BioIT 2015

DataDirect™ NETWORKS
INFORMATION IN MOTION®

Intelligent Infrastructure Approaches for Emerging Life Sciences Data Management Issues at Scale

Complete Data Lifecycle Solutions for Life Sciences Research including High-Performance Analysis and Secure Collaboration

George Vacek, PhD, MBA

Global Director, Life Sciences
gvacek@ddn.com
@DrV_DDN

Why DDN for the Life Sciences

DDN technology drives more actionable results in less time for more researchers than any other storage solution

- Single Platform for Ingest, Analysis, Search, Collaboration, Archive
- Deep Expertise in Life Sciences and Key Technologies
- Significantly Better Performance for Key Workflows
  - Genome Sequencing
  - Imaging
  - Modeling / Simulation

DataDirect™
NETWORKS
INFORMATION IN MOTION™

DDN – Building the Right Solution

DDN Technology

Best of Class standalone and integrated solutions

DataDirect NETWORKS
INFORMATION IN MOTION

Application Acceleration, Burst Buffer — **Infinite Memory Engine*™**

Petascale Lustre® Storage

EXAScaler™

Up to 100,000 Clients
Open Source
Leading metadata & small file performance

Enterprise Scale-Out GPFS File Storage

GRIDScaler™ / GS7K*

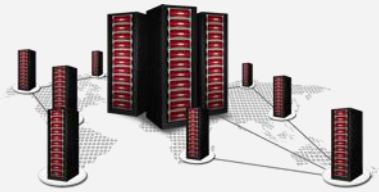
Up to ~16,000 Clients
Integrated Object Storage, Tape
NFS & CIFS
Snapshots, Replication

Content Distribution, Cloud Storage, Active Archive

WOS®

Web Object Scaler

- 32 Trillion Unique Objects
- Data Protection & Distribution via Replication or Erasure Coding
- Lowest latency data access and rebuild



Options

- Single 4U appliance up to 64 sites
- OCP-compliant Hardware
- Software-only

Unified Management

DirectMon™

Core Storage Platforms

SFA™12KX

48GB/s, 1.7M IOPS
1,680 Drives in 2 Racks
Embedded Computing

SFA™7700X

12.5GB/s, 450K IOPS
60 Drives in 4U
396 Drives in 20U

Flexible Drive Configuration

All Flash, Hybrid or HDD-only

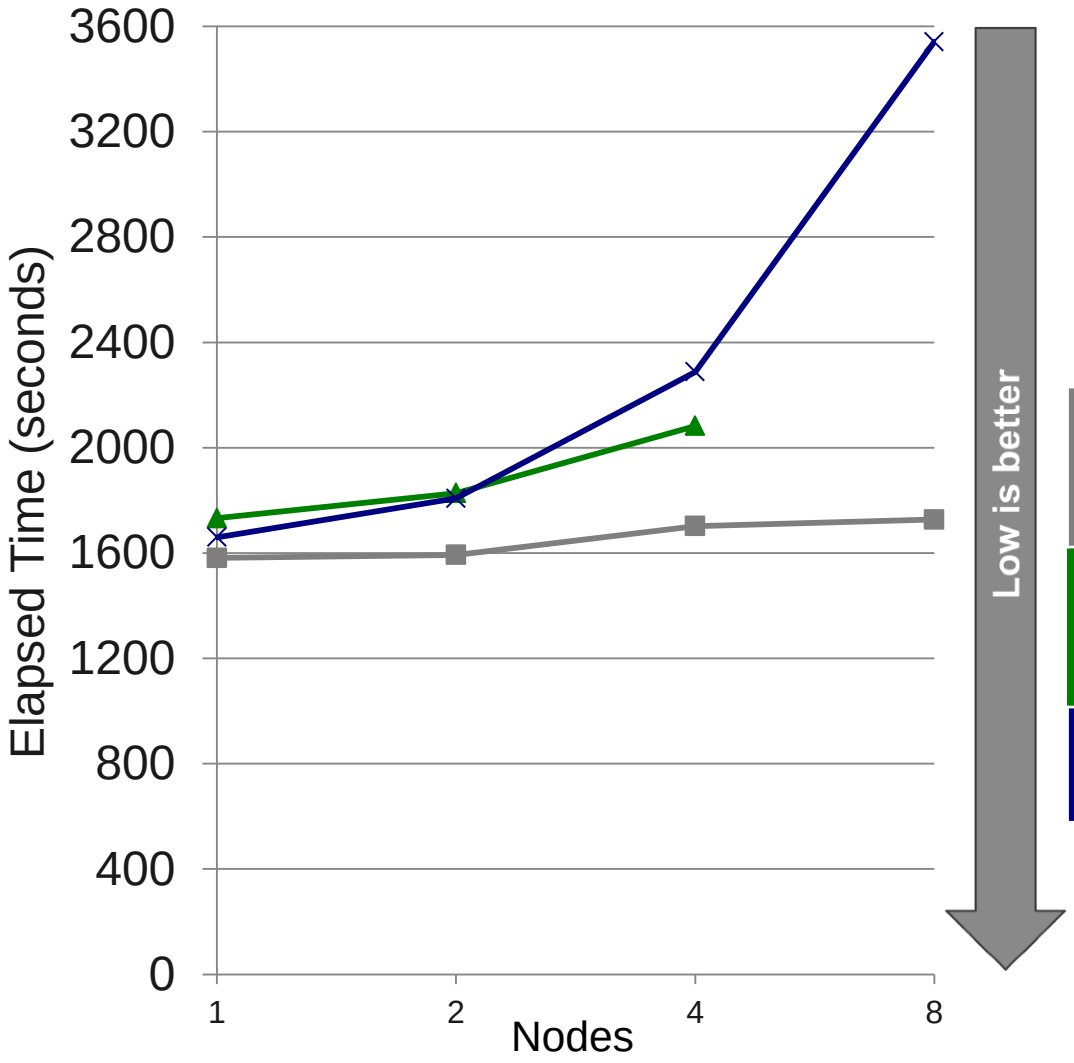
SAS **SSD** **SATA**

Automated Flash Caching

SFX™ Adaptive Transparent Flash Cache
SFX API Gives Users Control
[pre-staging, alignment, by-pass]

BWA Network Attached Storage Test

Effect of adding fully loaded systems



BWA-SW v0.6.1
Genome ERR003014

Sandy Bridge – RH EL 6.1 (16c)
Architecture: E5-2670 (2.6 GHz)
Cache: 20MB / 8 cores
Node: 2-processor 16-core SL230s
16 single thread jobs / node

GRIDScaler - array using 2 **DDN SFA 10k** controllers connected to QDR, then connected to FDR

NFS mounted (via FDR IB) **xfs file system** on a 4disk RAID0 stripe (node1 exports to node[2-N])

NFS Mounted (via FDR IB) **ext4 file system** on a RAID6 array

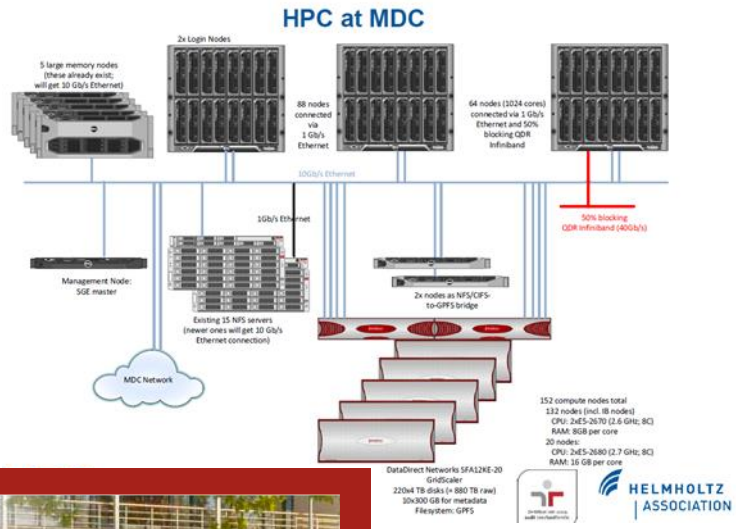
Max Delbrück Center for Molecular Medicine

7x Faster Variant Calling with DDN

DataDirect NETWORKS
INFORMATION IN MOTION

- Variant calling with GATK
  - 247 BAM files
- Legacy NFS server 36 hours
- GRIDScaler SFA12KE-20 5 hours
- Application important in clinical use
  - Could not practically be run on NFS
- Many instances can now run in parallel without loss of performance

7x Faster Variant Calling
on DDN At Max Delbrück Center for Molecular Medicine



"If someone else was putting load on the NFS server, it was typical that the [work] never completed"

- Alf Wachsmann, CIO
Max Delbrück Center for Molecular Medicine

Video at www.youtube.com/watch?v=URVXHb5GAI4

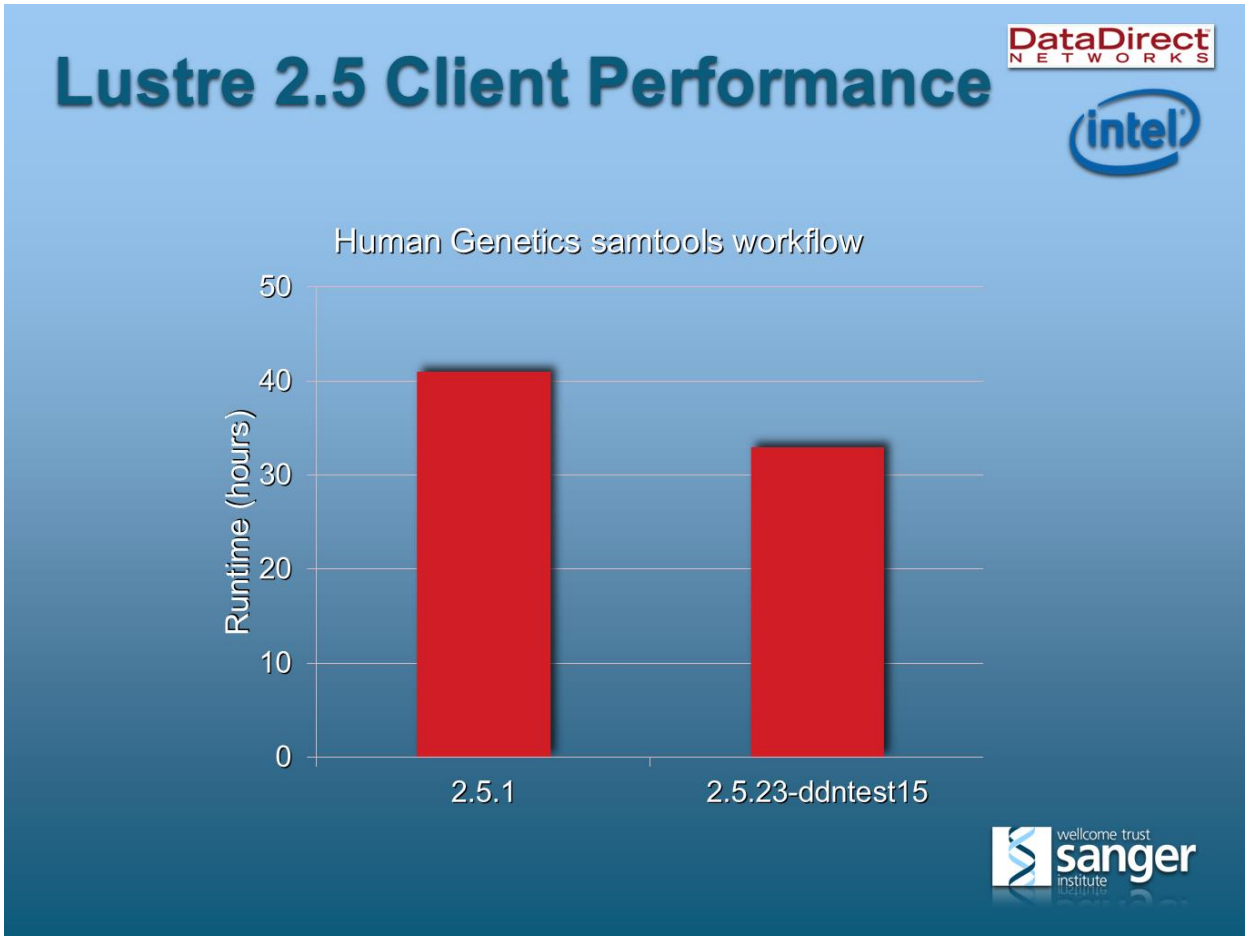
wellcome trust sanger institute

DataDirect NETWORKS
INFORMATION IN MOTION

Samtools 20% faster with DDN optimizations

"As the scale [of computing] got bigger, so did the need to support our Lustre file system with the fastest, most reliable storage available."

- Tim Cutts, Head of Scientific Computing, Wellcome Trust Sanger Institute



Video: www.youtube.com/watch?v=bCWdAaW3PVA

DataDirect™ NETWORKS
INFORMATION IN MOTION™

Customer Case Studies

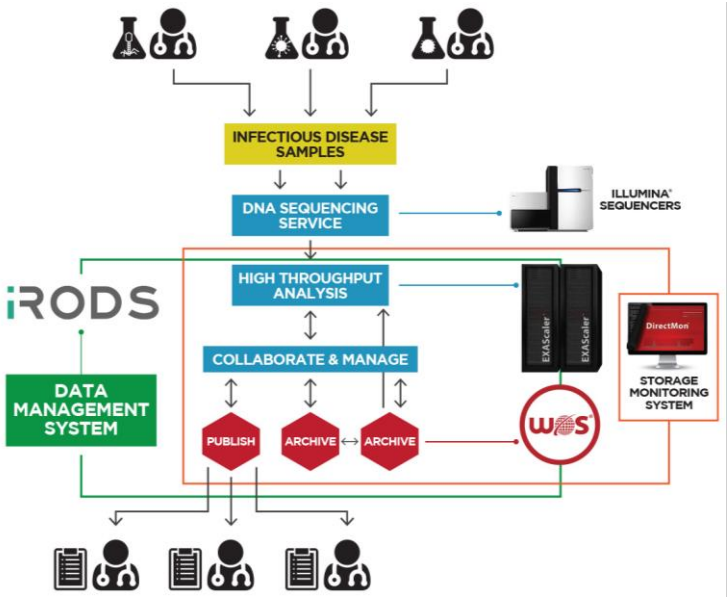
Diagnosis and Surveillance of Infectious Diseases
Public Health England

Public Health England

DataDirect Networks
Information in Motion

"PHE realizes the benefits of HPC and big data storage and is using both to set standards for the rest of the world to follow. PHE is pioneering use of DNA bacterial sequence data to provide a public service. It's the first project of its type in the World."

- Julian Fielden
Managing Director, OCF



- Cost effective public health intervention
  - ID potentially aggressive pathogens
  - Public sharing of data
- Fast, Effective Genome Analysis
  - 2 @ Illumina HiSeq
  - Thousands of patient samples / week
- 300TB SFA with EXAScaler
  - 16x increase in throughput
  - Variant calling & de novo assembly
- 360TB WOS
  - Collaboration between PHE sites
  - Active archive for reanalysis
- iRODS
  - Automate workflows
  - Organize and search project data

BioIT World – Wednesday, April 22, 12:15pm
Dr. Anthony Underwood, Public Health England

Integrated Imaging Program at Albert Einstein

Albert Einstein College of Medicine OF YESHIVA UNIVERSITY

DataDirect NETWORKS INFORMATION IN MOTION

Challenges

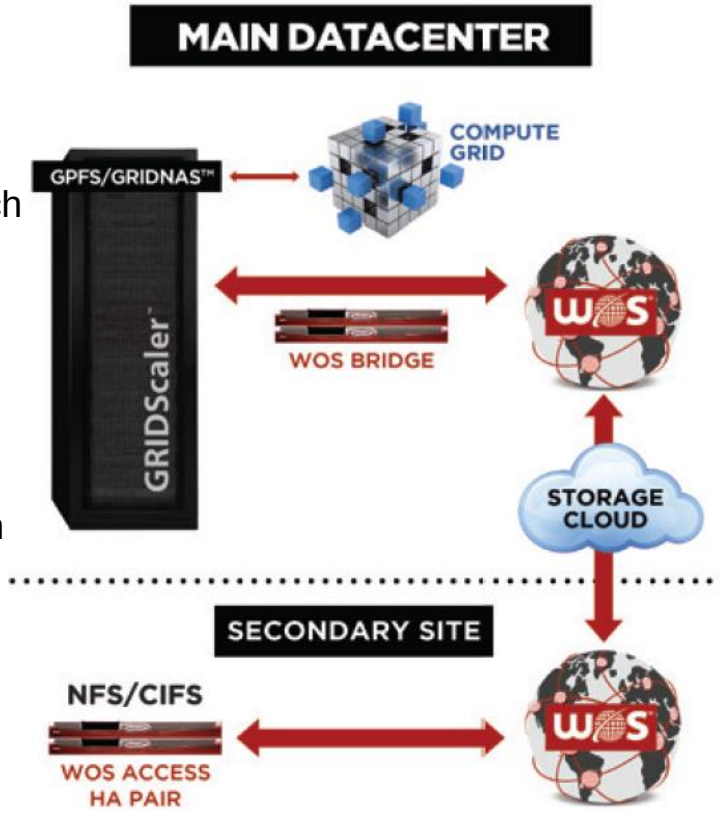
- Scale to meet 1TB / week data growth
- Coordinate IT driven by decentralized, department-based research
- IT support with limited staff

GRIDScaler + WOS

- Fully integrated single architecture
- Single namespace and ID management for CIFS, NFS, GPFS
- Automated policies minimize management and reduce TCO
- Performance and scalability for high data rate capture and growth

Specimen images into research cloud

- Active archive frees up space
- Enables research collaboration across institutions
- Distribute to remote sites for disaster protection
  - Replication and erasure coding



"WOS integrates seamlessly into our high performance storage environment, by federating our data sets from across the multiple data centers in our research community, and helping us to meet retention and data protection requirements."

- Shailesh M. Shenoy

Director of Engineering and Operations for Einstein's Integrated Imaging Program

VirginiaTech
Virginia Bioinformatics Institute

Rapid response to emerging crises

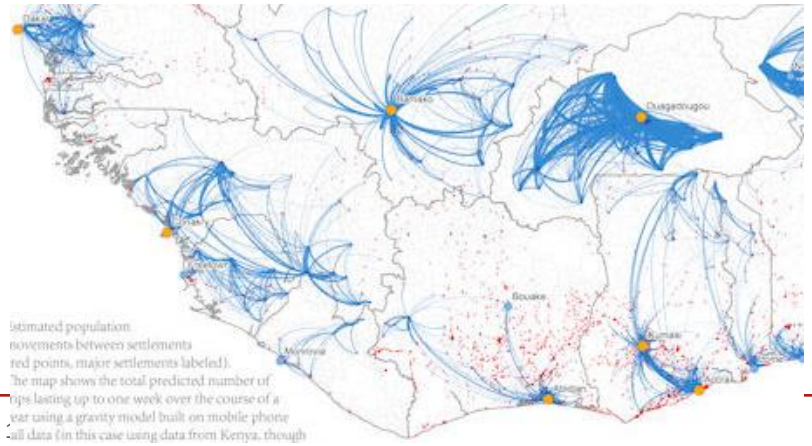
DataDirect NETWORKS
Information in Motion

"With DDN, we've attained a fast, reliable parallel file system to handle all our different workloads."

- Dr Kevin Shinpaugh
Director of IT and HPC, VBI

"As with all epidemics, time is of the essence. The ability to have all the necessary data at our fingertips is essential to delivering rapid answers to a series of tough questions."

- Dr Bryan Lewis
Computational Epidemiologist, VBI



Computational epidemiology requires

- Scalable storage for highly detailed synthetic regional and global populations
  - Demographics
  - Family structures and social networks
  - Activities and travel patterns
- Hundreds of simulations with variable parameters for each of dozens of models
- Flexible, high performance, data-heavy and I/O parallel processing

1PB SFA GRIDScaler

- Supports both data and compute-intensive workflows
- Industry leading density and reliability

Rapid response to emerging outreak crises

- Ebola recommendations within 48 hours
  - Defense Threat Reduction Agency (DTRA)
  - National Institutes of Health (NIH)
  - World Health Organization (WHO)
  - West Africa's Ministries of Health (MOH)

Select Life Sciences Customers

DataDirect NETWORKS
INFORMATION IN MOTION

wellcome trust sanger institute
National Institutes of Health
ThermoFisher Scientific
Mount Sinai Hospital
MERCK
東北メディカル・メガバンク機構 Tohoku Medical Megabank Organization
MDC Max Delbrück Center for Molecular Medicine Berlin-Buch
Public Health England
HHMI Howard Hughes Medical Institute
OICR Ontario Institute for Cancer Research
Stanford Center for Genomics and Personalized Medicine
Weill Cornell Medical College
VirginiaTech Virginia Bioinformatics Institute
HudsonAlpha Institute for Biotechnology
King Abdullah University of Science and Technology
UCL
Keck School of Medicine of USC
SIB Swiss Institute of Bioinformatics
astellas Leading Light for Life
NIBMG
Purdue University
Scripps
UNC Lineberger
McGill
Einstein Albert Einstein College of Medicine of Yeshiva University
Harvard Medical School
USF Health
UF University of Florida
RIKEN
Kazusa DNA Res. Inst.
National Cancer Institute
Chugai
GenomeQuébec
D E Shaw & Co
Human Genome Center Institute of Medical Science, University of Tokyo
University of Oxford

"One of our requirements was that our provider must understand what we are trying to achieve and recognize what would best meet our needs."

- Tim Cutts, Head of Scientific Computing
Wellcome Trust Sanger Institute

DataDirect™ NETWORKS
INFORMATION IN MOTION™

Summary

Complete Life Sciences Data Lifecycle

DataDirect NETWORKS
INFORMATION IN MOTION

Deeper Processing

Massive Ingest

Extract
Collaborate

Store at Greater Scale

Accelerate Discovery Throughout the Data Lifecycle

Ingest → Analyze → Search → Collaborate → Archive

Industry Leading Performance, Density, Capacity and TCO

DataDirect™
NETWORKS
INFORMATION IN MOTION™

Thank You!

See more at Booth #233 or ddn.com