



Cray Storage Update LUG 2018
Nathan Schumann, Sr. Product Manager



Safe Harbor Statement



This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Cray and ClusterStor



Alignment

Commitment

Integration

COMPUTE

STORE

ANALYZE

Cray ClusterStor Platforms



ClusterStor L300



- Lustre 2.7
- 12 Gbit SAS enclosures
- Broadwell based ESMs
- EDR IB
- OPA
- 100 GbE
- 2x 40 GbE
- 6/8/10 TB HDDs

Large sequential I/O workloads

ClusterStor L300N



- NXD I/O Manager
- 3.2 TB SSDs
- Advanced MMU

Many applications with mixed I/O patterns

COMPUTE

STORE

ANALYZE

ClusterStor L300N: Any Workload, Any Time



Goal: Support Widest Application I/O Variety Within Budget

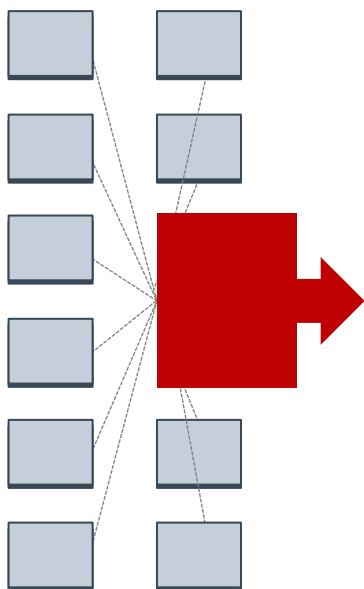
Challenge: Lost Productivity Due To Poor I/O Predictability

Solution: Smart I/O Management Provides Consistent Performance

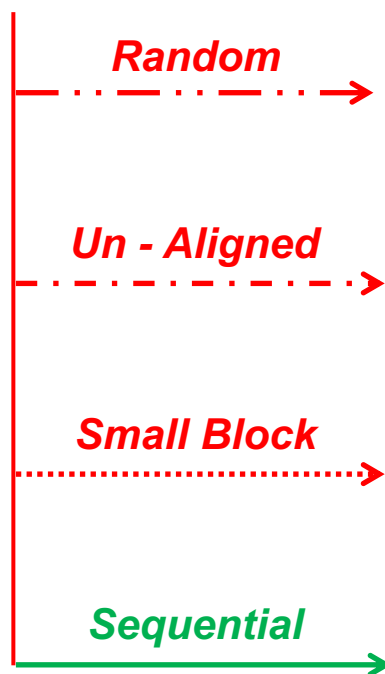
Result: I/O Optimized For Each Application I/O Profile

Compute Cluster Applications

*Sensor Data Capture & Processing
Data mining - Simulations
Modelling - Software development
Visualization of complex data -
Rapid mathematical calculations*



Mixed I/O Patterns =
Unpredictable Application Performance



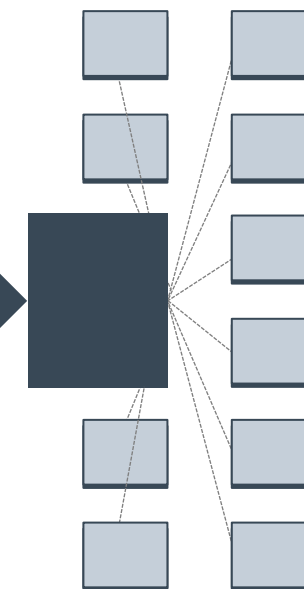
ClusterStor NXD =

Transparent Redirection of I/O to Appropriate SSD or HDD Medium



ClusterStor NXD =

Predictable and Fastest Compute Application Time-to-Solution

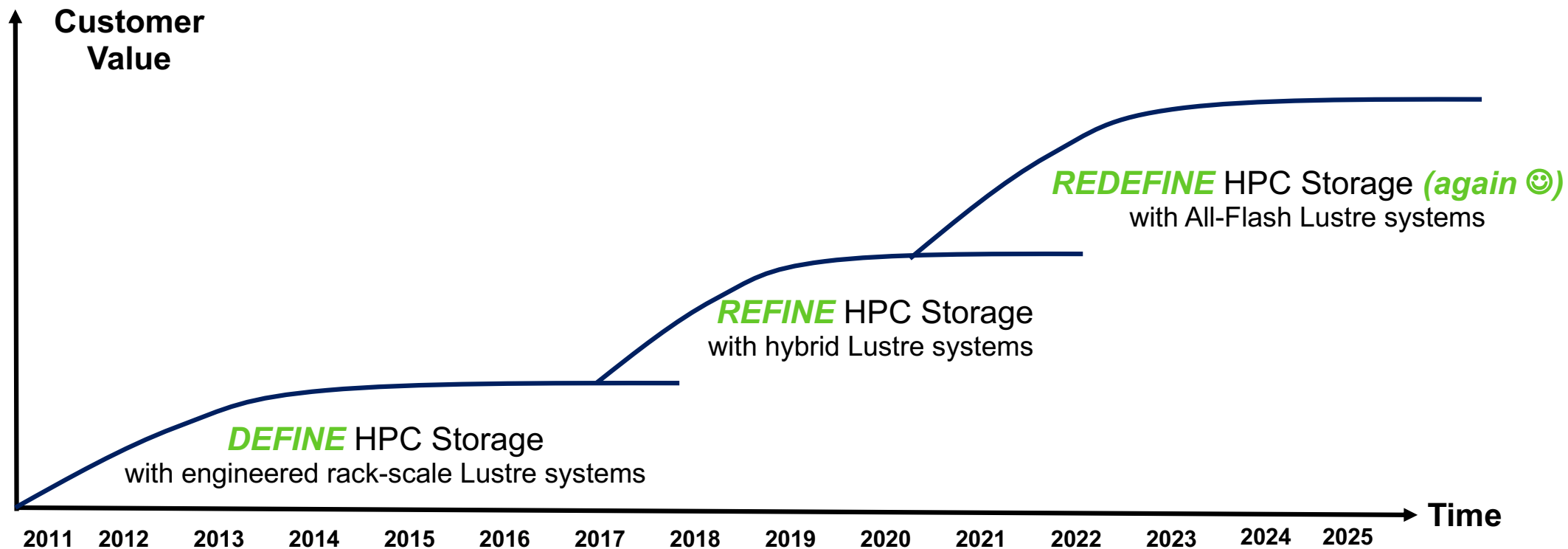


COMPUTE

STORE

ANALYZE

Three Horizons of ClusterStor



All HDD

HYBRID

ALL FLASH

Primary HPC Storage

COMPUTE

STORE

ANALYZE

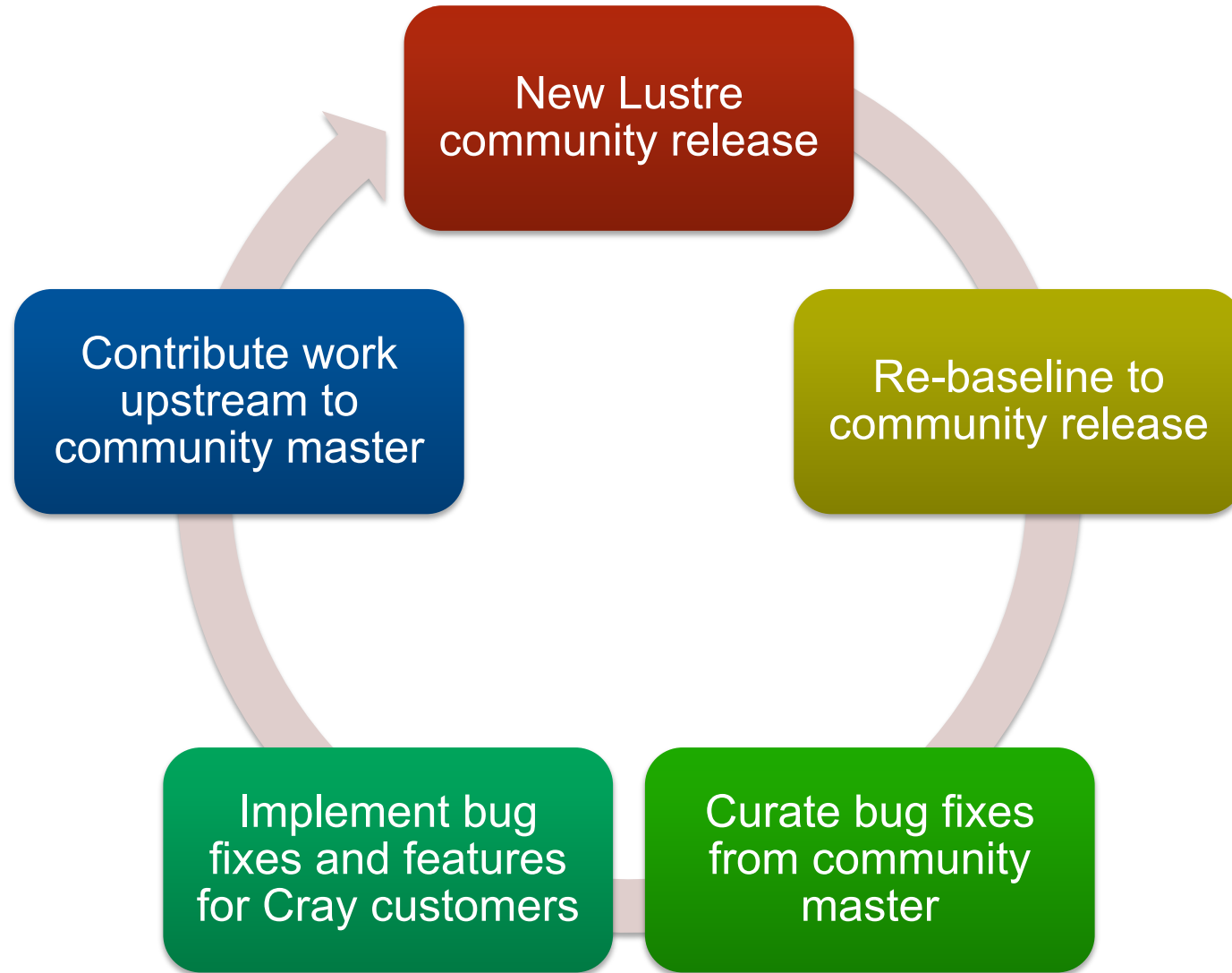
Lustre Improvements for Flash

Reduce client code overhead	Reduce network RPCs	Reduce MDS overhead
<p>Small IO improvements</p> <ul style="list-style-type: none">• Tiny writes [2.11]*• Fast reads [2.11]*	<ul style="list-style-type: none">• Lock Ahead [2.11]*• Immediate short IO [2.11]*• Small overwrites [2.11]*• Data on MDT [2.11]• Size on MDT*	<ul style="list-style-type: none">• DNE.2 [2.8] (reduce MDS bottlenecks)• Progressive file layouts [2.10] (more compact layouts)

Other improvements:

- OST/MDT server tunings for SSD
- Multi-channel PTLRPC (QoS)

ClusterStor Lustre Release Adoption



COMPUTE

STORE

ANALYZE

Lustre Features Coming to ClusterStor



Feature Name	Lustre Release	ClusterStor Release
Lock ahead	2.11	2.x
Multiple MD modify RPCs per client	IEEL 3.0, 2.8	3.0
DNE phase IIb (async commit and recovery)	2.8	3.next
UID/GID mapping	2.9	3.next
Subdirectory mounts	2.9	3.next
Large bulk I/O	2.9	3.next
Multirail LNet	2.10	3.next
Progressive file layouts	2.10	3.next
Project quotas	2.10	3.next
Data on MDT (DoM)	2.11	3.next
File-level redundancy (FLR) - delayed resync	2.11	3.next

The central graphic consists of three overlapping circles. The leftmost circle is filled with a blurred image of a computer terminal displaying multi-colored text (green, yellow, red, white) on a dark background, representing data or code. The middle circle is a solid dark blue and contains the text "Optimization" and a quote. The rightmost circle is filled with a blurred image of a highway at night, showing long-exposure light trails from cars in red and white, representing data flow or analysis.

Optimization

If you can't
measure it, you
can't improve it.

Job Performance Details and Scoring



Jobs

Searchable fields for the most important data to the job step level

for CLUSTERSTOR™ admin

snx11154

98 jobs loaded in 0.595 seconds

2/14/18 12:00 PM to 2/14/18 5:51 PM

Job ID	apid	User ID	Application	Start Time	End Time	Duration	Avg. I/O Size	Metadata Ops
220966.sdb	3406431	1356	mdtest	2018-02-14 11:21:59	2018-02-14 12:39:48	1h 17m 49s	4.1kB	7.0M
220967.sdb	3406431	1356	mdtest	2018-02-14 11:58:53	2018-02-14 12:03:28	4m 35s	N/A	220.1k
220968.sdb	3406431	1356	IOR	2018-02-14 12:00:20	2018-02-14 12:01:57	1m 37s		
221072.sdb	3406433	1356	IOR	2018-02-14 12:02:13	2018-02-14 12:06:00	3m 47s		
221073.sdb	3406435	1356	lustre_ko_sanit	2018-02-14 12:06:16	2018-02-14 12:06:18	2s		
221074.sdb	3406437	1356	mdtest	2018-02-14 12:06:21	2018-02-14 12:10:47	4m 26s		
221075.sdb	3406439	1356	IOR	2018-02-14 12:06:25	2018-02-14 12:14:54	8m 29s		
221076.sdb	3406441	1356	mdtest	2018-02-14 12:10:56	2018-02-14 12:12:04	1m 08s		
221077.sdb	3406443	1356	mdtest	2018-02-14 12:12:20	2018-02-14 12:18:17	5m 57s		
221078.sdb	3406445	1356	IOR	2018-02-14 12:15:10	2018-02-14 12:16:37	1m 27s		

Avg. I/O Size

- 1.0MB
- 650.3kB
- 15.5kB
- 1.0MB

Scoring

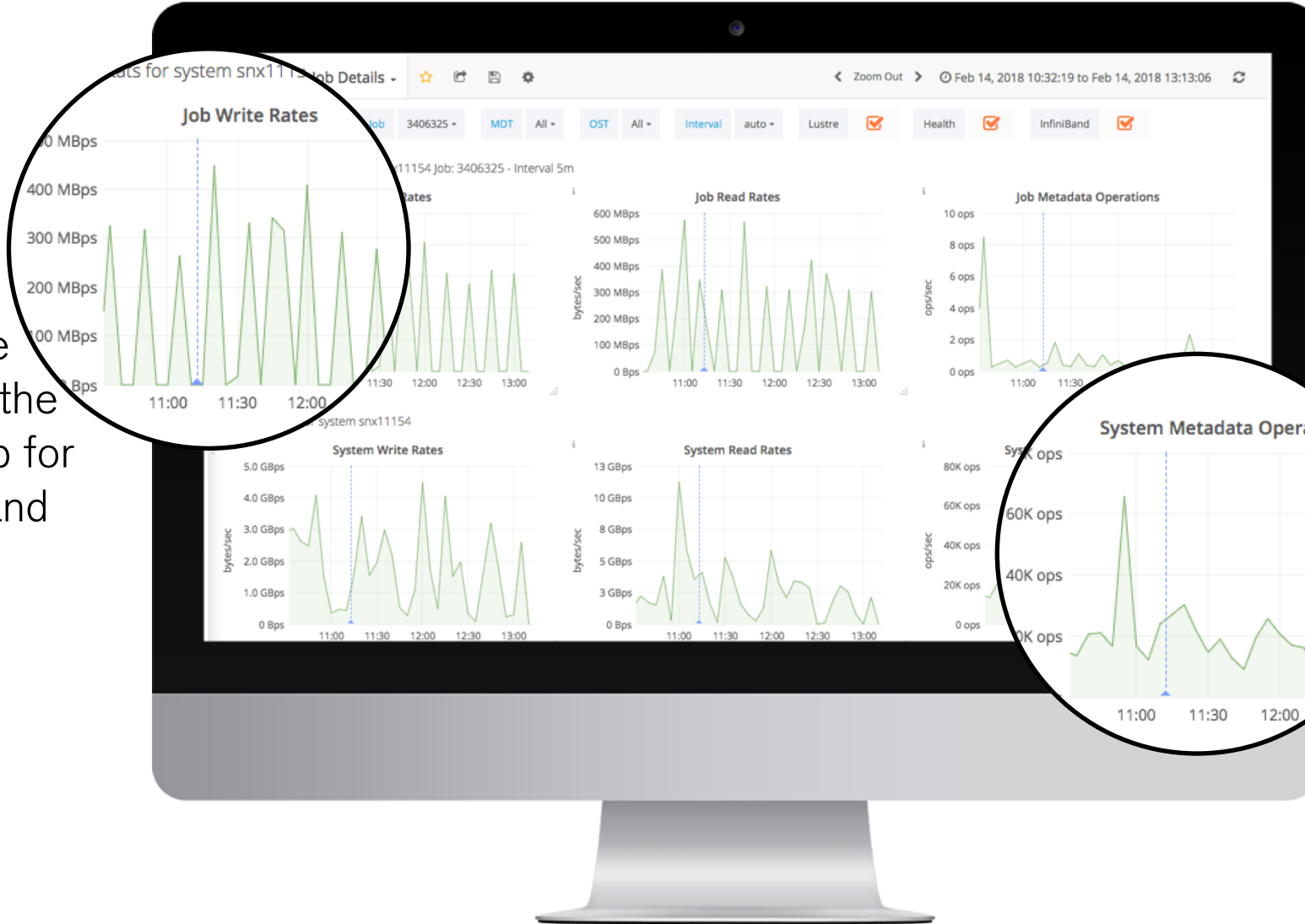
Know which jobs might be causing issues with visual cues

Performance Visualization and Comparison



Visualize

Performance graphs over the life of the job for write, read, and metadata operations



Compare

Compare this job to the rest of the system at a glance

COMPUTE

STORE

ANALYZE

Questions?