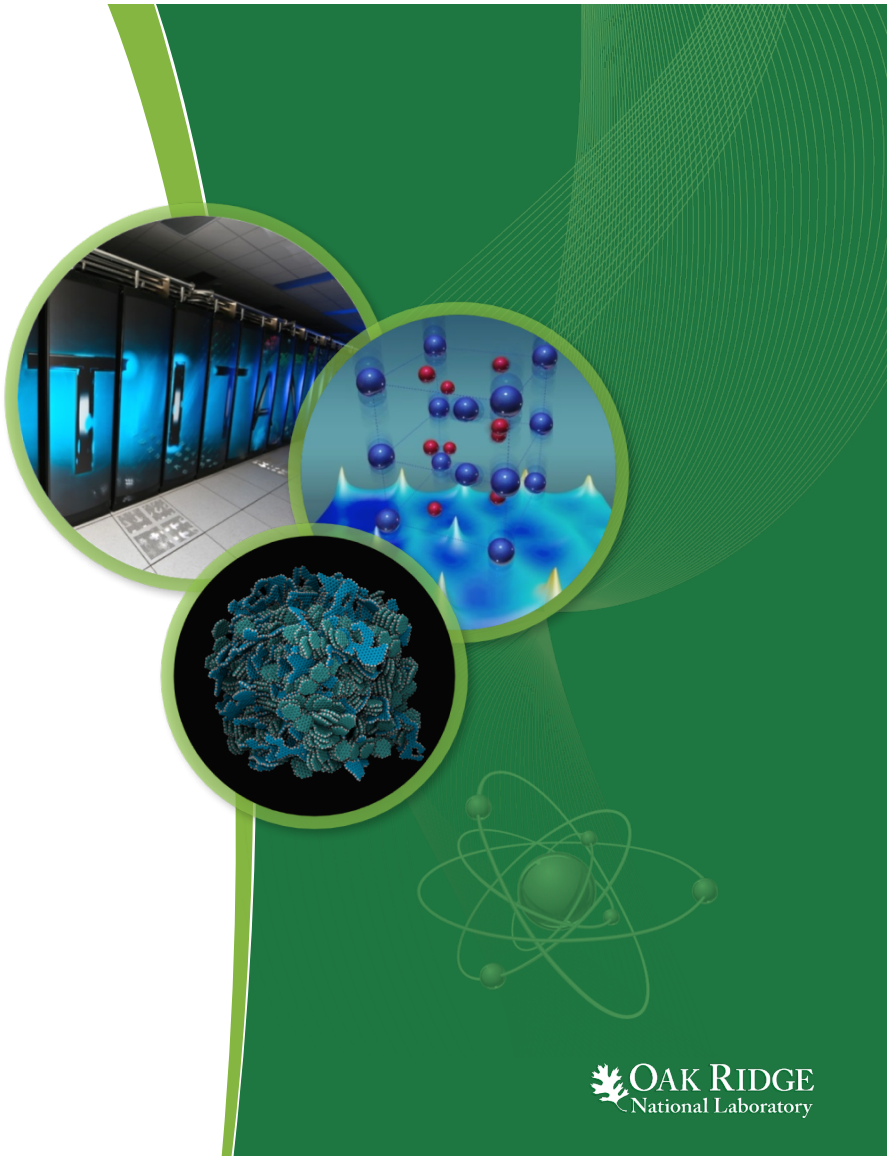
Managing HPC Active Archive Storage with HPSS RAIT at Oak Ridge National Laboratory

Quinn Mitchell

HPC UNIX/LINUX Storage Systems



ORNL is managed by UT-Battelle
for the US Department of Energy

U.S. Department of Energy user facilities in Top500 HPC

- **Oak Ridge Leadership Computing Facility**
  - Titan: #1 on 2012/11, now #4 @ 17.6PF
- **Argonne Leadership Computing Facility**
  - Mira: #3 on 2012/6, now #9 @ 8.6PF
- **National Energy Research Scientific Computing Center**
  - Edison: #18 on 2014/6, now #72 @ 2.6PF
  - Cori: #5 on 2016/11, now #6 @ 14PF
- **Lawrence Livermore National Lab**
  - Sequoia: #1 on 2012/06, now #5 @ 17.2PF
- **Los Alamos & Sandia National Labs**
  - Cielo: #6 on 2011/06, now #107 @ 1.4PF
  - Trinity #6 on 2015/11, now #10 @ 8.1PF

Other non DOE HPC Top10 around the world

- National Supercomputing Center in Wuxi (China)
  - Sunway TaihuLight: #1 on 2016/06, now #1 @ 93PF
- National Super Computer Center in Guangzhou (China)
  - Tianhe-2: #1 on 2013/06, now #2 @ 33.9PF
- Swiss National Supercomputing Centre (Switzerland)
  - Piz Daint: #114 2012/11 @ .2 PF, now #3 @ 19.6PF
- Joint Center for Advanced HPC (Japan)
  - Oakforest-PACS: #6 on 2016/11, now #7 @ 13.6PF
- RIKEN Advanced Institute for CS (Japan)
  - K computer: #1 2011/06, now #8 @ 10.5 PF

Oak Ridge Leadership Computing Facility

U.S. Department of Energy National User Facility

- 1,100 users from Universities, Government Labs, Industry, and around the world
- Resources are available to anyone through a peer-reviewed proposal process
- We have 3 times more users wanting access to Titan than we have resources to allocate.

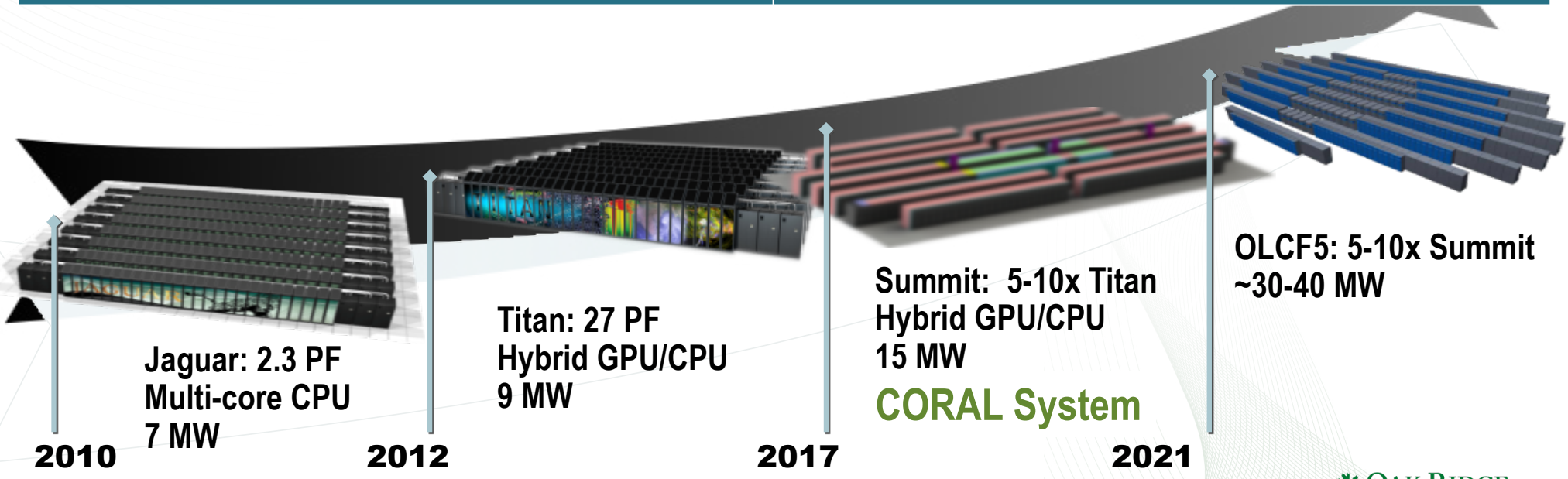
Oak Ridge Leadership Computing Facility

Mission: Deploy and operate the computational and data resources required to tackle global challenges

- **Providing world-leading computational and data resources and specialized services for the most computationally intensive problems**
- **Providing stable hardware/software path of increasing scale to maximize productive applications development**
- **Providing the resources to investigate otherwise inaccessible systems at every scale: from galaxy formation to supernovae to earth systems to automobiles to nanomaterials**
- **With our partners, deliver transforming discoveries in materials, biology, climate, energy technologies, and basic science**

Our mission requires that we continue to advance DOE's computational capability over the next decade on the roadmap to Exascale and beyond.

Since clock-rate scaling ended in 2003, HPC performance has been achieved through increased parallelism. Jaguar scaled to 300,000 cores. Titan has >560K. Summit will have more than 1.5 million.	Titan and beyond deliver hierarchical parallelism with very powerful nodes. MPI plus thread level parallelism through OpenACC or OpenMP plus vectors



OAK RIDGE National Laboratory

Collaboration for recent DOE HPC Acquisitions

CORAL: Collaboration of Oak Ridge, Argonne, and Livermore

- Summit system at OLCF in 2018
- Sierra system at LLNL in 2017/18
  - Both Summit and Sierra are IBM systems based on IBM Power9 and NVIDIA processors
- Theta system at ALCF in 2016
- Aurora system at ALCF in 2018
  - Both Theta and Aurora are Cray systems based on Intel Xeon Phi processors
- Next CORAL systems are exascale systems in 2021/22

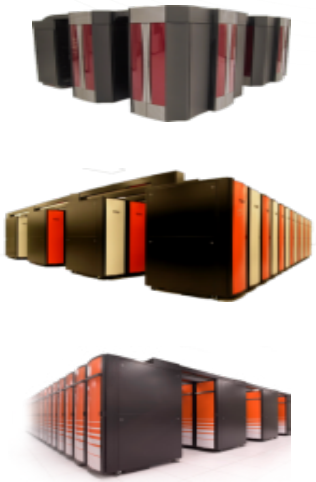
DOE Office of Science Upgrades At a Glance

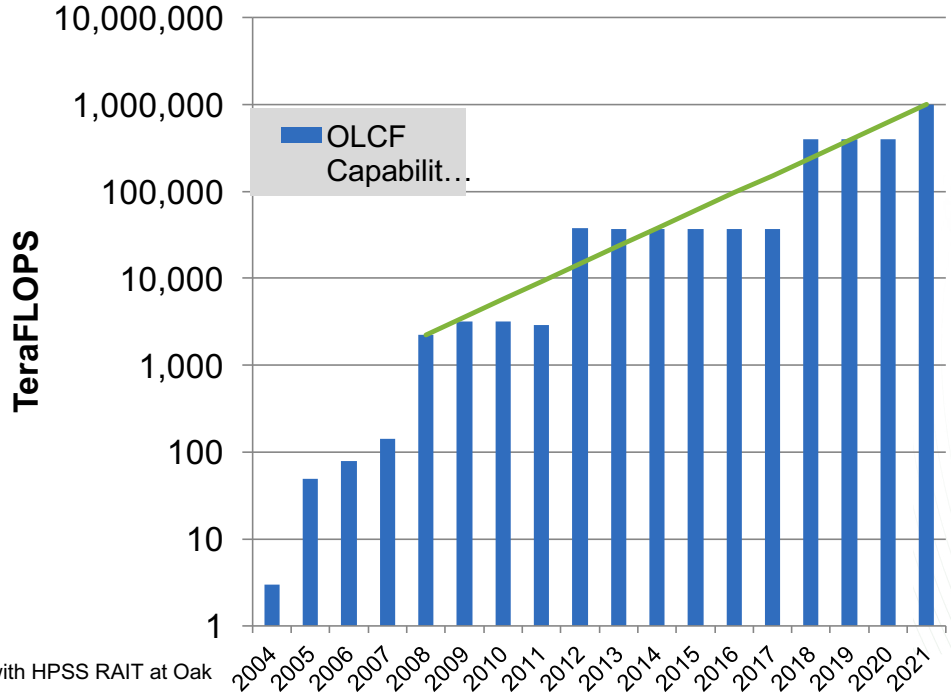
System attributes	OLCF Now	ALCF Now		NERSC Now		OLCF Upgrade	ALCF Upgrade
Name Planned installation	**TITAN**	**MIRA**	**Theta 2016**	**Edison**	**Cori**	**Summit 2017-2018**	**Aurora 2018-2019**
System peak (PF)	27	10	>8.5	2.6	~ 31	200	180
Peak Power (MW)	9	4.8	1.7	2	3.5	13.3	13
Total system memory	710TB	768TB	>480 TB DDR4 + High Bandwidth Memory (HBM)	357 TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	> 2.4 PB DDR4 + HBM + 3.7 PB persistent memory	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	1.452	0.204	>3	0.46	>3	> 40	> 17 times Mira
Node processors	AMD Opteron Nvidia Kepler	64-bit Power PC A2	Intel Knights Landing Xeon Phi many core CPUs	Intel Ivy Bridge	Intel Xeon Phi KNL Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Knights Hill Xeon Phi many core CPUs
System size (nodes)	18,688 nodes	49,152	>2,500 nodes	5,600 nodes	9,300 KNL nodes + 2,000 nodes in data partition	~4,600 nodes	>50,000 nodes
System Interconnect	Gemini	5D Torus	Aries	Aries	Aries	Dual Rail EDR-IB	2nd Generation Intel Omni-Path Architecture
File System	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS TM	10PB, 210 GB/s Lustre initial	7.6 PB 168 GB/s, Lustre®	28 PB 744 GB/s Lustre® , 1.5 TB/s Burst Buffer	250 PB 2.5 TB/s GPFS™	150 PB 1 TB/s Lustre®

https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/201704/20170418-Helland-ASCAC.pdf

The OLCF has increased our system capability by 10,000 times since our founding in 2004

- Strong partnerships with computer designers.
- OLCF users employ large fractions of machine resources for long periods of time.
- Science delivered through strong user partnerships to scale codes and algorithms.



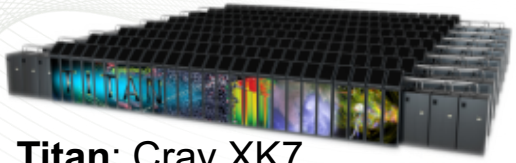


OAK RIDGE National Laboratory

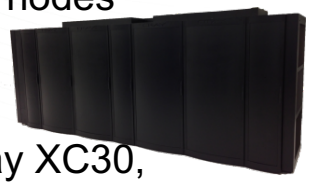
Titan & Summit Differences that change archive storage requirements

- Application performance of 5-10x over Titan
  - Applications generate much more data and larger files requiring more bandwidth – *Need increased bandwidth and metadata performance*
- Very large memory – *Drives the amount of data stored*
  - Summit has ~8x more memory than Titan
  - But only twice the DRAM and a new non-volatile memory layer
- 250 PB GPFS scratch file system – *Reduces frequent transfers*
  - 8x larger than Titan's 32PB Lustre file
  - 2.5x more bandwidth

OLCF Simulation Environment



Titan: Cray XK7, Hybrid CPU/GPU 18,688 nodes

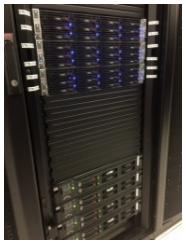


EOS: Cray XC30, Xeon Processors, 736 Nodes

Rhea: Linux cluster, Xeon Processors, 512 Nodes

SION Scalable I/O Network FDR Infiniband

Spider: 32 PB, 1 TB/s Lustre File System



DTN: Data Transfer Nodes

EVEREST: Science Visualization Facility

Everest: Visualization Cluster

Sith: Workflow Cluster

CADES: Compute And Data Environment for Science

HPSS: Archival Storage System

OAK RIDGE National Laboratory

What is HPSS?

- A 25-year old collaboration between IBM and five Department of Energy laboratories
- Extreme scale storage software for:
  - High performance data movement
  - Huge single namespace capacity
  - Efficient use of tape resources
- HPSS is hardware vendor neutral
- Half of Top500.org Top10 use HPSS

HPSS
High Performance Storage System

Berkeley Lab

OAK RIDGE National Laboratory

IBM

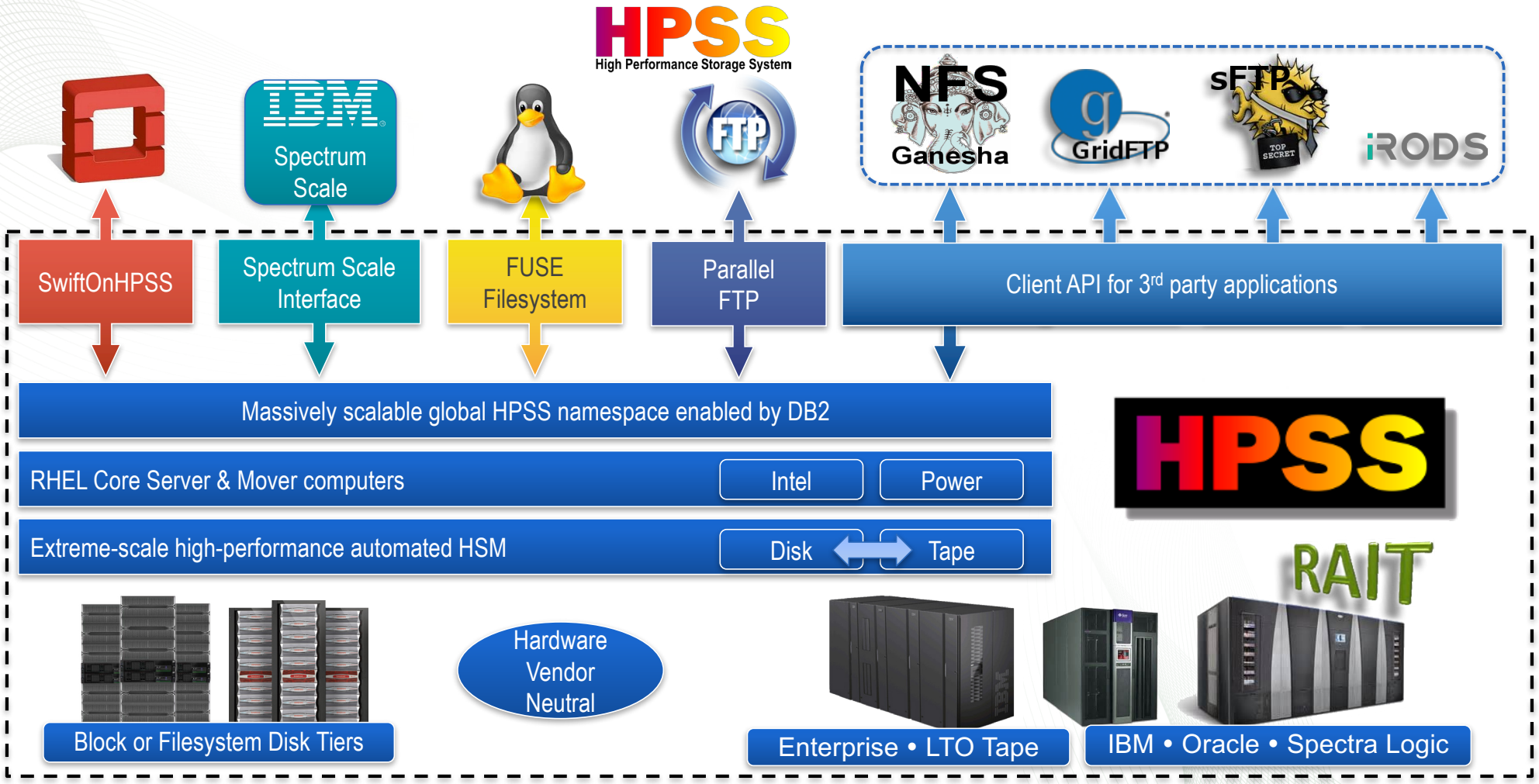
Los Alamos NATIONAL LABORATORY

University of California Lawrence Livermore National Laboratory

Sandia National Laboratories

Exabytes!

OAK RIDGE National Laboratory


HPSS
High Performance Storage System
NFS
Ganesha
GridFTP
sFTP
iRODS
Spectrum Scale
FTP
SwiftOnHPSS
Spectrum Scale Interface
FUSE Filesystem
Parallel FTP
Client API for 3rd party applications
Massively scalable global HPSS namespace enabled by DB2
RHEL Core Server & Mover computers
Intel
Power
Extreme-scale high-performance automated HSM
Disk
Tape
HPSS
RAIT
Hardware Vendor Neutral
Block or Filesystem Disk Tiers
Enterprise • LTO Tape
IBM • Oracle • Spectra Logic

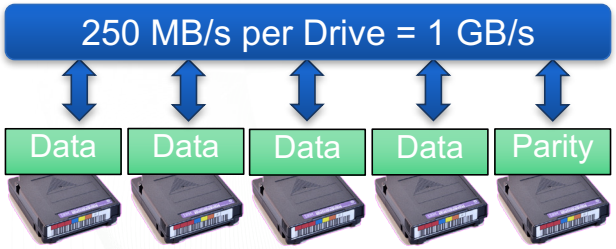
Extreme scale capacity and bandwidth

HPSS
High Performance Storage System

- A single instance of HPSS, at ECMWF (European Center for Medium-Range Weather Forecasts), grew from 184 PB to 270 PB over the past year – almost 236 TB/day
- Oak Ridge National Laboratory cut redundant tape costs by 75% with 4+1 HPSS RAIT and enjoys single file tape transfers well beyond 1 GB/s
- Météo France demonstrated tape ingest of 2.6 M files at 219 MB/s per tape drive, and 960 tape library mounts per hour!

ECMWF

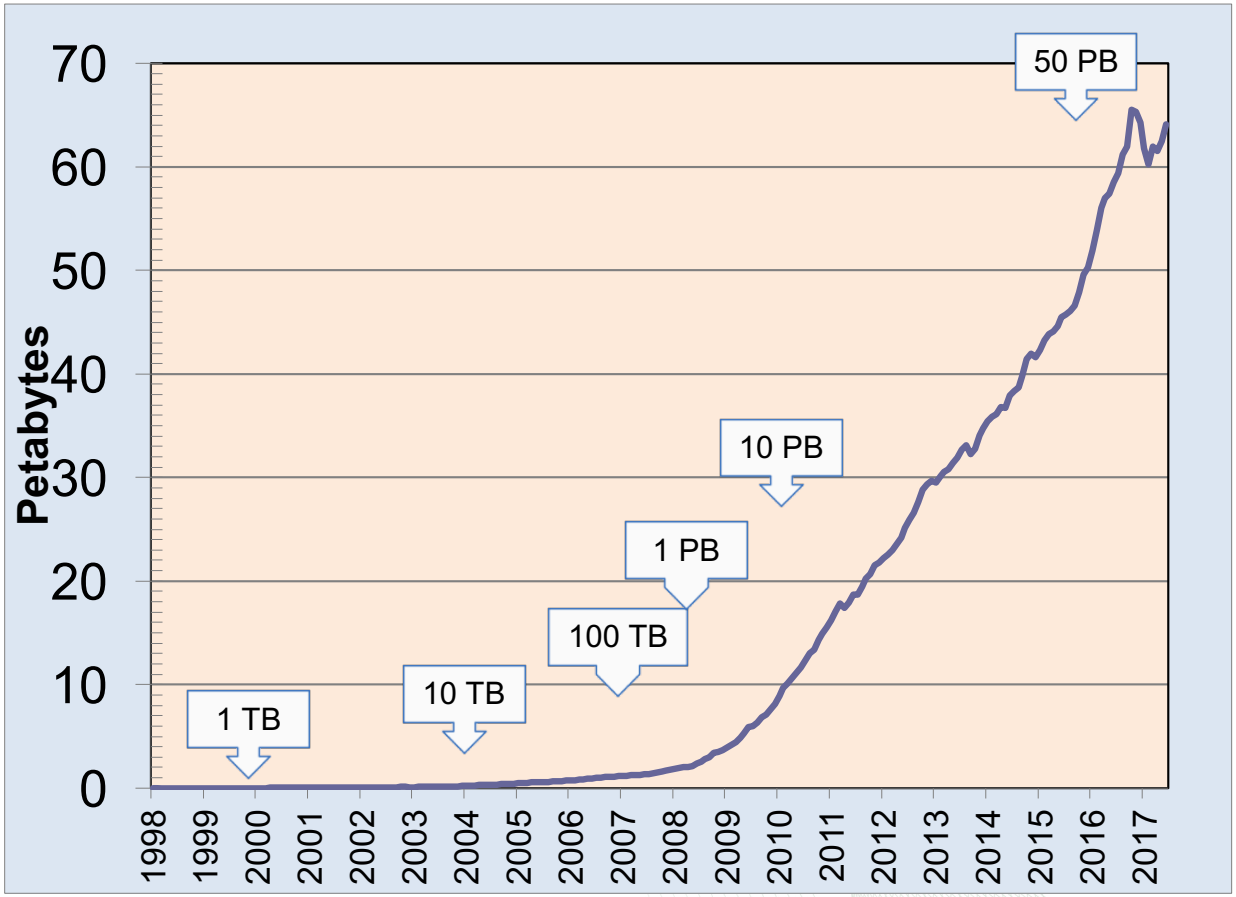
OAK RIDGE National Laboratory



METEO FRANCE

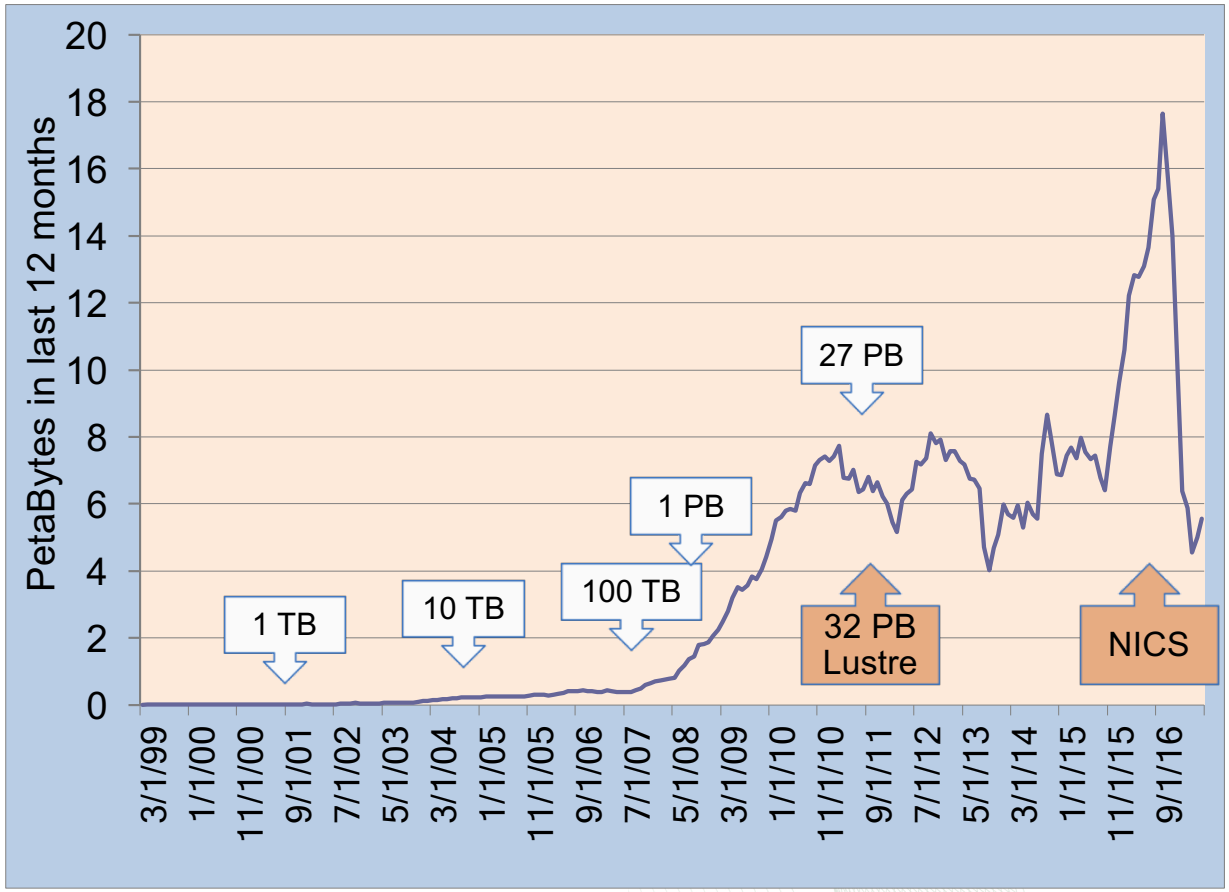
HPSS Data Growth at ORNL

- ORNL has been running HPSS since 1998
- Continuous growth in stored data over that time
- What happened in 2008 that changed the slope of the line?
  - We installed Jaguar - our first 1 PetaFLOPS computer
  - NICS storing more data



HPSS Annual Data Growth

- This chart shows the amount of data growth in the previous 12 months
- Exponential growth until Titan arrived – so what changed the game?
- We added a 32 PB Lustre file system and users didn't have to put everything into HPSS to save it.
- NICS ceased archive activities to HPSS.



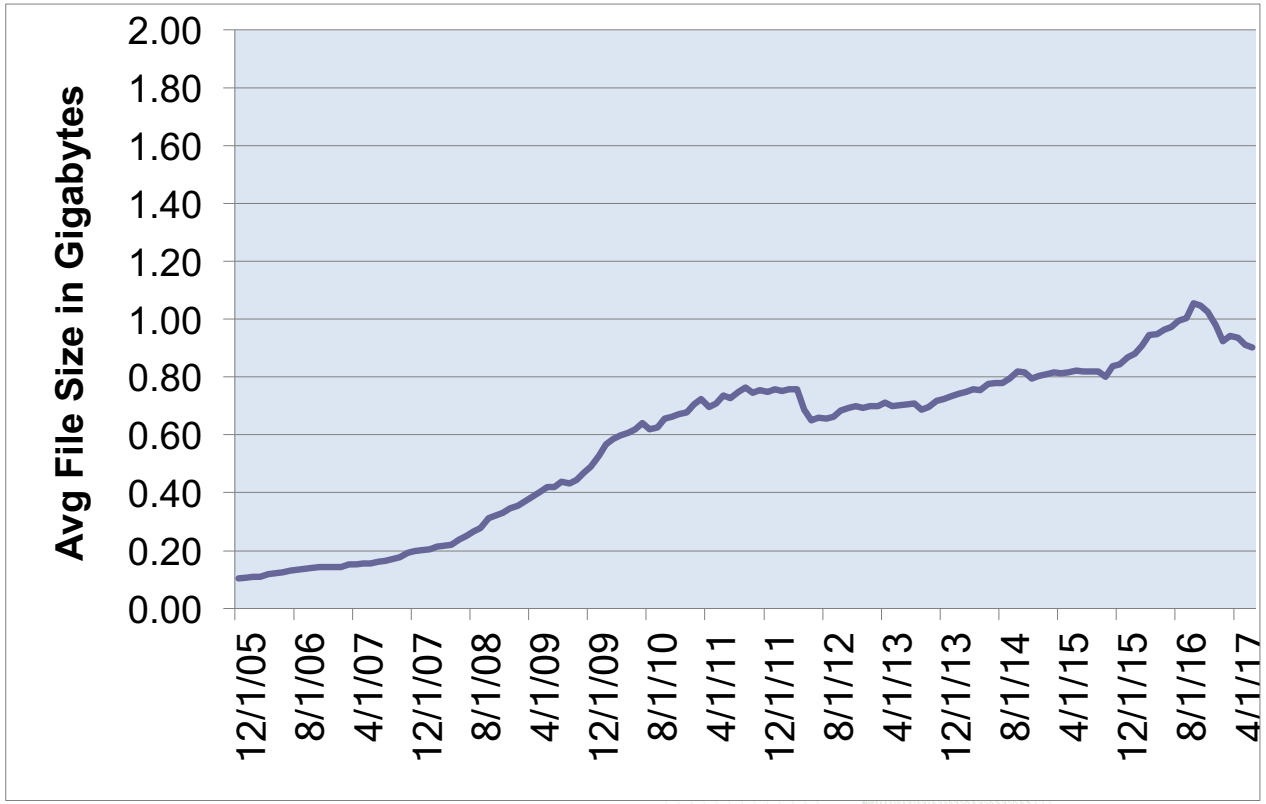
Profile of data stored in OLCF's HPSS

- 76 million files
- 58 PB of data
- 95% of files are smaller than 1 GB
- 95% of data is in files larger than 1 GB
- 20% of data is in files over 1 TB
- 30% of data retrieved in the first 30 days
- 70% of data never retrieved

Data current as of July, 2017

But there is an interesting trend

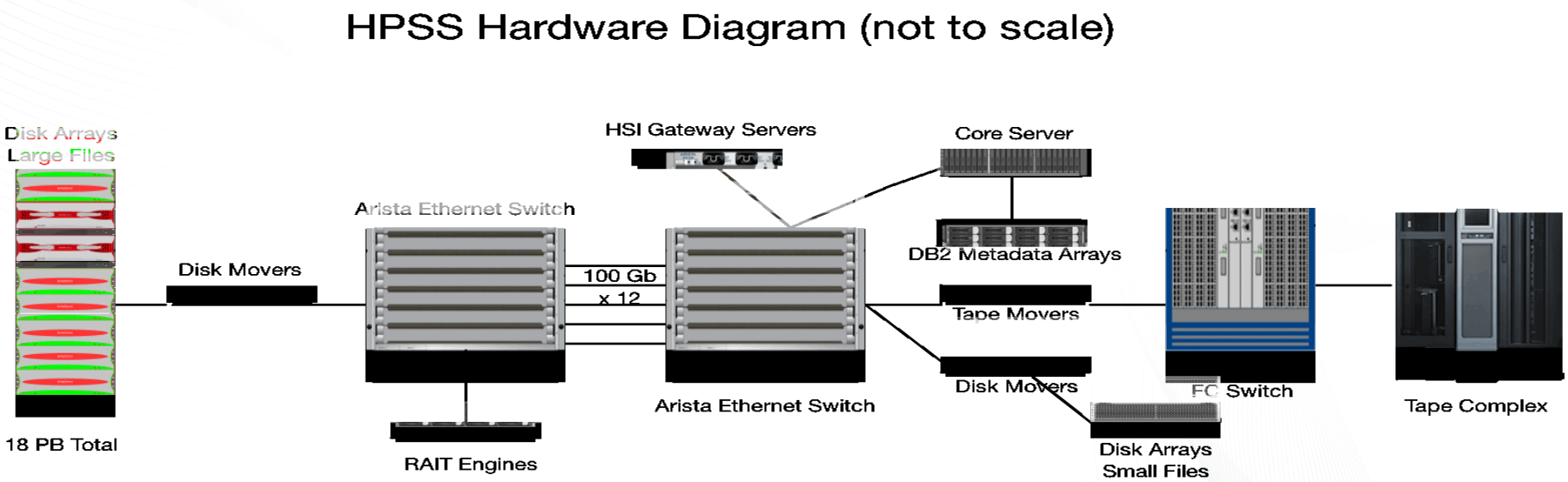
- In the last 10 years, the average file size has increased almost 10x



ORNL's HPSS Specifics

- Core Networks:
  - 100 GbE
  - FDR Infiniband
- Disk Cache:
  - 20 PB
  - 220 GB/s
- Data Transfer Nodes: 44
  - Globus
  - HSI
- Tape Storage:
  - 6 SL8500 Libraries
  - Tape drives
    - 120 T10K-D FC 16GB

Simplified HPSS Hardware Diagram

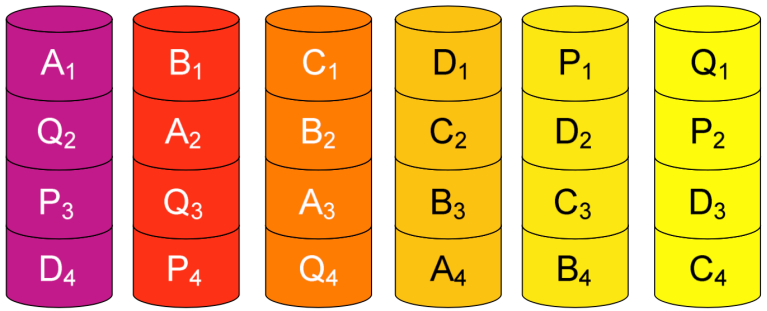


HPSS RAIT Overview

- Similar to RAID on disk, but for tape
- Time tested Reed Solomon Algorithm
- HPSS Notation
  - Data_Stripe_Count+Parity_Count
  - Ex. 4+2 four stripes and two parity
- Variable configuration
  - 2+1, 3+1... 8+1, 2+2, 3+2… 8+2
- Transfer speeds of striping, but with protection of multiple copies

Rotational Parity (RAIT 6)

4+2 Stripe (4 Data, 2 Parity)



HPSS RAIT at ORNL

- R/W speeds up to 5.31 GB/s with RAIT 4+1 T10KD

- HPSS Tape and RAIT configurations
  - 3 Copy : files smaller than 16MB, stored in aggregated sets
  - RAIT 2+1 : files between 16MB and 8GB, stored in aggregated sets
  - RAIT 4+1 : files larger than 8GB
  - 69 PB contained on 8,700 RAIT protected T10KD media
    - 58 PB data and 11 PB in parity

Data current as of July, 2017

Questions?