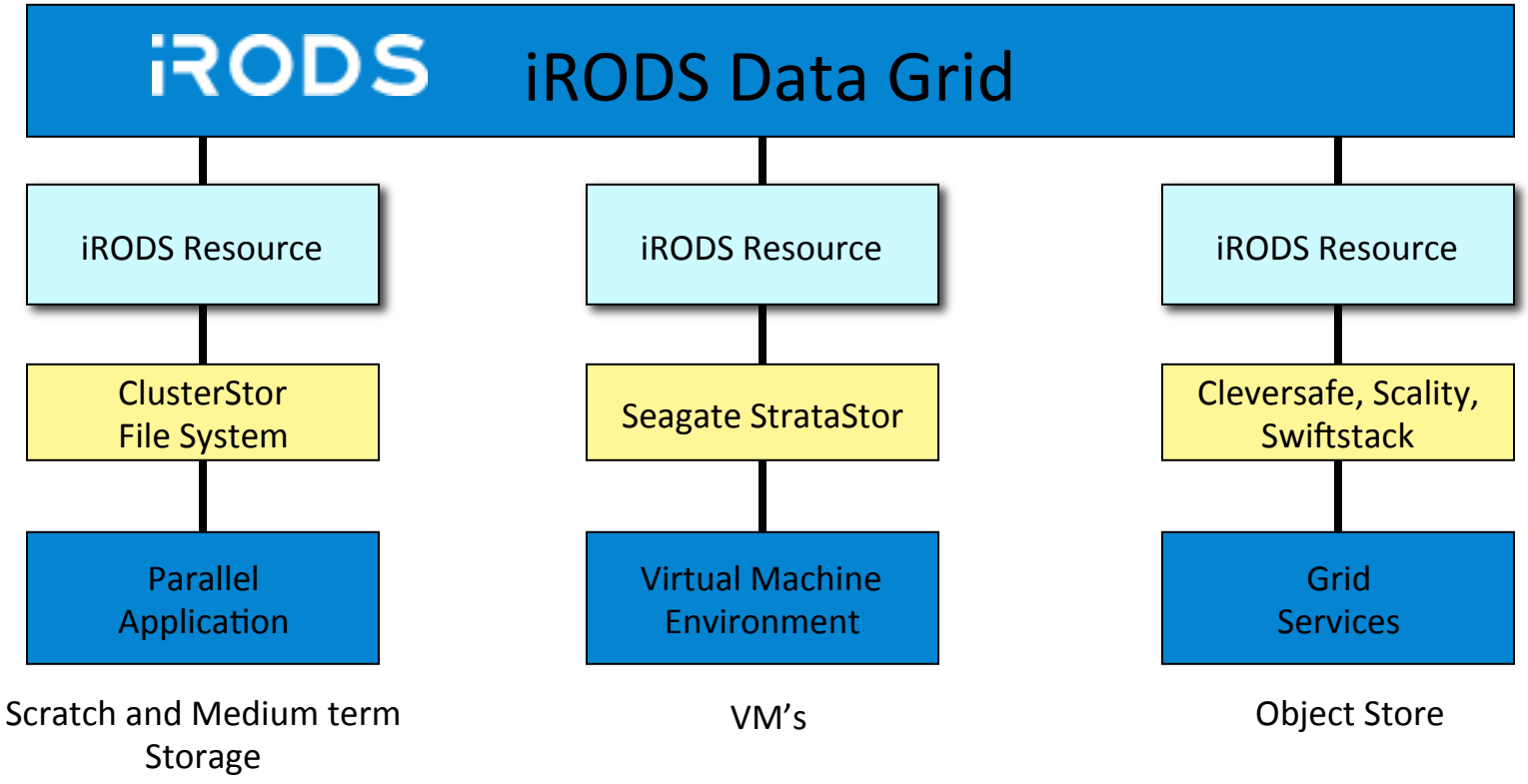
The converged Infrastructure Solution



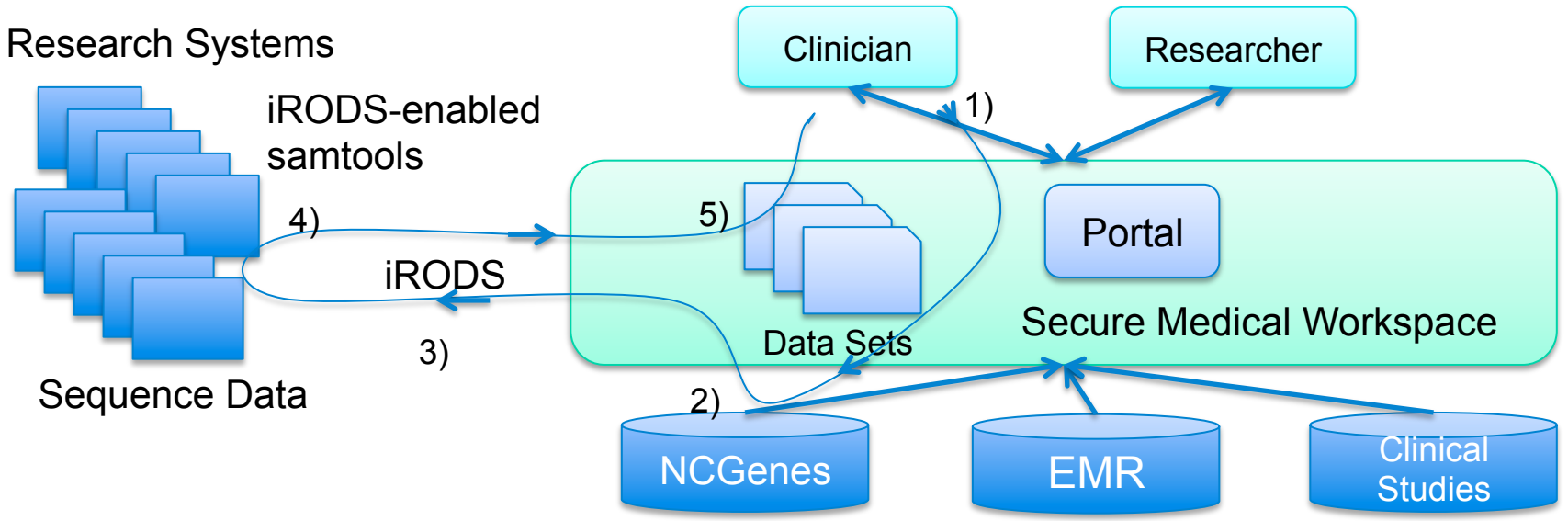
The converged Infrastructure Solution



Integration across Parallel File System and Grid Data services

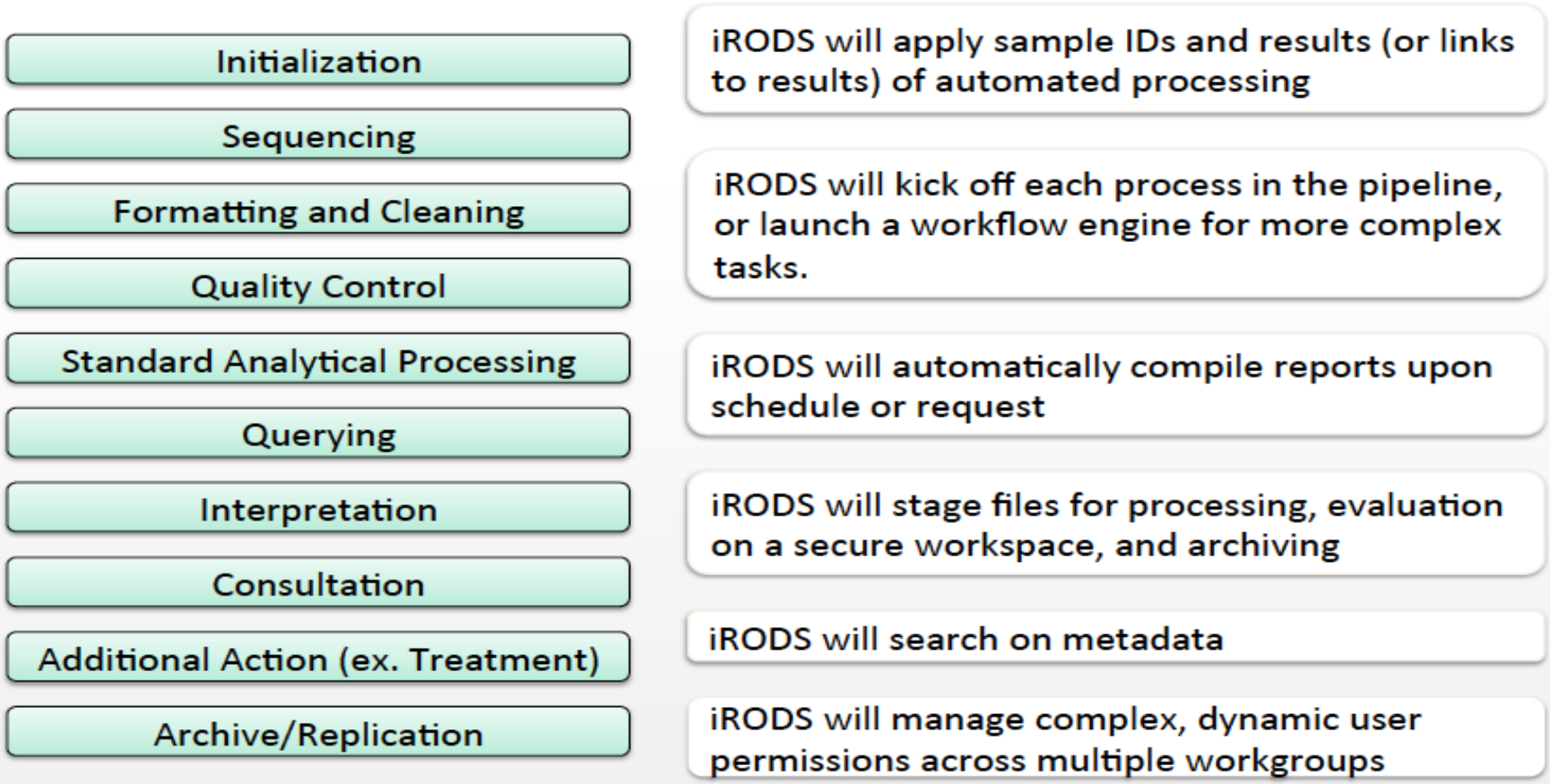


Secure Access to Data on the Clinical Side



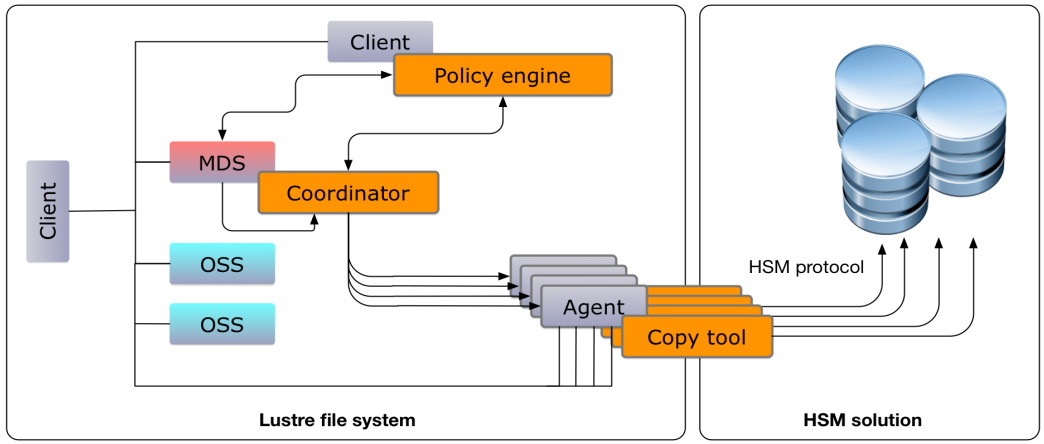
1) Clinician request for sequence reads on patient X
2) Patient id lookup to obtain subject id
3) Subject id lookup in iRODS
4) Data sets packaged in zip file and retrieved
5) Data unzipped and displayed within secure workspace

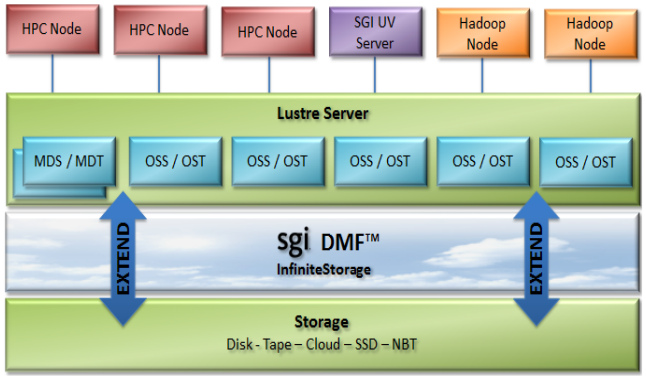
NGS Reference Implementation

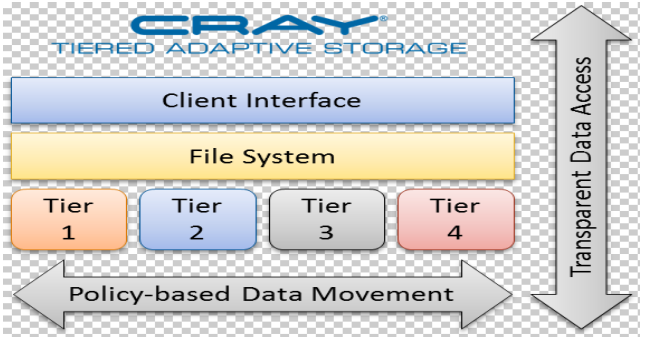


ClusterStor HSM Partners – iRODS, Cray TAS and SGI DMF

- ClusterStor v2.0 or greater w/Lustre 2.5 is HSM Ready
  - Requires HSM application and support from a partner
- HSM partner options
  - SGI DMF
  - Cray TAS
  - iRODS

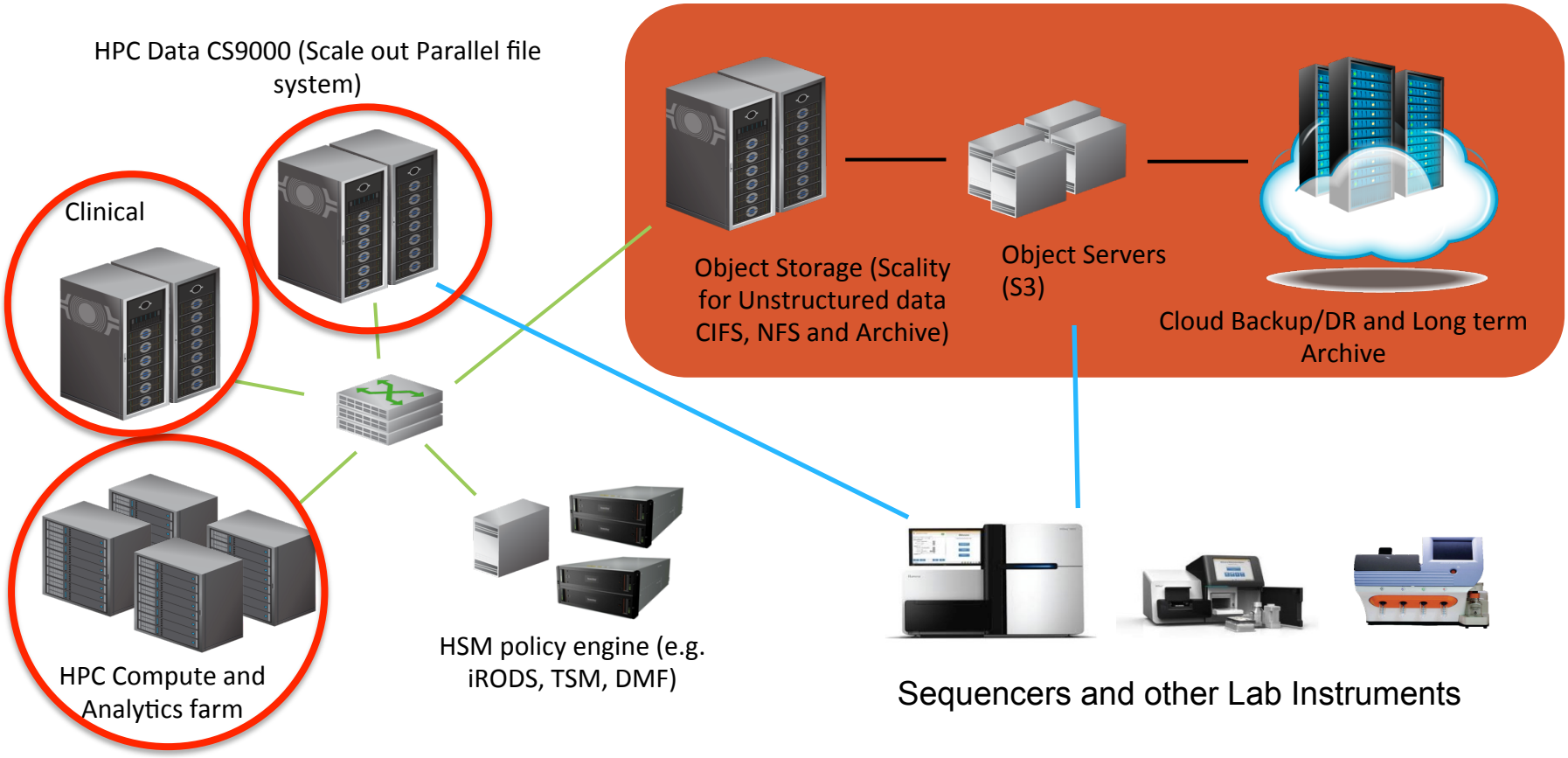




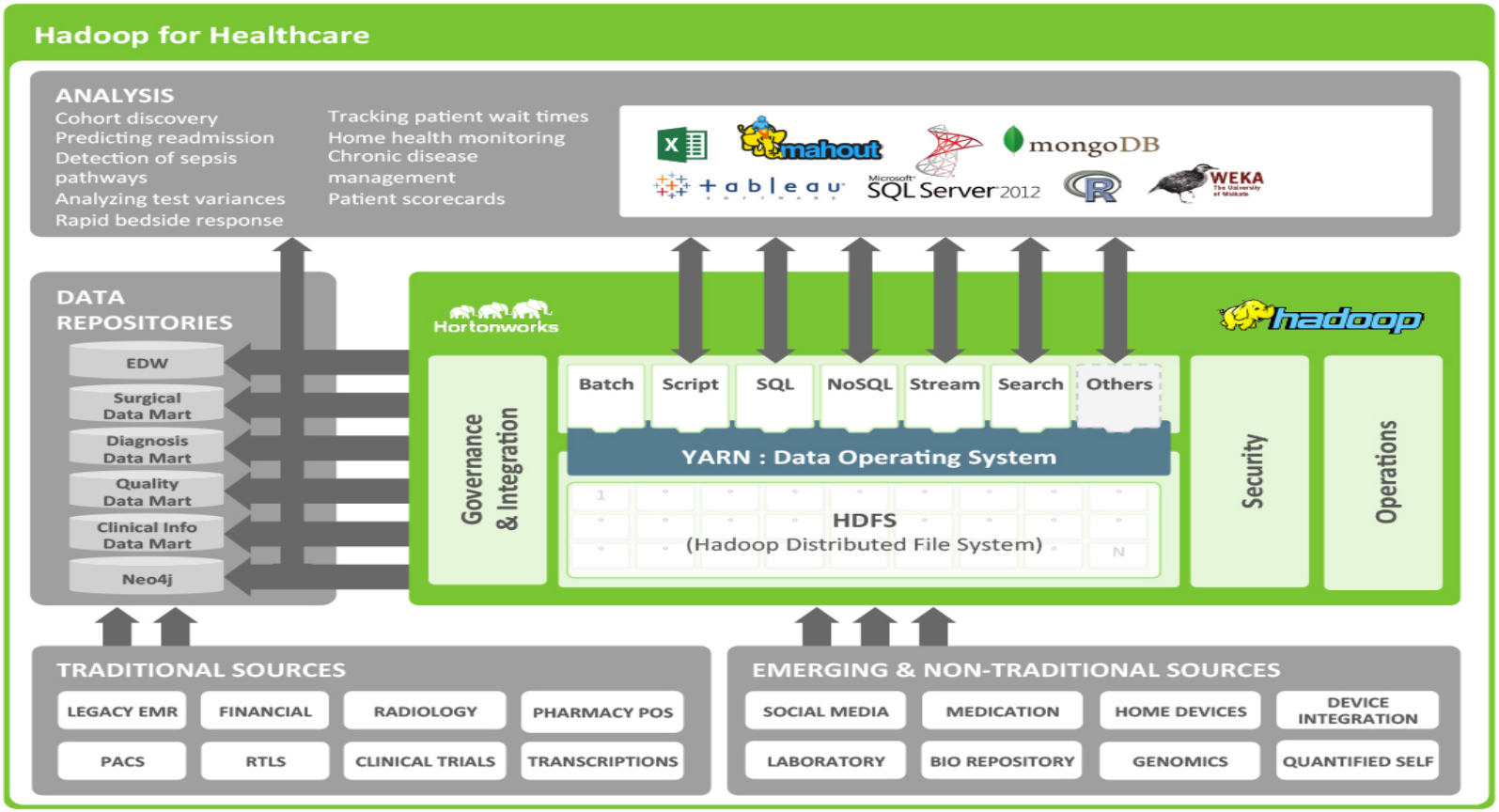


Seagate

The converged Infrastructure Solution



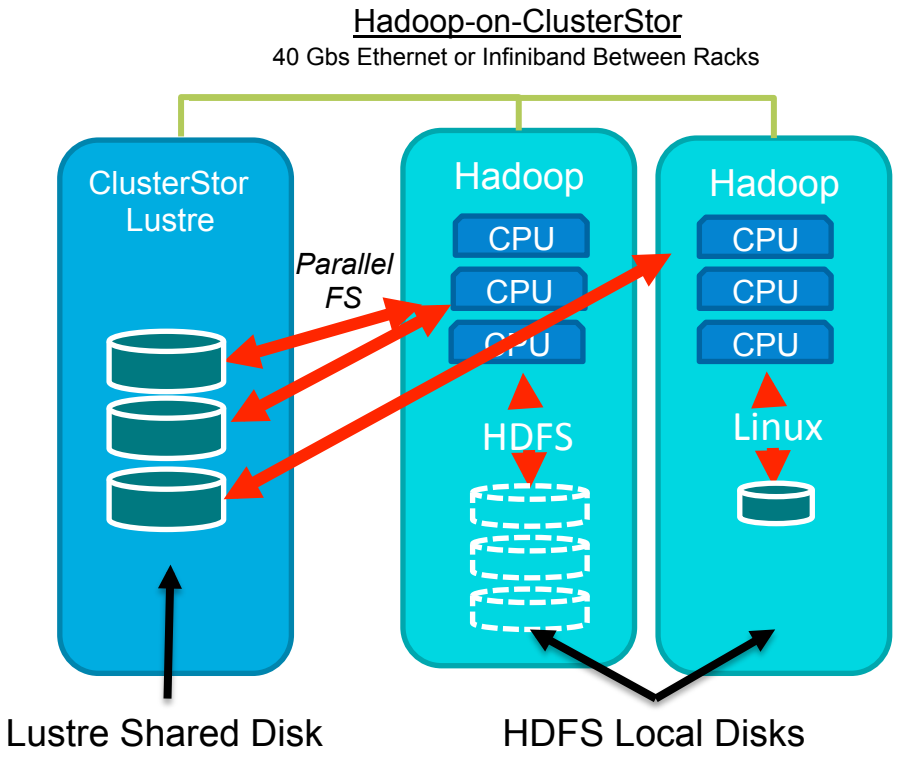
New predictive data Analytics architectures



Hadoop Workflow Accelerator on ClusterStor

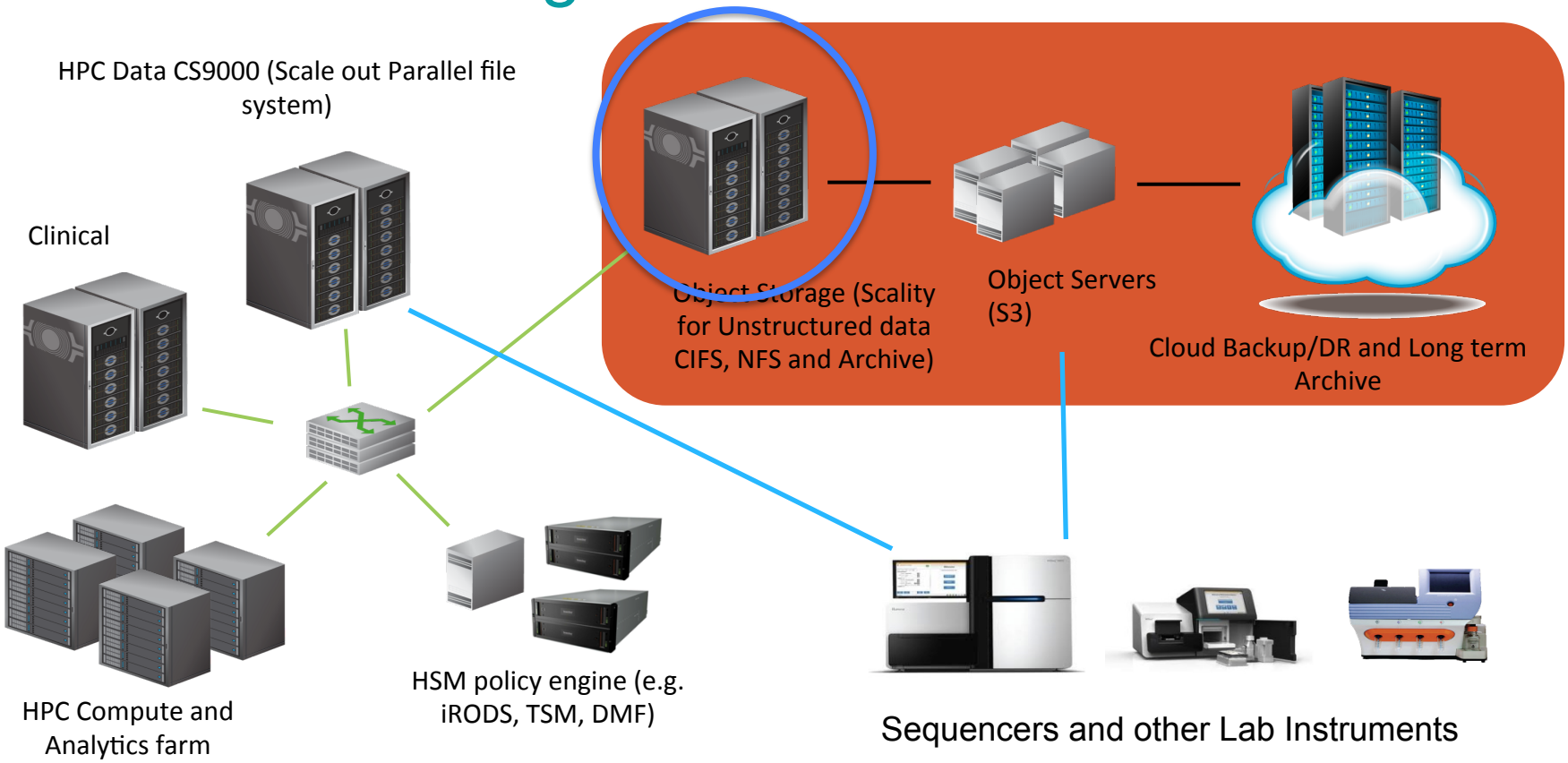
Overview

- Hadoop compute uses the ClusterStor storage system for primary input and output operations.
- MapReduce temporary files are saved using minimal local storage or can be configured to save to ClusterStor eliminating all local storage.
- The Hadoop Workflow Accelerator eliminates the reliance on direct attached storage, allowing for truly diskless compute nodes and independent scaling of compute and storage
- Hadoop Workflow Accelerator optimizes Hadoop performance for you workloads and applications while improving TCO.



Seagate

The converged Infrastructure Solution

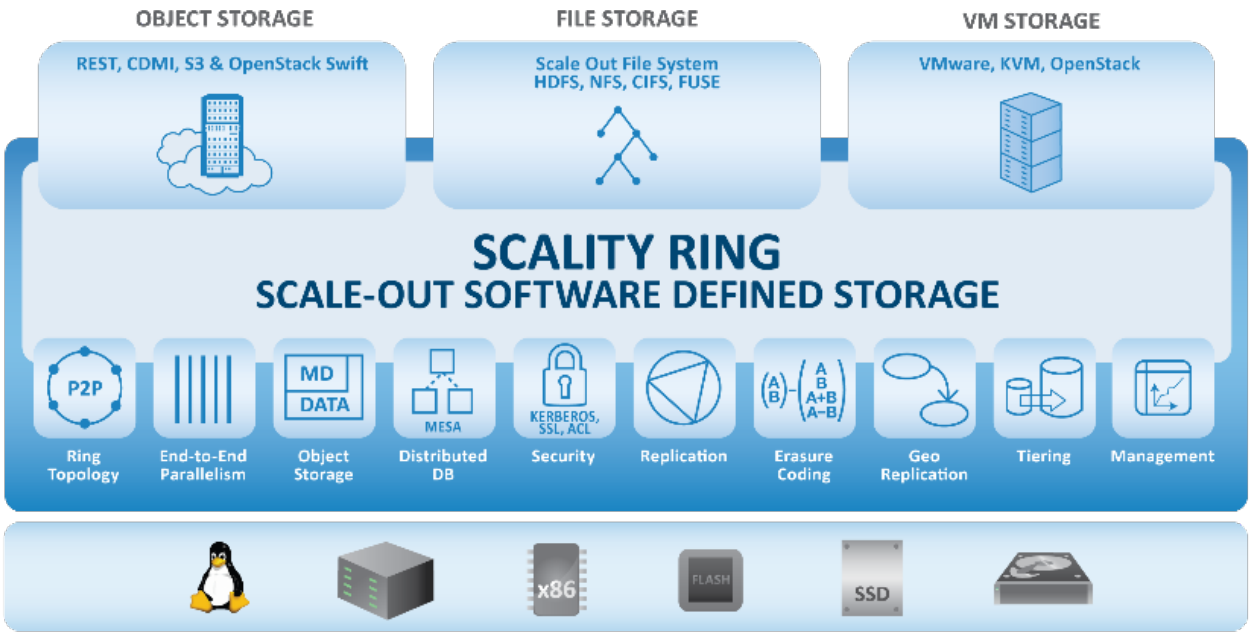


Software Defined CSS Validated Architectures

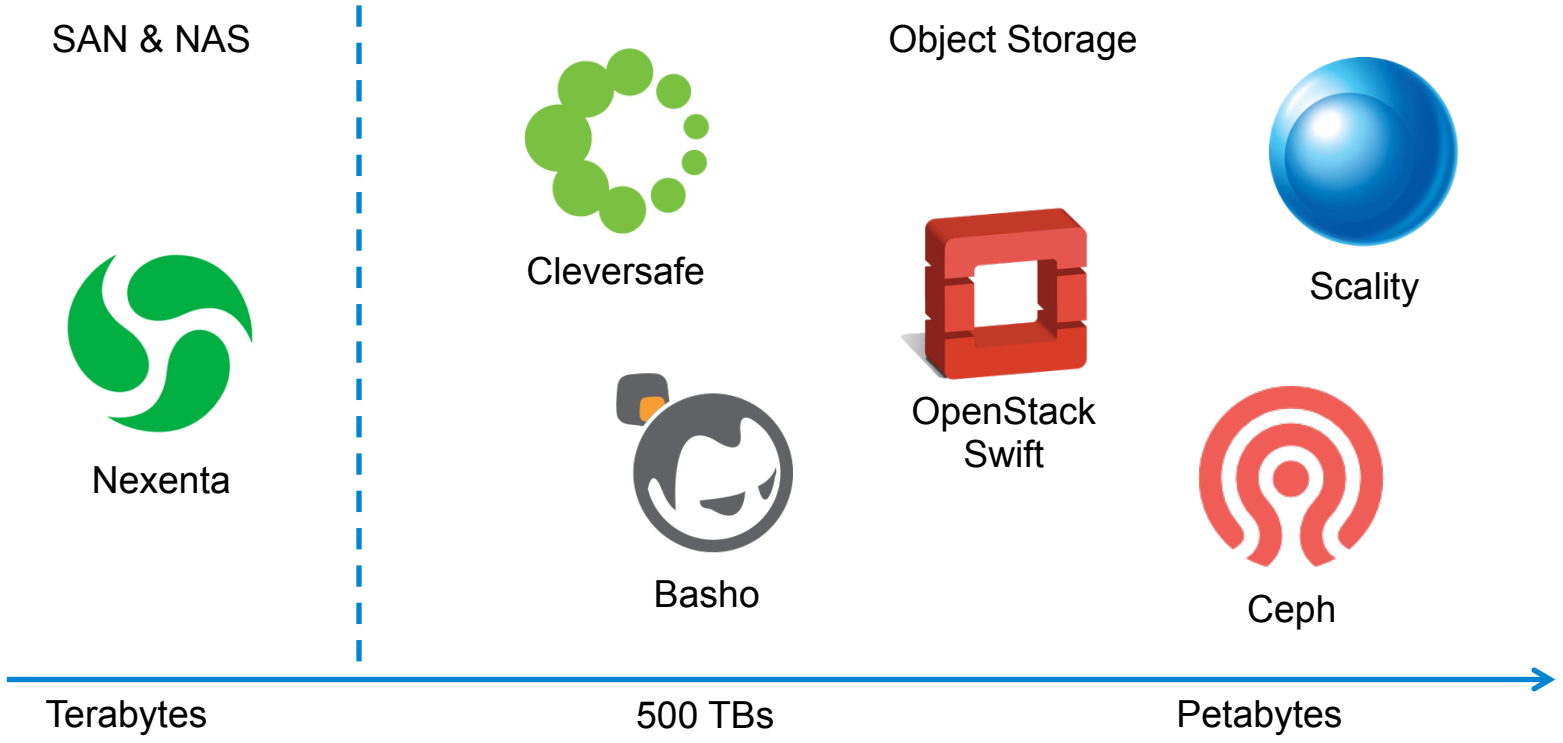
Scality is Validated & Supported By Seagate For Key Customer Use Cases

Validated Architecture

- Use Cases
- Tailored Hardware Configurations
- Shipped with SW installed
- Seagate options for Billing, etc
- Tested solutions
- Benchmarks
- End to end support
- Managed Services option



Software Defined CSS Validated Architecture Family

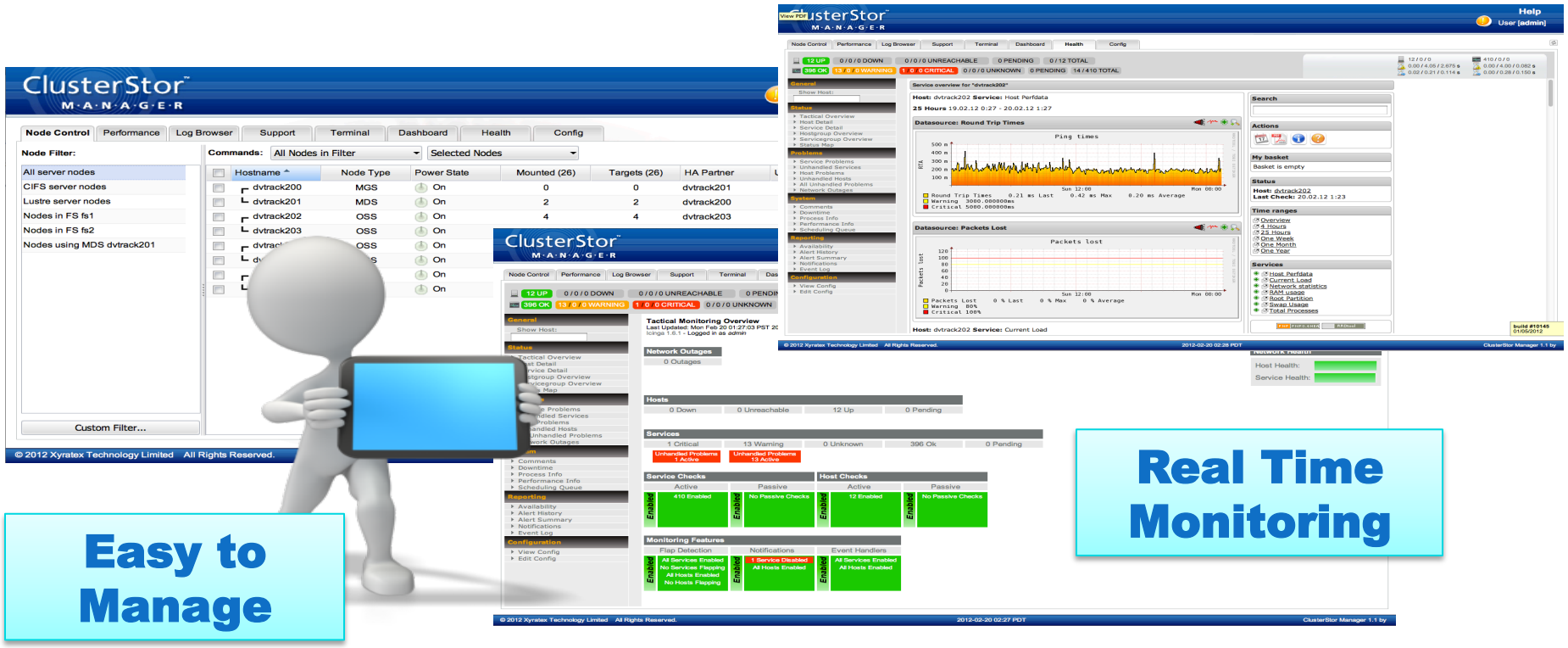


Seagate

Single Pane of Glass for Infrastructure and data Management

Fully Integrated Solution Visibility and Management

Low level diagnostics, embedded monitoring, proactive alerts



Thank you!