Data Storage for the era of Converged Big Data and HPC

Torben Kling Petersen, PhD
Principal Engineer

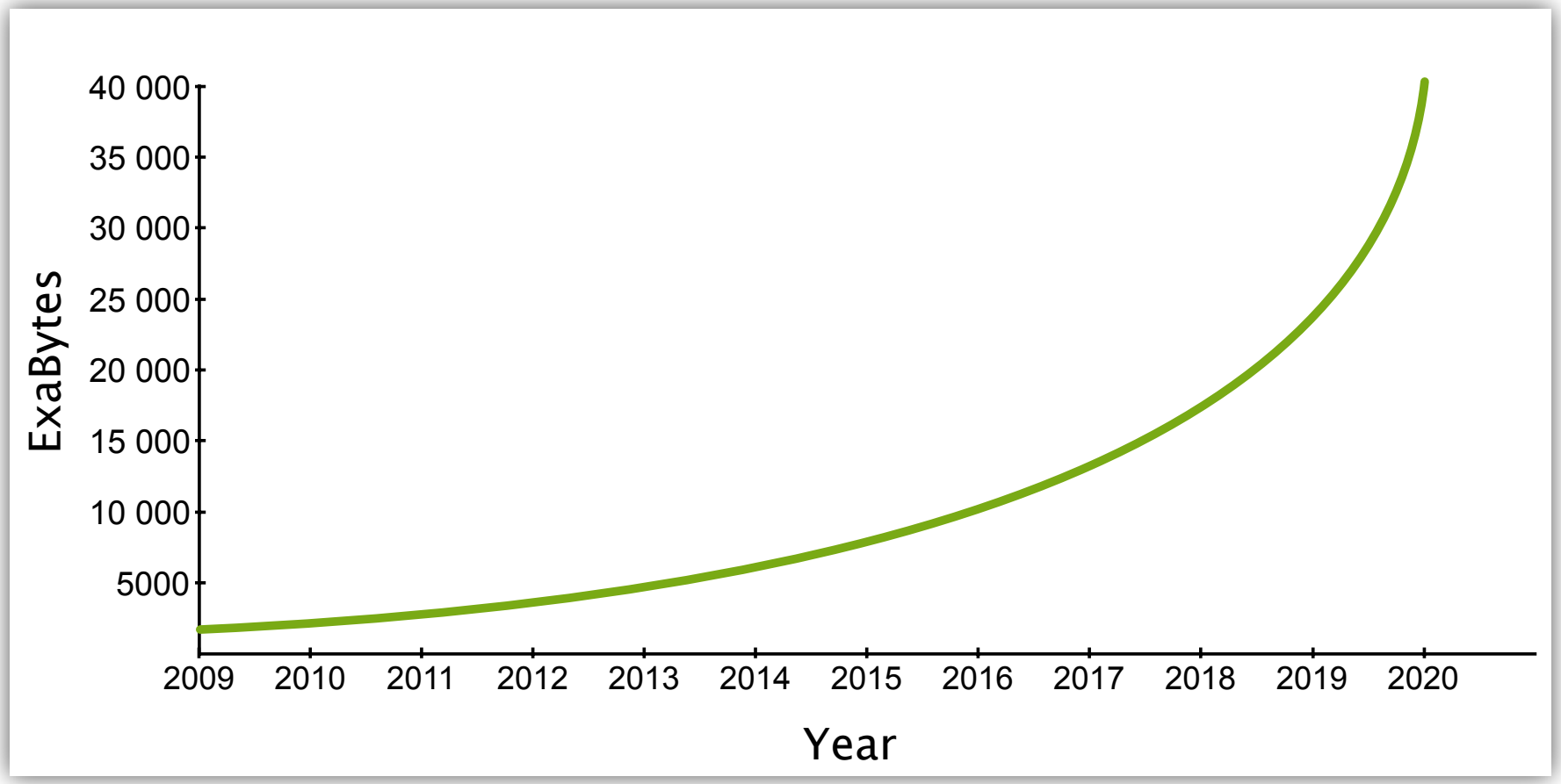
Seagate

Future High Performance Systems ….

Systems	2009	2018	Difference
System Peak	2 Pflop/sec	1 Eflop/sec	x 500
Power	6MW	20MW	x 3.33
System Memory	0.3 PBs	32-64 PBs	x 100 - 200
Node Compute	125 Gflop/s	1-15 Tflops/s	x 8 - 120
Node Memory BW	25 GB/s	2-4 TB/s	x 80 - 160
Node Concurrency	12	100 - 1000	x 8 - 80
Total Node Interconnect BW	3.5 GB/s	200-400 GB/s	x 50 - 100
System Size (Nodes)	18,700	100,000-1M	x 5 - 50
Total Concurrency	225,000	1,000,000,000	x 4400
Storage	**15 PB**	**500-1000 PB**	**x 30 - 60**
I/O	**0.2 TB/sec**	**60 TB/sec**	**x 300**

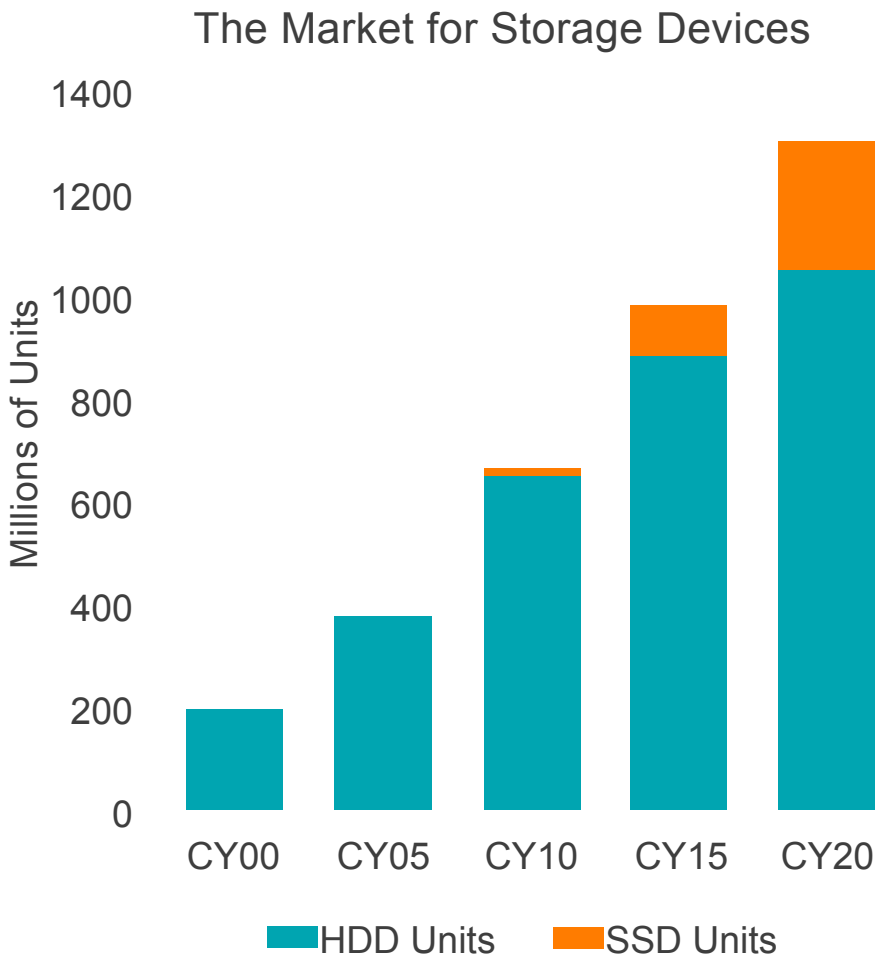
From J. Dongarra, “Impact of Architecture and Technology for Extreme Scale on Software and Algorithm Design,” Cross-cutting Technologies for Computing at the Exascale, February 2-5, 2010.

Predictions on Data Growth



From: IDC - THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East

The Market for All Storage Devices is Growing



By 2020…

HDD:

1 billion units

and

SSDs:

200 million units

**Source: Seagate Technology LLC projection*

Storage Solutions - Basic requirements …

… and some less basic ones ….

- Scalability and performance
- Cost and power efficiency
- Density
- Reliability – 5x9 ??
- Data security
- Data integrity – checksums or better
- Manageability – 1000+ storage nodes
- Supportability – systems in use for 4+ years
- TCO – analytics, statistics, predictive failure …

ClusterStor – An engineered solution

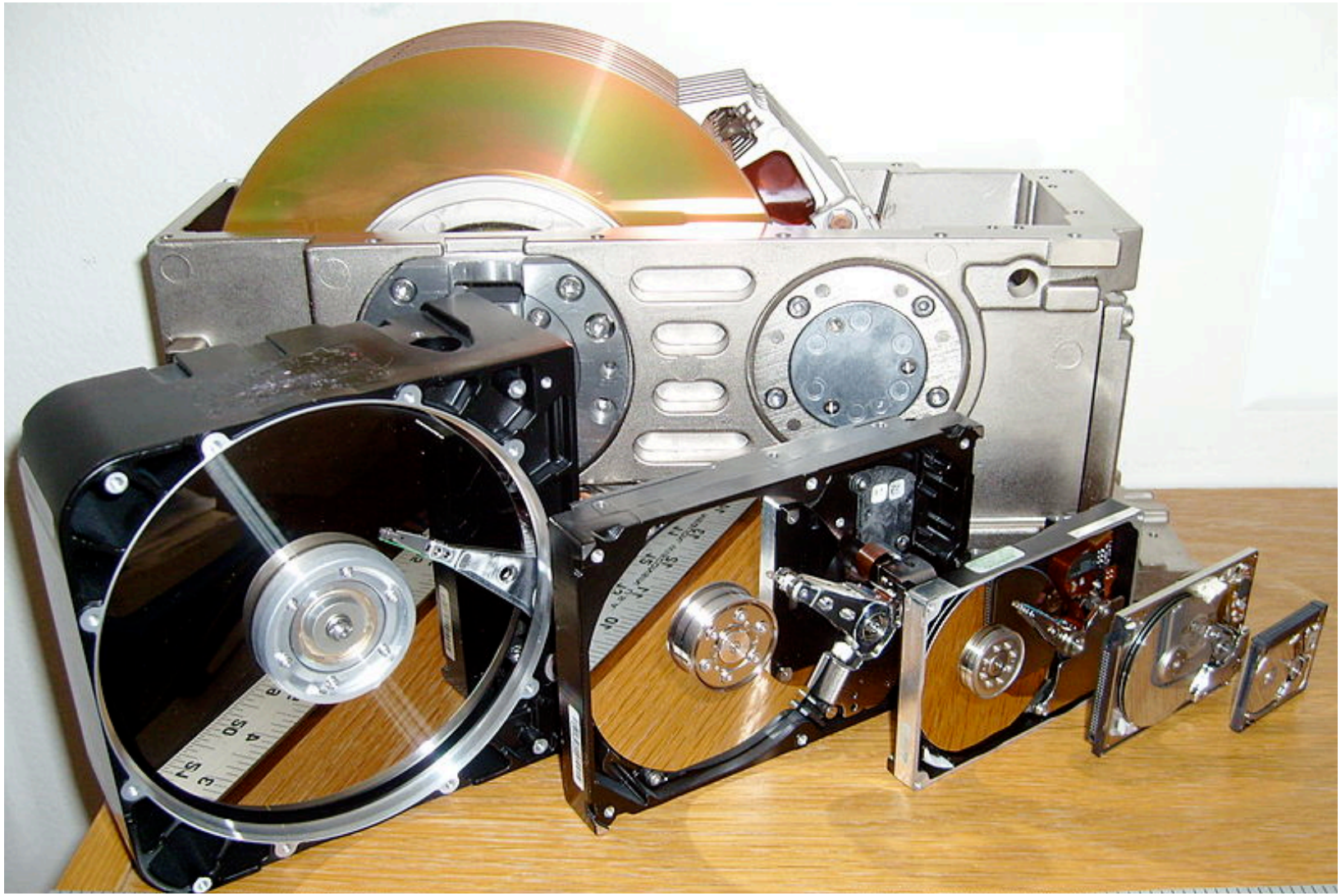
Complete Lustre appliance (.. almost)

- Designed for:
  - Extreme scalability - 100 PB+
  - Extreme performance >50 GB/rack
  - Extreme reliability – no SPOFs
  - Extreme supportability – FRUs
- Delivers Lustre with:
  - Fastest performance per rack
  - Highest density per rack
  - Complete solution
    - » Factory built and tested
    - » Industry leading GUI & CLI
    - » Best in class components

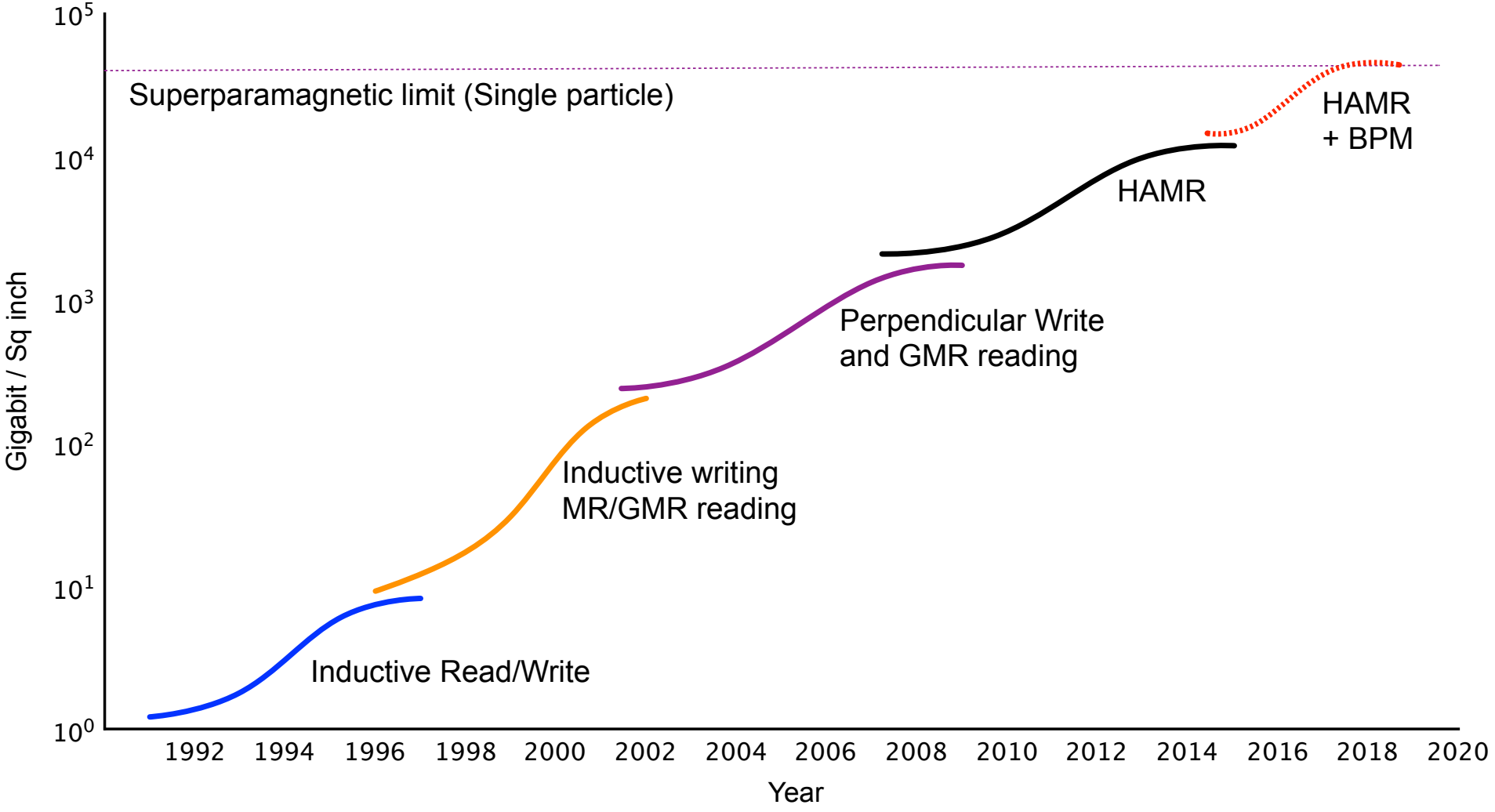
Future storage technologies

- Storage enhanced CPUs
- Disk fabrics
  - 12 Gbit SAS
  - 24 Gbit SAS
  - PCI-E direct attach
  - Ethernet attach (Kinetic drives)
- System interconnects
  - EDR IB
  - 40 & 100 Gbit ethernet
- HSM and enterprise features
- Disk drives, Solid State & NV-RAM technologies

Disk drive technologies



Areal density futures



Hard drive futures (2014 - 2015) …

- Sealed Helium Drives (Hitachi)
  - Higher density – 6 platters/12 heads
  - Less power (~ 1.6W idle) & Less heat (~ 4°C lower temp)
- SMR drives (Seagate)
  - Denser packaging on current technology
  - Aimed at read intensive application areas
- SSHD Hybrid drives (multiple vendors)
  - Enterprise edition
  - Transparent SSD/HHD combination (aka Fusion drives)
    - » eMLC + SAS

Hard drive futures (2015 - 2018) …

- HAMR drives (Seagate)
  - Using a laser to heat the magnetic substrate (Iron/Platinum alloy)
  - Projected capacity – 30-60 TB/ 3.5 inch drive …
  - 2016 timeframe ….
- BPM (bit patterned media recording)
  - Stores one bit per cell, as opposed to regular hard-drive technology, where each bit is stored across a few hundred magnetic grains
  - Projected capacity – 100+ TB / 3.5 inch drive …

* Current estimates

Secure data storage

Scalable solutions

Linux security features ….

RedHat SE Linux

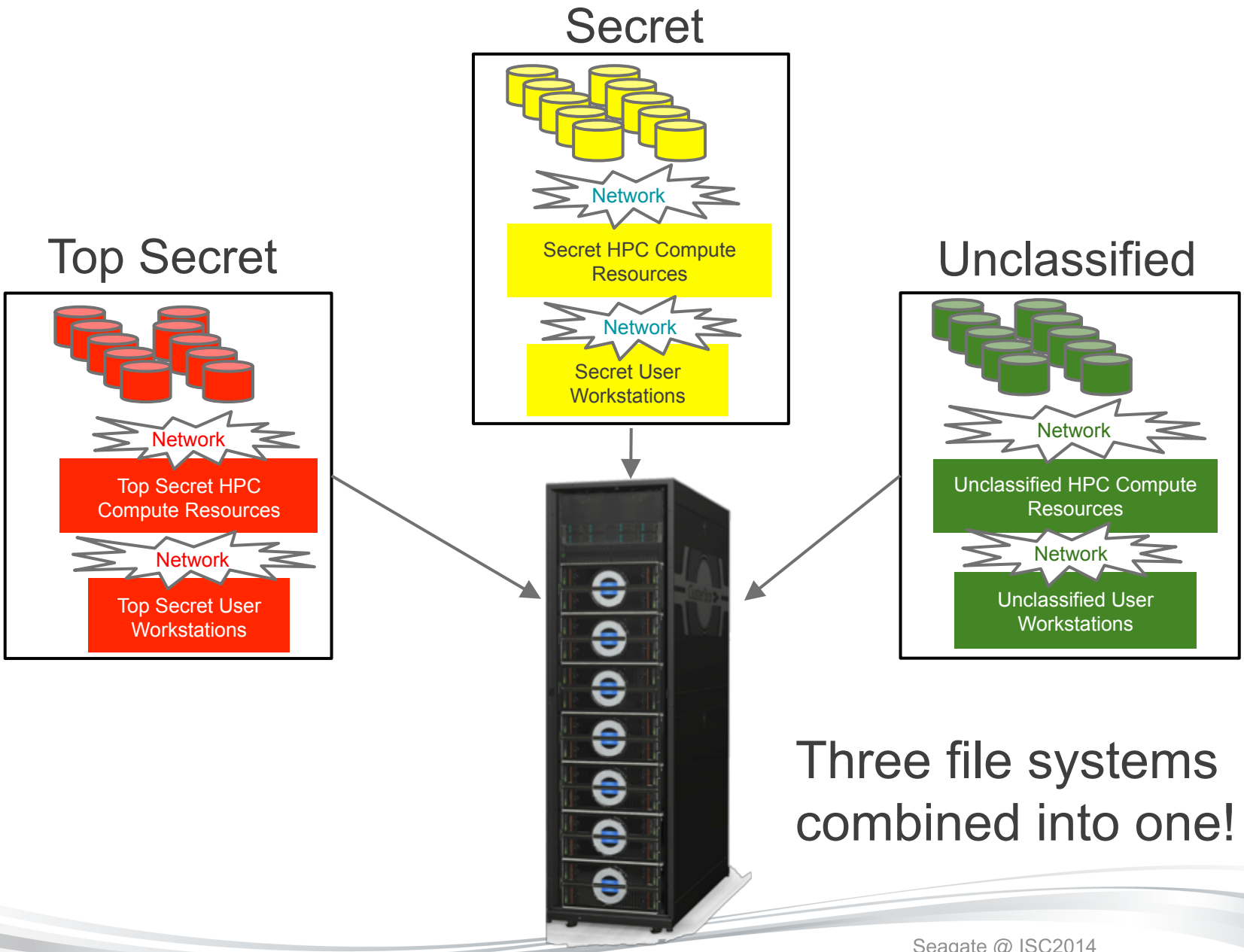
8.1.2. Environmental Requirements

Before installing the Lustre software, make sure the following environmental requirements are met.

- *(Required)* ***Disable Security-Enhanced Linux**** ***(SELinux) on all Lustre servers and clients.*** The Lustre software does not support SELinux. Therefore, the SELinux system extension must be disabled on all Lustre nodes. Also, make sure other security extensions (such as the Novell AppArmor* security system) and network packet filtering tools (such as iptables) do not interfere with the Lustre software.

- Options ??
  - Separate file systems based on different solutions ?
  - Create new version of Lustre ?
  - Make Lustre work ?

Customer Value Proposition

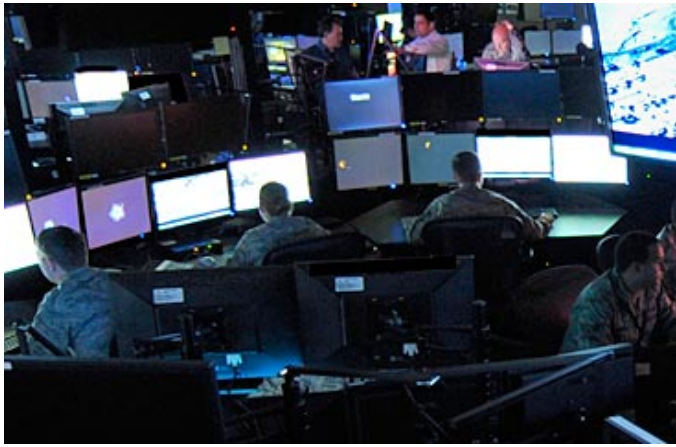


ClusterStor SDA Features

- **Industry first secure scale-out parallel file system designed to enable administrators to achieve ICD 503 (DCID 6/3 PL4) compliance**
  - Complete and explicit audit trails
  - Know who is doing what on the system
- **Predictable linear performance and storage capacity scale**
  - Linear scale-out file system performance over 1 terabyte per second
  - Single file system namespace over 90 petabytes usable data capacity
- **Integrated solution with end-to-end security administration, diagnostics and management**
  - Reduces operational and management cost
  - Relieves data center floor space, power and cooling constraints
  - Factory integrated, pre-configured, tested and supported by Xyratex

ClusterStor SDA Benefits

- **Enables Multilevel Security (MLS) while satisfying exploding need for massive scale**
  - **Overcome legacy barriers and bottlenecks**
  - **Unprecedented scale in performance and capacity**
  - **Industry unique end-to-end security administration solution with fully integrated diagnostics and management**
- **Greatly improve intelligence center security, productivity and efficiency**
  - **Protects against both external and internal threats**
  - **Significantly increase mission productivity and agility**
  - **Relieve floor space, power and cooling constraints**
  - **Reduce operational and management complexity**
- **Significantly reduce CapEx and OpEx cost**
  - **Reduce capital equipment acquisition cost and complexity**
  - **Reduce security administrative cost**



The way forward

Is good hardware enough ??

Current scalable file systems …

- Parallel filesystems
  - GPFS
  - Lustre
  - Ceph
  - BeeGFS
- Cloud solutions
  - Amazon S3
  - Google Cloud Storage
  - M$ Azure

… non of these will be able to scale sufficiently!

… but advanced middleware such as E10
can help extend the life of these solutions …

Object based storage – Next Gen

- A traditional file system includes a hierarchy of files and directories
- Accessed via a file system driver in the OS
- Object storage is “flat”, objects are located by direct reference
- Accessed via custom APIs
  - Swift, S3, librados, etc.
- The difference boils down to 2 questions:
  - How do you find files?
  - Where do you store metadata?
- Object store + Metadata + driver ***is a file system***

Object Storage Backend: Why?

- **It's more flexible.** Interfaces can be presented in other ways, without the FS overhead.
  A generalized storage architecture vs a file system
- **It's more scalable.** POSIX was never intended for clusters, concurrent access, multi-level caching, ILM, usage hints, striping control, etc.
- **It's simpler.** With the file system-"isms" removed, an elegant (= scalable, flexible, reliable) foundation can be laid

The way forward ..

- Object Storage based solutions offers a lot of flexibility:
  - Next-generation design, for exascale-level size, performance, and robustness
  - Implemented from scratch
    - » "If we could design the perfect exascale storage system..."
  - Not limited to POSIX
  - Non-blocking availability of data
  - Multi-core aware
  - Non-blocking execution model with thread-per-core
  - Support for non-uniform hardware
  - Flash, Solid State, non-volatile memory, NUMA ….
    - » Transparent !!
  - Using abstractions, guided interfaces can be implemented
    - » e.g., for burst buffer management (pre-staging and de-staging).

Solutions …

- Size does matter …..

  2014 – 2016 >20 proposals for 40+ PB file systems

  Running at 1 – 4 TB/s !!!!
- Heterogeneity is the new buzzword
  - Burst buffers, data capacitors, cache off-loaders …
- Mixed workloads are now taken seriously ….
- Data integrity is paramount
  - T10-PI/DIX is a decent start but …
- Storage system resiliency is equally important
  - PD-RAID need to evolve and become system wide

Transparent multi-tier storage is absolute key

Seagate

Thank you
for listening to a
madman's ramblings …