



Seagate ExaScale HPC storage ...

Possibility or pipedream ?

Torben Kling Petersen, PhD
Principal Engineer
Seagate Systems Group, HPC



Defining ExaScale HPC Storage

Systems (J. Dongarra, 2009)	2009	2018
System Peak	2 Pflop/sec	1 Eflop/sec
Power	6 MW	~20 MW
System Memory	0.3 PBs	32 - 64 PBs
Node Compute	125 Gflop/s	1,2 or 15 Tflops/s
Node Memory BW	25 GB/s	2 - 4 TB/s
Node Concurrency	12	1,000 – 10,000
Total Node Interconnect BW	3.5 GB/s	200 - 400 GB/s (1:4 or 1:8 from memory BW)
System Size (Nodes)	18,700	O(billion) [O(10) to O(100) for latency hiding]
Total Concurrency	225,000	1,000,000,000
Storage	15 PB	500-1000 PB (>10x system memory is min)
I/O	0.2 TB	60 TB/s (how long to drain the machine)
MTTI	Days	Minutes

What does that mean in today's numbers ?

Systems (T Kling Petersen, 2016)	2018	Today
Storage	1000 PB	10 TB NL-SAS 120,000 HDDs (RAID6) 232 racks @ 3,5 MW
I/O	60 TB/s	500 racks ~ 6,000 EDR ports
MTTI	Minutes	Days

Obvious Questions: Compute would require 30x the current no 1 ??
Using more than 500 MW !!

What file system ? Lustre ?? GPFS ?? Ceph?? DAOS ??

Are Flash technologies going to save the day ??

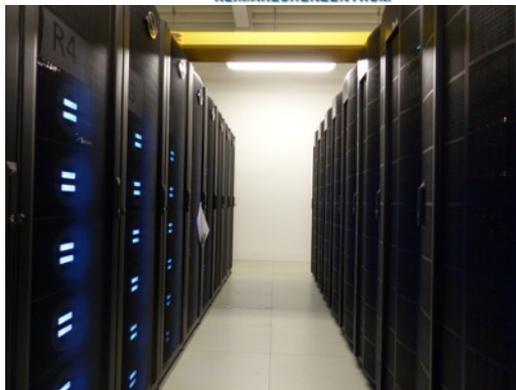
What applications could use these capabilities ??

How to manage and support systems of this size ??

... And as Storage is the **CRITICAL** building block?

Component	Technology	Requirements
Networks	HDR/OPA phase 2	Storage is the ONLY component that saturates 100 Gbit today
Flash tier	Burst buffers	Balanced I/O (R/W) Intelligent data placement/movement
Capacity tier	Faster/larger HDDs	Near line storage HAVE to keep up with the flash tier
Archiving	Tape replacement	Archives cannot be 1000x slower than capacity tiers ...
Data mgmt	File systems etc	Semi automated data management can not keep up with the requirements !!
Reliability	MTTI	Current enterprise features cannot deliver the required system reliability

Powered by  SEAGATE



55 PB Lustre File System
~500 GB/s Lustre File System



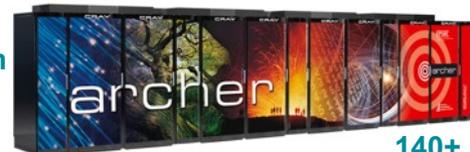
500+ GB/s Lustre File System



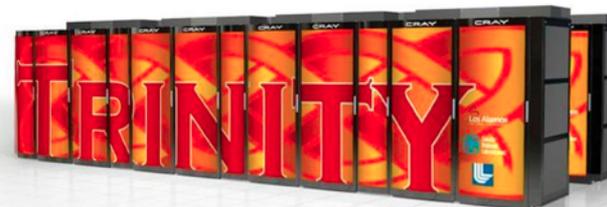
130+ GB/s Lustre File System



20 PB Lustre File System
1+ TB/s aggregate I/O



140+ GB/s Lustre File System



1.6 TB/sec Lustre File System



1 TB/sec Lustre File System

Real storage leadership



Rank	Name	Computer	Site	Total Cores	Rmax (TFLOPS)	Rpeak (TFLOPS)	Power (KW)	File system	Size	Perf
1	Tianhe-2	TH-IVB-FEP Cluster, Xeon E5-2692 12C 2.2GHz, TH Express-2, Intel Xeon Phi	National Super Computer Center in Guangzhou	3120000	33,862,700	54,902,400	17808	Lustre / H2FS	12.4 PB	~750 GB/s
2	Titan	Cray XK7 , Opteron 6274 16C 2.2GHz, Cray Gemini interconnect, NVIDIA K20x	DOE/SC/Oak Ridge National Laboratory	560640	17,590,000	27,112,550	8209	Lustre	10.5 PB	240 GB/s
3	Sequoia	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom Interconnect	DOE/NNSA/LLNL	1572864	17,173,224	20,132,659	7890	Lustre	55 PB	850 GB/s
4	K computer	Fujitsu, SPARC64 VIIIfx 2.0GHz, , Tofu interconnect	RIKEN AICS	705024	10,510,000	11,280,384	12659	Lustre	40 PB	965 GB/s
5	Mira	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	DOE/SC/Argonne National Lab.	786432	8,586,612	10,066,330	3945	GPFS	28.8 PB	240 GB/s
6	Trinity	Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect	DOE/NNSA/LANL/SNL	301056	8,100,900	11,078,861		Lustre	76 PB Powered by SEAGATE	1,600 GB/s
7	Piz Daint	Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x	Swiss National Supercomputing Centre (CSCS)	115984	6,271,000	7,788,853	2325	Lustre	2.5 PB Powered by SEAGATE	138 GB/s
8	Shaheen II	Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect	KAUST, Saudi Arabia	196,608	5,537,000	7,235,000	2,834	Lustre	17 PB Powered by SEAGATE	500 GB/s
9	Hazel Hen	Cray XC40, Xeon E5-2680v3 12C 2.5GHz, Aries interconnect	HLRS - Stuttgart	185088	5,640,170	7,403,520		Lustre	7 PB Powered by SEAGATE	~100 GB/s
10	Stampede	PowerEdge C8220, Xeon E5-2680 8C 2.7GHz, IB FDR, Intel Xeon Phi	TACC/ Univ. of Texas	462462	5,168,110	8,520,112	4510	Lustre	14 PB	150 GB/s

n.b. NCSA Bluewaters 24 PB 1100 GB/s (Lustre 2.1.3)

The Concept: Fully integrated, fully balanced, no bottlenecks ...

ClusterStor Scalable Storage Unit

- Intel Ivy bridge or Haswell CPUs
- EDR, 100 GbE & 2x40GbE, all SAS infrastructure
- SBB v3 Form Factor, PCIe Gen-3
- Embedded RAID & Lustre support

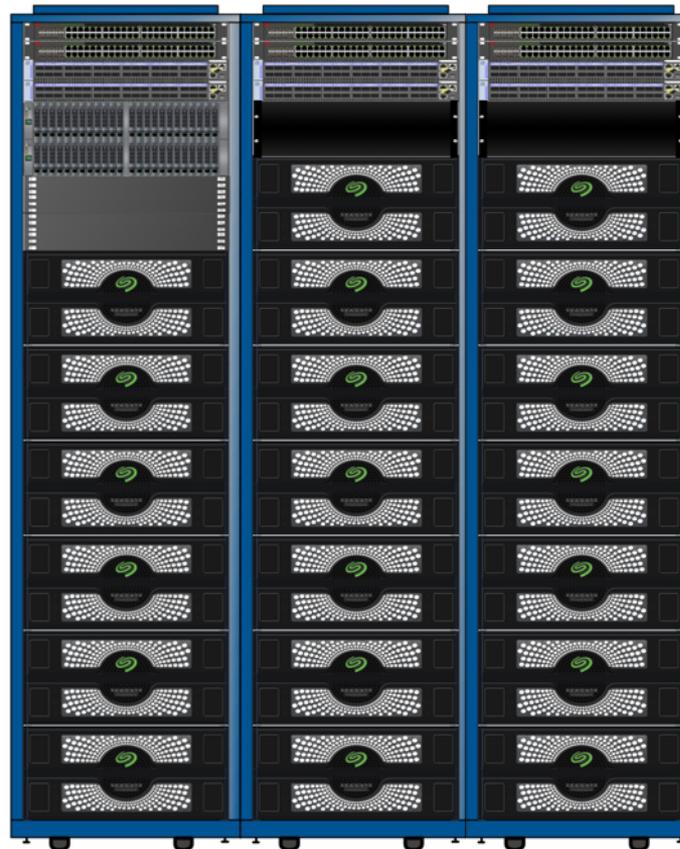
ClusterStor/Sonexion Manager

Lustre 2.5 / 2.7
IBM Spectrum Scale 4.2

Data Protection Layer
(PD-RAID/Grid-RAID)

Linux OS

Unified System Management
(GEM-USM)





Lustre solutions



Seagate and Intel join forces on Lustre®

Agreement signed February 19



- Seagate are transitioning from OpenSFS to Intel IEEL Lustre as the baseline
 - Beginning with the Lustre 2.7 release planned for 2H 2016
 - Seagate distribution will contain our specific Lustre features including more than 260 patches mainly focused on running Lustre at extreme scale
 - Lustre on ClusterStor/Sonexion is a super set distribution
- Seagate Lustre Dev and Support team will continue to support our customers
 - Largest support capability in the Industry (Intel's Lustre team + Seagate Lustre team)
 - Seagate support will work with customers and escalate any IEEL issues to Intel.
- Seagate will continue to improve Lustre
 - Continue to test and improve the quality of Lustre 2.7+ particularly at scale
 - Seagate will develop some unique Lustre features

CS-3584 - Scalable Storage Unit (SSU) – OSS/NSD

- Ultra HD - CS-3584 SSU – dual OSS or NSD
 - 5U84 Enclosure – completely H/A
 - Two (2) trays of 42 HDD's each with 12 Gbit SAS
 - Dual-ported 3.5" NL SAS & SSD HDD Support
 - 300+ MB/s SAS available bandwidth per HDD
 - Pair of H/A Embedded Application Servers
 - L300 = 12 - 18 GB/sec IOR over IB
 - IB F/EDR or 40/100 GbE Network Link
 - Data Protection/Integrity (Grid-RAID, 8+2)
 - Grid-RAID - 2 OSS's per SSU, 1 OST's per OSS
 - 2x SSD OSS journal disks for increased performance
 - 64 Usable Data Disks per SSU
 - 2 – 8 TB drives supported



Only 5° C delta
with drawer open



Embedded
server modules

New Platform 300 Embedded Application Server

Mezz/daughter
slot/connector

PCI HBA
(EDR/Omni)
Slot/connector

12Gbit SAS
mezz/daughter
card
installed

EDR HBA
Installed

New Object Storage Server/NSD

- PCI Slot for Network HBA
- Intel Omni-Path or Mellanox EDR



ClusterStor Grid-RAID Declustered Parity - Geometry

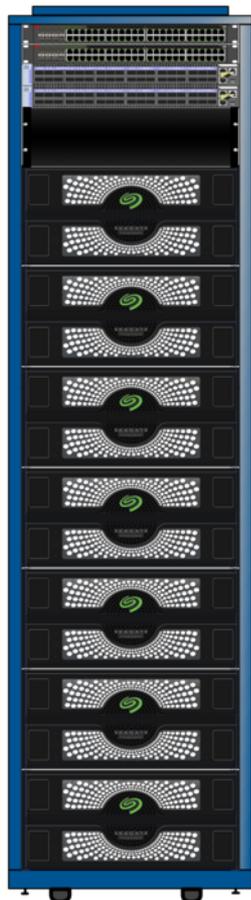
- PD RAID geometry for an array is defined as:
 - P drive (N+K+A)
 - example: 41 (8+2+2)
- P = total number of disks in the array
- N = number of data blocks per stripe
- K = number of Parity blocks per stripe
- A = number of distributed spare disk drives
- Benefits:
 - Balanced disk usage within an array
 - 1 OST per OSS (less context switching etc)
 - Much faster re-builds (< 2 hours with HPC drives)
 - Performance benefits vs MD-RAID

1	2	2	1	2	2	0	3	1	2	5	0	2	1	1	2	1	3	3	8	0	3	3	0	3	1	2	1	0	3	0	3	0	3	0	1	2	1	0	2	0					
5	6	6	5	6	6	4	7	4	6	5	4	6	5	6	5	6	5	6	5	7	7	8	3	7	7	4	7	5	6	4	6	4	7	4	7	4	5	5	4	6	4				
9	9	10	9	10	10	8	11	8	10	8	10	9	9	10	9	11	11	8	11	8	10	9	7	10	8	10	8	10	8	9	9	9	9	8	10	8	9	8	10	8					
13	13	13	13	14	14	12	15	12	14	5	12	14	13	13	14	12	15	15	5	11	15	15	12	14	12	14	13	11	14	12	14	12	15	11	13	13	13	12	14	12					
16	17	17	17	18	18	16	19	16	18	5	16	18	17	17	18	16	19	18	5	15	19	19	16	18	16	17	17	15	18	15	18	16	19	15	17	17	17	16	18	16					
20	21	21	21	22	22	20	22	20	21	5	19	22	21	21	22	20	23	22	5	19	23	23	20	21	20	19	22	19	22	20	23	19	21	21	21	20	22	20							
24	25	25	25	26	26	24	26	24	25	5	23	26	24	25	26	24	26	26	5	23	27	24	24	26	24	25	24	23	26	23	26	23	27	23	25	26	25	24	26	24					
28	29	29	29	30	30	28	30	28	29	5	27	30	28	29	29	28	30	30	5	27	30	31	28	30	28	29	28	27	30	27	31	27	29	29	28	28	29	27							
32	33	33	32	34	34	32	34	32	33	5	31	33	32	33	33	32	34	34	5	31	34	34	32	34	32	33	32	31	34	31	34	31	35	31	33	33	32	31	33	31					
36	37	37	36	38	37	35	38	36	37	5	35	37	36	37	37	36	38	38	5	35	38	38	36	38	36	37	36	37	36	38	35	38	35	38	35	36	37	36	35	37	35				
40	39	42	40	40	39	42	42	41	40	39	41	41	42	39	5	41	41	41	39	42	40	42	40	42	39	41	41	42	39	40	40	39	41	39	40	42	39	5	41	40					
44	43	46	44	43	42	46	46	45	44	43	45	45	45	43	5	45	45	44	43	46	44	46	44	46	43	45	45	46	43	44	44	43	45	43	44	46	43	5	45	44					
48	47	50	47	47	46	50	49	48	47	49	48	49	47	5	49	49	48	46	49	48	50	48	47	49	49	50	47	48	48	47	49	47	49	50	47	47	49	48	5	49	48				
51	51	54	51	51	50	53	54	53	52	51	53	52	53	51	5	53	53	52	50	53	52	54	52	54	51	53	53	54	51	52	52	51	53	50	52	54	51	5	52	52					
55	58	58	55	55	54	57	58	57	56	55	57	56	57	54	5	57	57	56	54	57	56	58	56	58	55	57	56	58	55	56	55	56	57	56	58	55	56	57	55	5	56	56			
59	58	62	59	59	58	61	62	61	60	59	61	60	61	58	5	60	61	60	58	61	60	62	60	62	60	61	59	61	60	62	59	60	59	61	58	59	61	59	61	59	5	60	60		
63	62	66	63	63	62	65	66	65	64	63	65	64	65	62	5	64	65	64	62	65	63	65	64	65	63	65	64	66	63	64	63	63	64	62	63	65	62	5	64	64					
67	66	70	67	67	66	69	70	68	68	66	69	68	69	66	5	68	69	68	66	69	67	69	67	69	67	68	68	67	68	67	68	66	67	68	66	67	68	66	5	68	68				
71	70	73	71	71	70	73	74	72	71	70	73	72	73	70	5	72	72	72	70	73	71	73	71	73	71	73	71	73	72	73	70	72	71	71	72	70	71	73	70	5	72	72			
75	74	77	75	75	74	77	77	76	75	74	77	76	77	74	5	76	76	76	74	77	75	77	75	77	75	77	75	77	74	75	75	74	76	74	75	77	74	75	77	74	5	76	76		
79	80	80	81	80	78	78	81	78	81	79	79	79	80	79	5	80	80	81	5	81	78	80	79	80	81	80	78	81	78	79	78	78	80	81	79	81	78	78	79	5	80	80			
83	84	84	85	84	82	82	85	82	85	83	83	83	83	83	5	84	84	85	5	85	82	84	82	84	85	84	82	84	81	83	82	82	84	85	83	85	82	82	83	5	84	84			
87	88	88	89	88	86	86	89	86	88	87	87	87	87	87	5	88	88	89	5	89	86	88	86	88	89	88	86	87	88	86	87	88	85	87	85	86	88	86	89	86	86	87	5	88	88
91	92	92	93	91	90	93	90	92	91	90	91	91	91	91	5	92	92	93	5	93	89	92	90	91	93	92	90	92	89	91	89	90	92	93	90	92	90	91	5	92	92				
95	95	96	95	94	94	97	94	95	94	94	95	95	95	95	5	96	96	97	5	97	93	96	94	95	97	96	94	95	93	94	96	93	94	96	93	94	95	5	96	96					
99	99	100	100	99	98	98	100	98	100	99	98	98	98	99	5	100	100	101	5	101	97	100	98	99	101	99	100	97	99	97	98	100	101	98	100	97	99	5	100	100					
103	103	103	104	103	102	102	104	102	104	102	102	102	102	103	5	104	104	105	5	105	101	104	102	103	104	103	101	104	101	103	101	102	104	105	102	104	101	101	5	103	103				
107	107	107	108	107	105	106	108	106	108	106	106	106	106	107	5	108	108	109	5	109	105	108	106	107	108	105	107	106	108	105	107	106	108	105	107	106	108	105	107	5	107	107			
111	111	111	112	111	109	110	112	110	112	110	110	110	110	111	5	112	112	113	5	112	109	111	110	111	112	111	109	112	109	111	109	109	111	112	110	112	109	109	5	111	111				
115	115	115	116	115	113	113	116	114	116	114	114	114	114	115	5	115	116	116	5	116	113	115	116	115	116	113	114	113	113	115	116	114	113	113	115	116	114	113	5	115	115				

PDRAID [41 (8+2+2), 3 Tiles: Permuted Layout.

L300 - File system performance Rack Aggregates/Totals

Expansion racks



	# drives: (HDDs/SSDs)	8TB HDD TBs: (U/R)	IOR perf GB/s*	Power kW
SSU #6	574/ 14	3580 / 4592	Up to 84	14.9
SSU #5	492 / 12	3072 / 3936	Up to 72	12.6
SSU #4	410 / 10	2560 / 3280	Up to 60	10.9
SSU #3	328 / 8	2048 / 2624	Up to 48	9.2
SSU #2	246 / 6	1536 / 1968	Up to 36	7.4
SSU #1	164 / 4	1024 / 1312	Up to 24	5.7
SSU #0	82 / 2	512 / 656	Up to 12	4.0

A close-up, artistic photograph of a hard drive's internal components. The central focus is a single, highly reflective silver platter. A read/write head is positioned just above the surface of the platter, with a thin layer of air between them. The background shows other platters and the complex mechanical structure of the drive, all bathed in a cool blue light with some red highlights from the drive's internal LEDs.

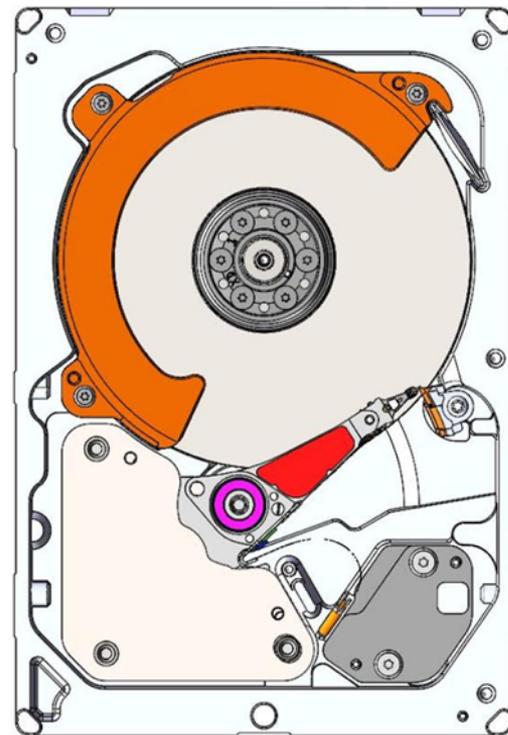
Platform 300
HPC Disk Drive
HAMR tech



Enterprise Performance 3.5 HDD

High level product description

- 4TB, 10K RPM, 5D, 3.5" FF HDD
- Performance increases across the board vs. 7200 RPM
 - Large block & small block
 - Random & sequential
 - Reads & writes
- 2M hr MTBF and 750 TB/yr workload ratings
- Targeting ~13W max. typical operating power
 - PowerBalance™ setting for ~2W lower available
- Configuration: 4Kn with 12Gb/s SAS SED
 - Seeding market with initial product offering
- Available with Seagate ClusterStor NOW



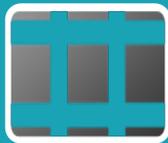
ClusterStor L300 HPC 4TB SAS HDD

HPC Industry First; Best Mixed Application Workload Value



Performance Leader

World-beating performance over other 3.5in HDDs: *Speeding data ingest, extraction and access*



Capacity Strong

4TB of storage for big data applications



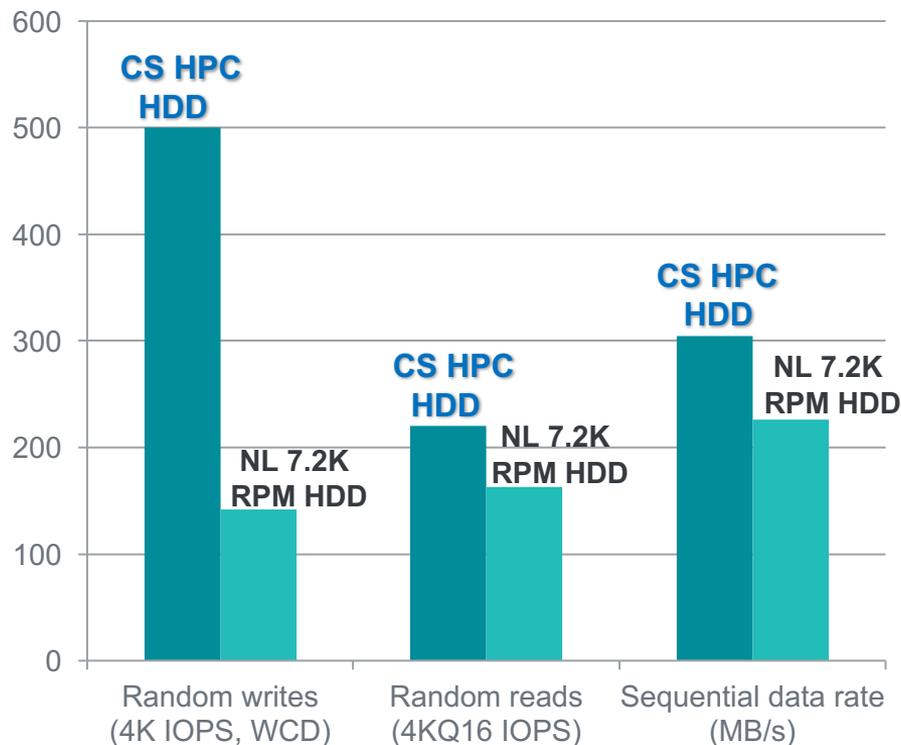
Reliable Workhorse

2M hour MTBF and 750TB/year ratings for reliability under the toughest workloads your users throw at it



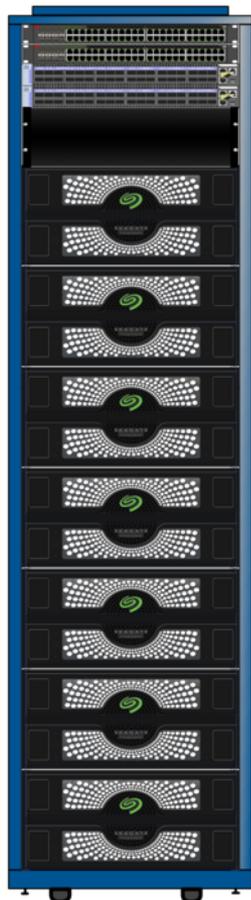
Power Efficient

Seagate's PowerBalance feature provides significant power benefits for minimal performance tradeoffs



L300 - File system performance Rack Aggregates/Totals

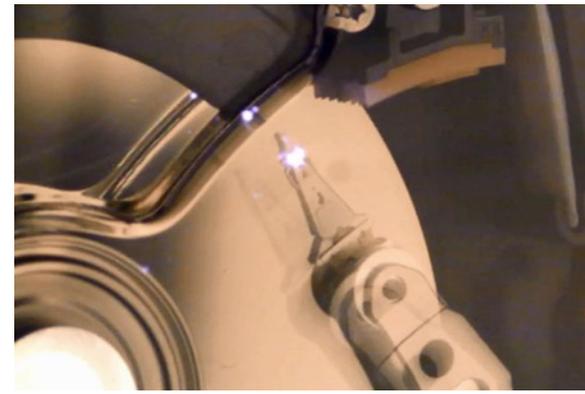
HPC drive base



	# drives: (HDDs/SSDs)	4TB HDD TBs: (U/R)	IOR perf GB/s*	Power kW
SSU #6	574 / 14	1792 / 2240	Up to 126	14.9
SSU #5	492 / 12	1536 / 1920	Up to 108	12.6
SSU #4	410 / 10	1280 / 1600	Up to 90	10.9
SSU #3	328 / 8	1024 / 1280	Up to 72	9.2
SSU #2	246 / 6	768 / 960	Up to 54	7.4
SSU #1	164 / 4	512 / 640	Up to 36	5.7
SSU #0	82 / 2	256 / 320	Up to 18	4.0

Hard drive futures ...

- HAMR drives (Seagate)
 - Using a laser to heat the magnetic substrate (Iron/Platinum alloy)
 - Possible capacity – 15 - 30 TB/ 3.5 inch drive ...
 - 2016 timeframe (first shipments)
- 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
 - Theoretical capacity – 100+ TB / 3.5 inch drive ...





The “Data Capacitor”

Utilizing
Seagate Flash technology



Seagate 1200.2 SAS SSD technology



Enterprise-focused Feature Set

- Enterprise Grade Performance & Features
 - 24Gb/s Active-Active (High I/O performance)
 - Wide capacity range (200GB to 4TB-class) with multiple endurance options in one platform
 - Multi-host, dual port supports
“No Single Point of Failure”
- Enterprise Grade Data Protection
 - T10-DIF End-to-End ECC Internal and External
 - No danger of ‘Silent Data Corruption’
 - Power loss data protection (PLDP) provides mechanism to save data/operations in process
 - Encryption to NSA Standard, SED and FIPS-compliance prevents unauthorized access to stored data
- Enterprise Grade Endurance
 - 5-Year Drive Life Even Under Write-Intensive Workloads



Best Fit Applications

Server Virtualization

Examples: VMware vSphere, Microsoft Hyper-V, Linux KVM, Zen



Databases

Examples: OLTP, Oracle, SAP, SQL-Server, Exchange, NO-SQL, MySQL, MongoDB



HPC applications

Examples: Lustre, Spectrum Scale, BeeGFS, etc ...

Software Defined Storage

Examples: Microsoft Storage Spaces, Nexenta, VMware vSAN



Nytro[®] PCIe Flash Accelerator Cards



Lowest Latency and Highest Efficiency

- Latency-Optimized
 - Controller with DRAM for minimized latency
 - Consistently high performance and low latency
- Density-Optimized
 - Maximum capacity & performance within a form factor
 - Performance scales with Queue Depth / Thread Count
 - IO intensive and virtualized workloads
- Thermally-Optimized
 - Single-planar, NAND-down design for optimal cooling
 - Read-intensive and power / thermally sensitive applications



XP6500

10 GB/s W/R !!



XP6302



XP6209 &
XP6210

Best Fit Applications

Transactional DB



Virtualized and IO-intensive



Dense environments



Enterprise



Nytro[®] XF1440 / XM1440 PCIe SSDs



Balanced Power and Performance

Innovative Data Center Storage Solutions

- PCIe Gen3 delivers higher sustained transfer speeds
- NVMe protocol for consistent response time
- Multiple form factors: SFF 2.5" 7mm and M.2
- Addressing read intensive and mixed workloads



Reducing Total Cost of Ownership

- \$/Watt cost advantage
- Power/performance optimized solutions (<12.5W)

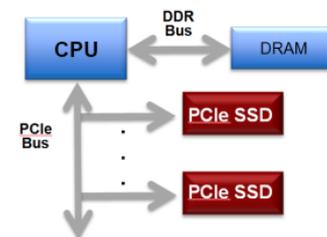
Delivering Enterprise Class Features

- TCG enterprise security
- Instant Secure Erase
- Power loss data protection
- Hot plug capability on SFF 2.5"
- Robust test infrastructure

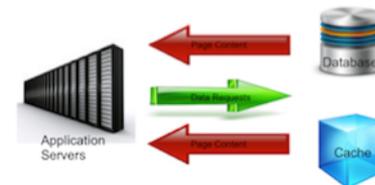


Best Fit Applications

Direct Attached Storage



Caching



Tiering

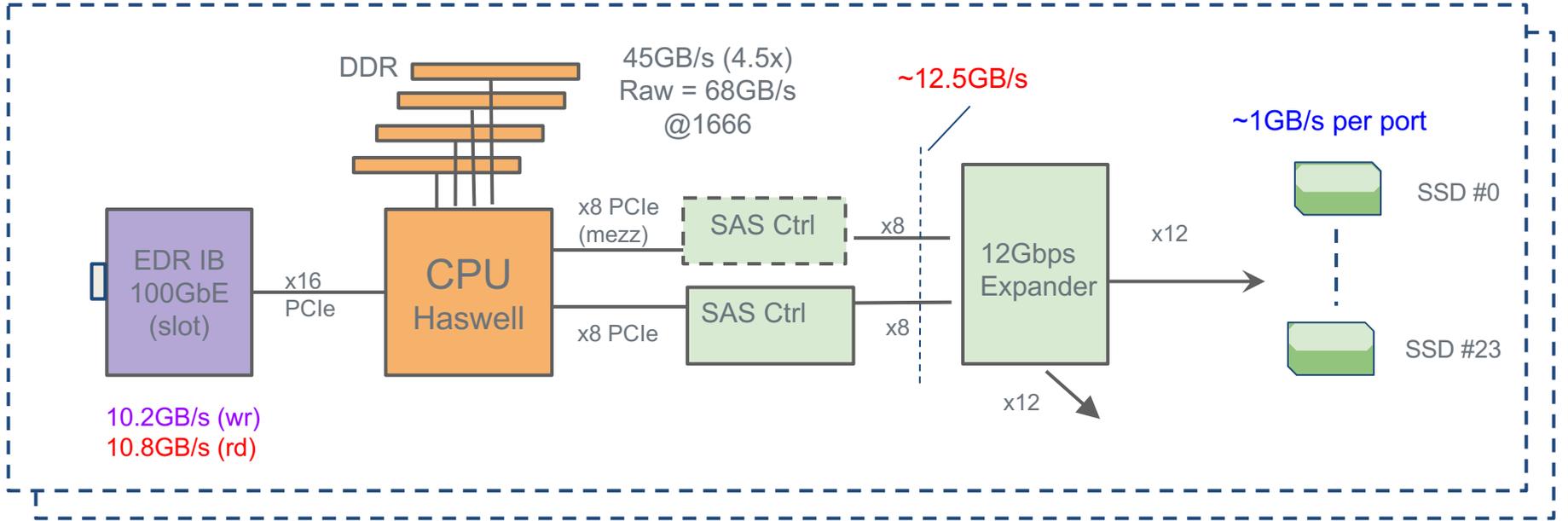


Flash tier - “The Data Capacitor” concept

- Seagate enclosures
 - OneStor 2U24 – 12G SAS
- Laguna Seca EAMs
 - Single Socket CPUs (Haswell)
 - 4 DIMMs per CPU
 - EDR/OmniPath support
- Next gen SAS SSDs
 - Capacity up to 15.4 TB
 - DWPD ~1 to 3
 - Up to 20 GB/s per enclosure
- Dedicated OSS/NSD pair



Dataflow – "The Data Capacitor"



10.2GB/s (wr)
10.8GB/s (rd)

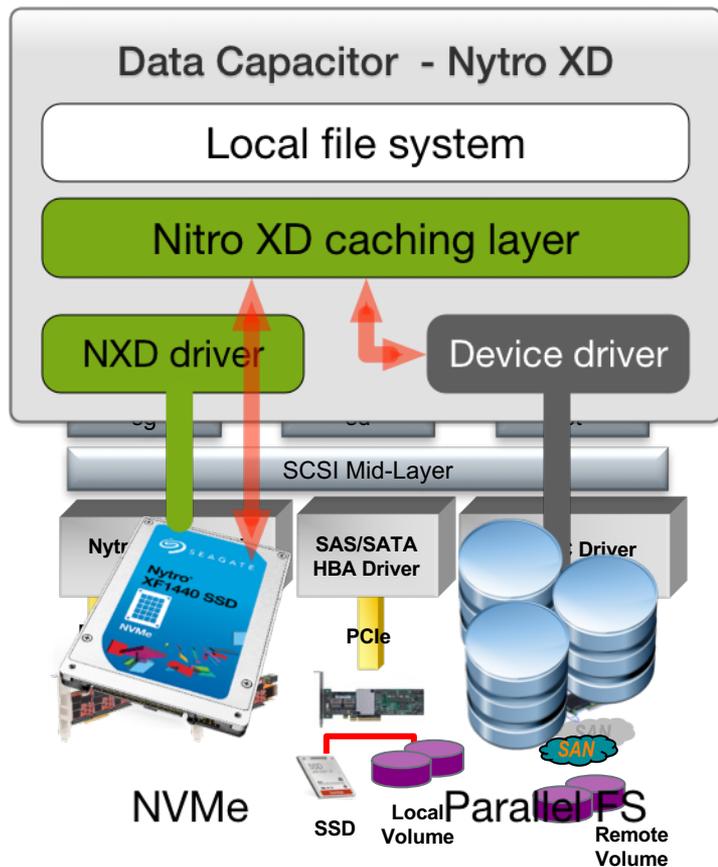
x2 for SSU



~20 GB/s per 2 RU
(over 2 EDR IBs)

Nytro XD Architecture – Data capacitor concept

Linux Driver Architecture



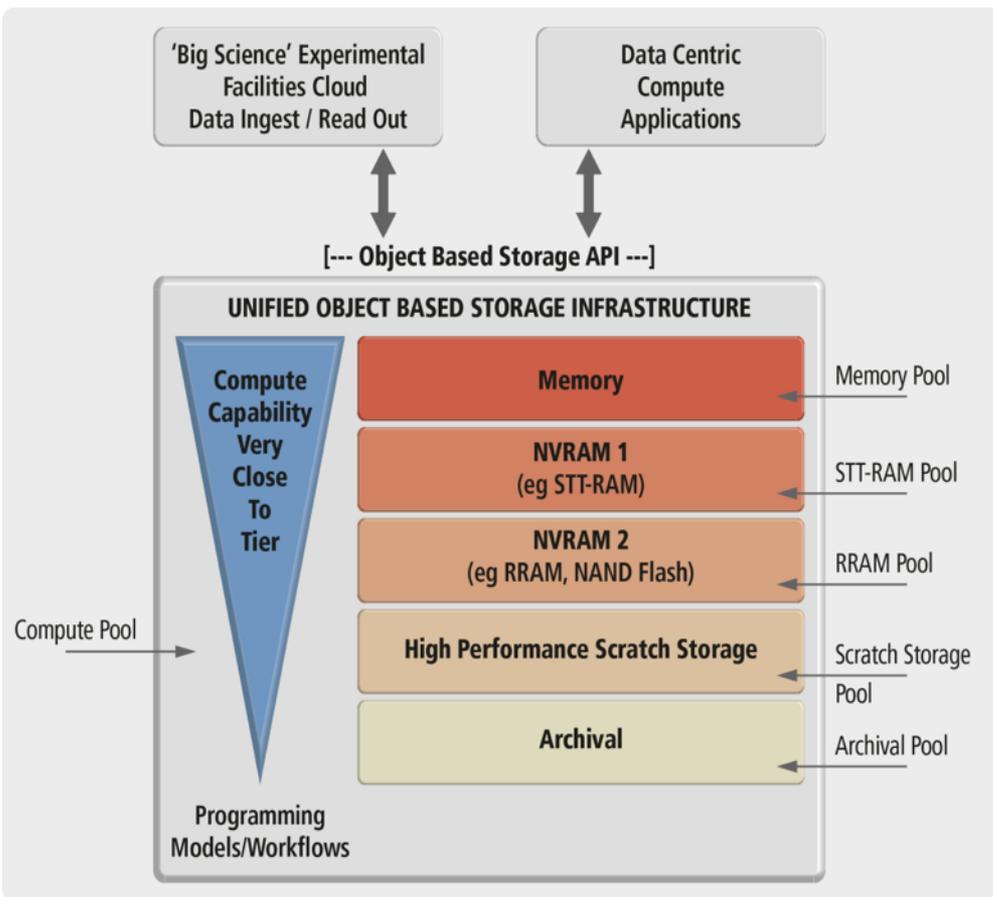
- Filter driver and OS dependent functions implemented as device mapper target driver
- Core caching library compiled as a Linux kernel module with well defined APIs
- Work at the block layer be transparent to file system and applications
- Hardware agnostic, can work with any block device
- Consumes Flash devices and provides Caching function across DAS/SAN volumes
- Core caching function is implemented as an OS agnostic portable library with well defined interfaces
- Filter Driver in OS stack intercept's IO and routes through Cache Management Library for Caching functions



Object Storage based archiving solutions



SAGE - Percipient Storage Overview



- **Goal**

- Build the data centric computing platform

- **Methodology**

- Commodity Server & Computing Components in I/O stack
- New NVRAM Technologies in I/O stack
- Ability for I/O to Accept Computations
 - Incl. Memory as part of storage tiers
- API for massive data ingest and extreme I/O

ClusterStor A200 Active Archive Product Overview

Combined with ClusterStor HSM or TSM to provide automatic policy-driven data migration & retrieval

Object API & portfolio of network based interfaces (POSIX, pNFS, CIFS, S3, HDF5, non-POSIX ...)

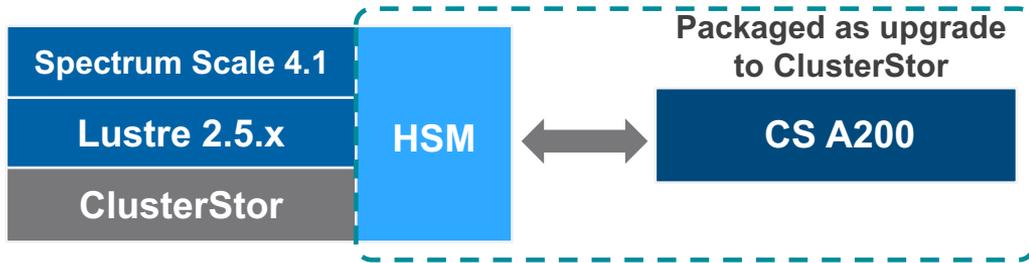
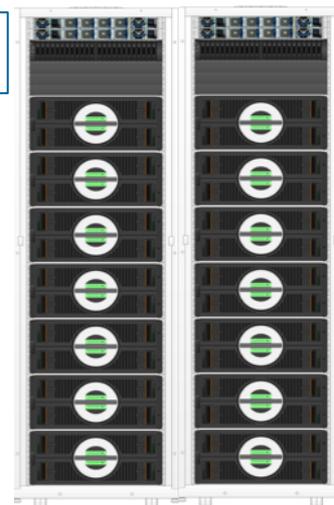
Unlimited scalability (file system size up to 2^{214} bytes)
High density storage up to 3.6PB* **usable** per rack

Utilizes **network erasure coding** to provide high levels of data availability and data durability

No single points of failure, resilient across single maintenance events

Dual 10Gb Ethernet node connectivity
IB as an option

ClusterStor A200



* moving to 5+ PB/rack in late 2016

Seagate innovation - SMR drives

Backed by Seagate Object store

SMR Drives

Shingled Technology increases capacity of a platter by 30-40%

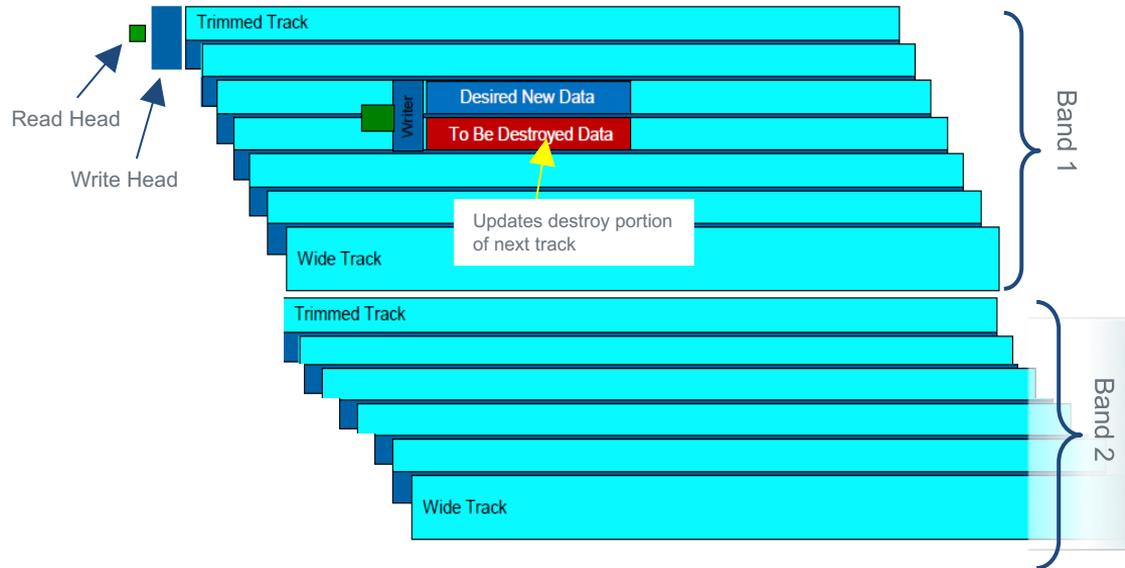
- › Write tracks are overlapped by up to 50% of write width
- › Read head is much smaller & can reliably read narrower tracks

SMR Drives are optimal for object stores as most data is static/WORM

- › Updates require special intelligence and may be expensive in terms of performance
- › Wide tracks in each band are often reserved for updates

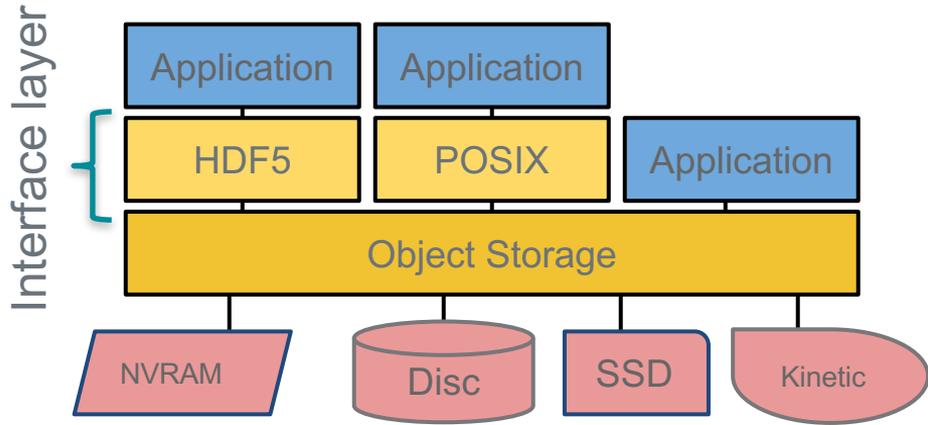
CS A200 manages SMR Drives directly to optimize workflow & caching

- › A200 avoids the "Read-Update-Write" problem by using **Copy-On-Write !!**

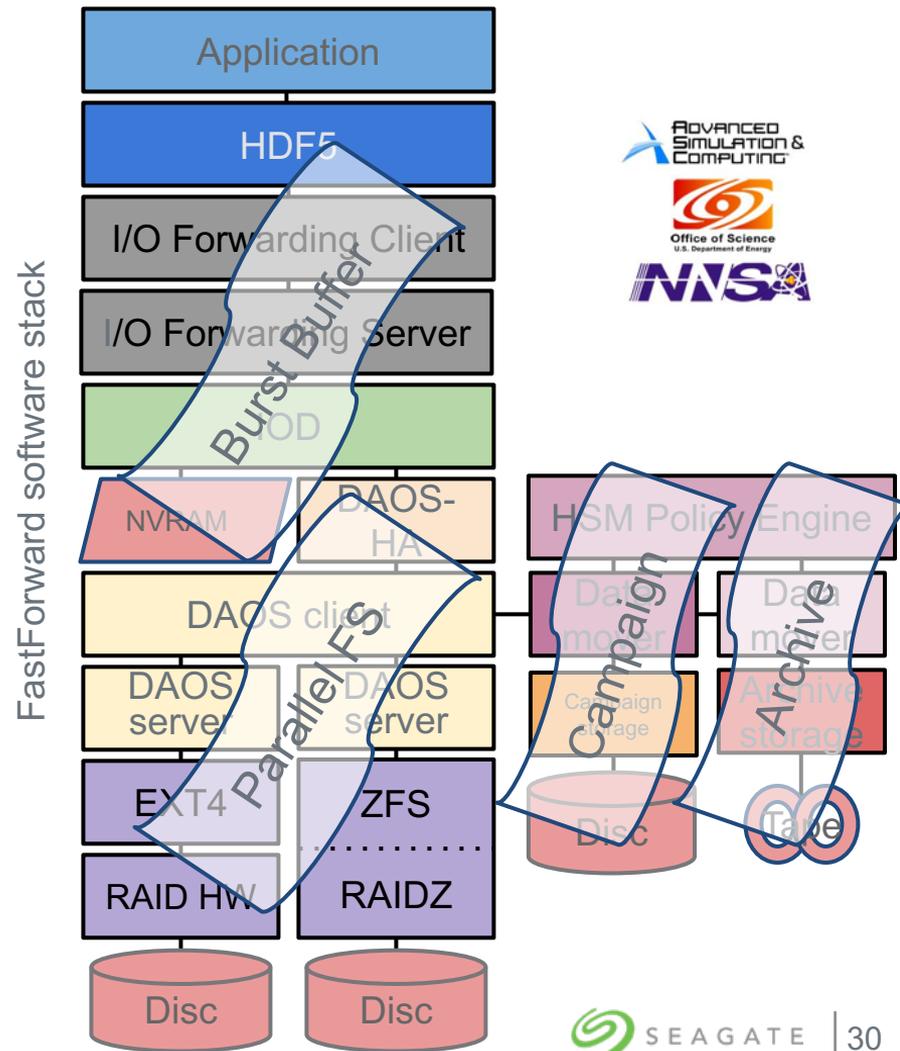


ExaScale Storage

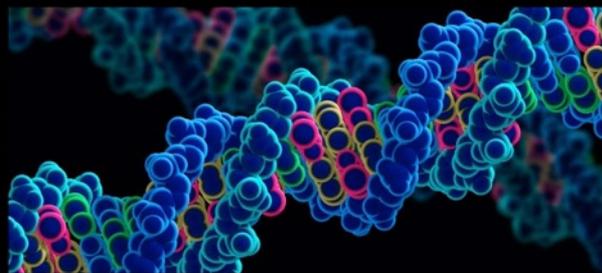
Simplifying the software stack



SAGE software stack



Seagate is HPC Storage



Unmatched speed and efficiency from the
Trusted Leader in HPC storage

