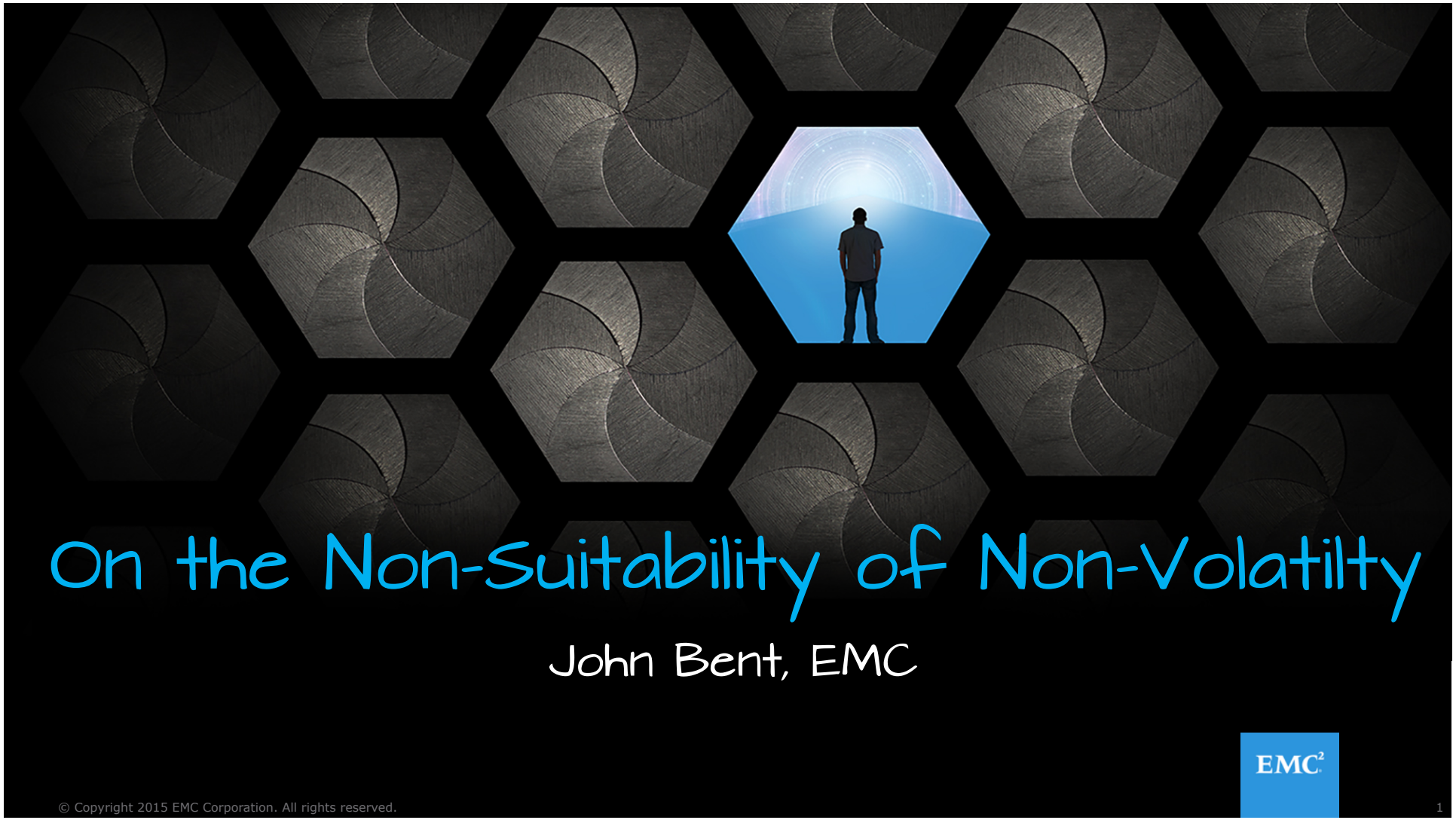


On the Non-Suitability of Non-Volatilty

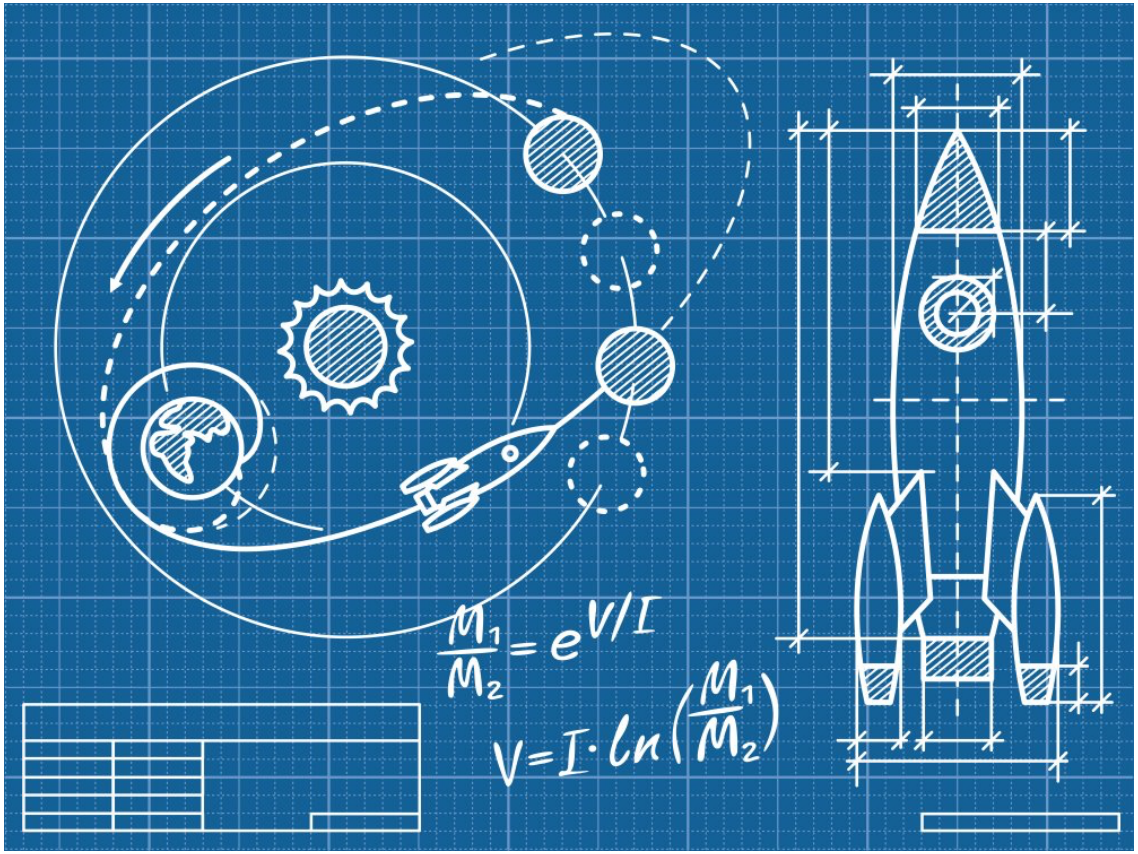
John Bent, EMC

© Copyright 2015 EMC Corporation. All rights reserved.

CO-AUTHORS

- Los Alamos National Lab
  - Brad Settlemyer
  - Nathan DeBardeleben
- EMC
  - Sorin Faibish
  - Uday Gupta
  - Dennis Ting
  - Percy Tzelnic

© Copyright 2015 EMC Corporation. All rights reserved.

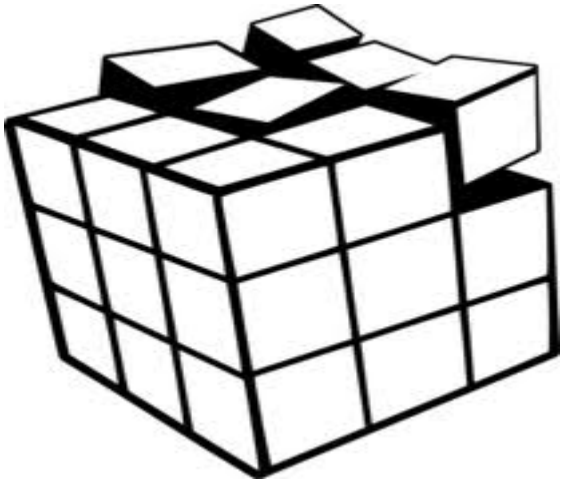




© Copyright 2015 EMC Corporation. All rights reserved.

HIGH-PERFORMANCE COMPUTING

simulation



© Copyright 2015 EMC Corporation. All rights reserved.

System attributes	2010	"2015"		2023-2025	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1day)	

© Copyright 2015 EMC Corporation. All rights reserved.

failure



© Copyright 2015 EMC Corporation. All rights reserved.

checkpointing
(and restart)

© Copyright 2015 EMC Corporation. All rights reserved.



Operating Systems

R. Stockton Gaines Editor

Time, Clocks, and the Ordering of Events in a Distributed System

Leslie Lamport
Massachusetts Computer Associates, Inc.

The concept of one event happening before another in a distributed system is examined, and is shown to define a partial ordering of the events. A distributed algorithm is given for synchronizing a system of logical clocks which can be used to totally order the events. The use of the total ordering is illustrated with a method for solving synchronization problems. The algorithm is then specialized for synchronizing physical clocks, and a bound is derived on how far out of synchrony the clocks can become.

Key Words and Phrases: distributed systems, computer networks, clock synchronization, multiprocess systems

CR Categories: 4.32, 5.29

Fig. 1.

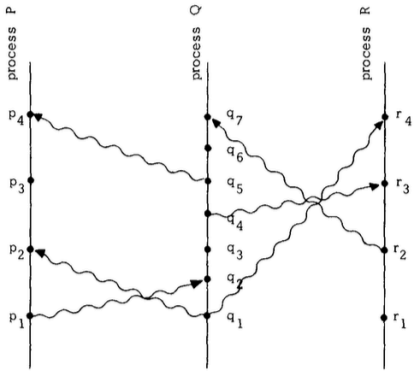
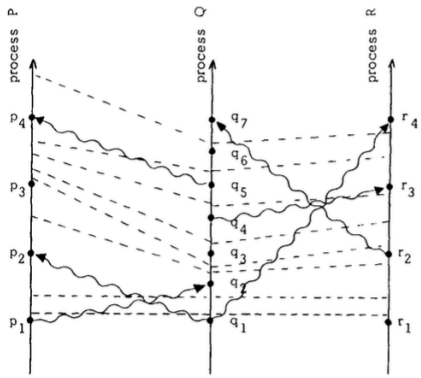


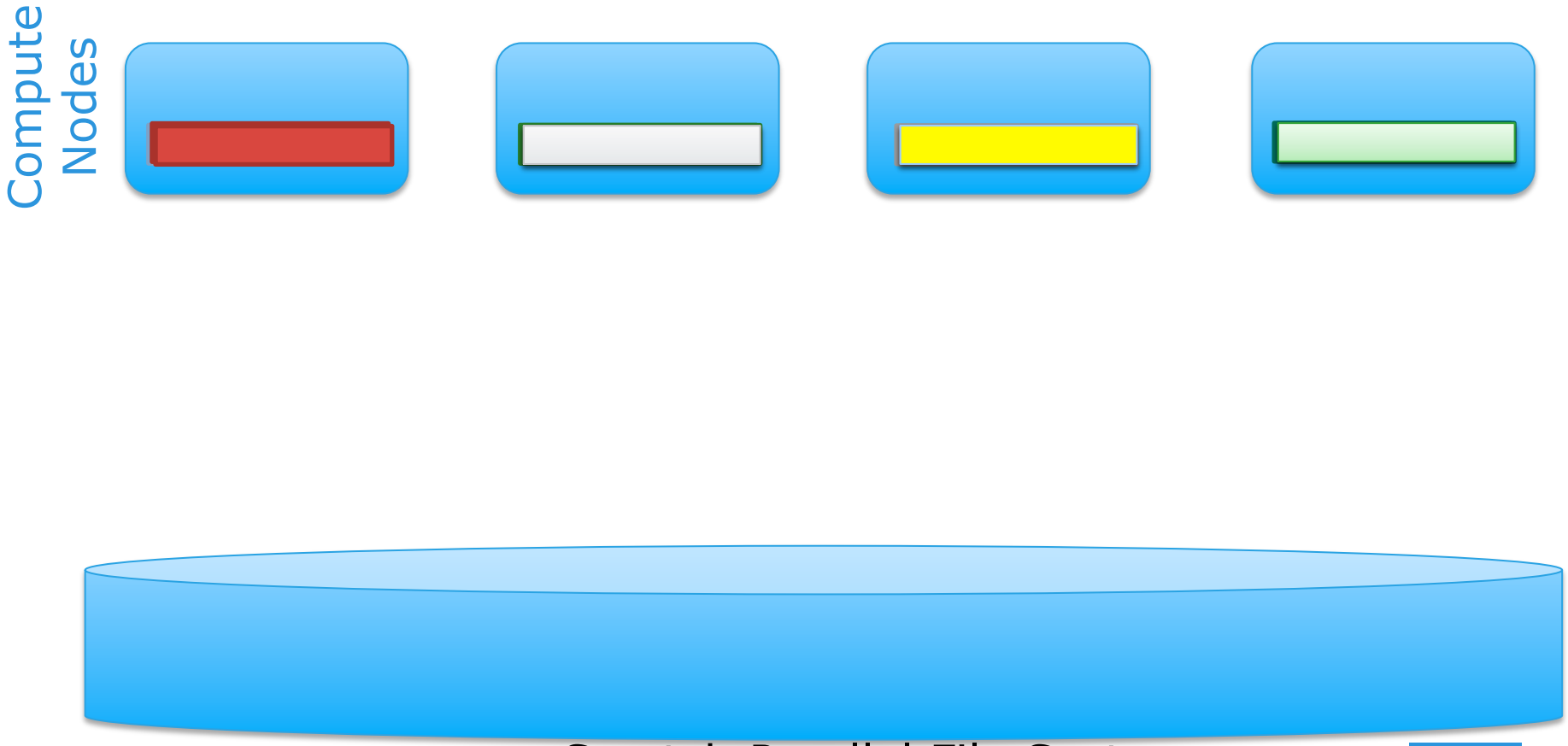
Fig. 2.





EMC²

HPC CHECKPOINTING



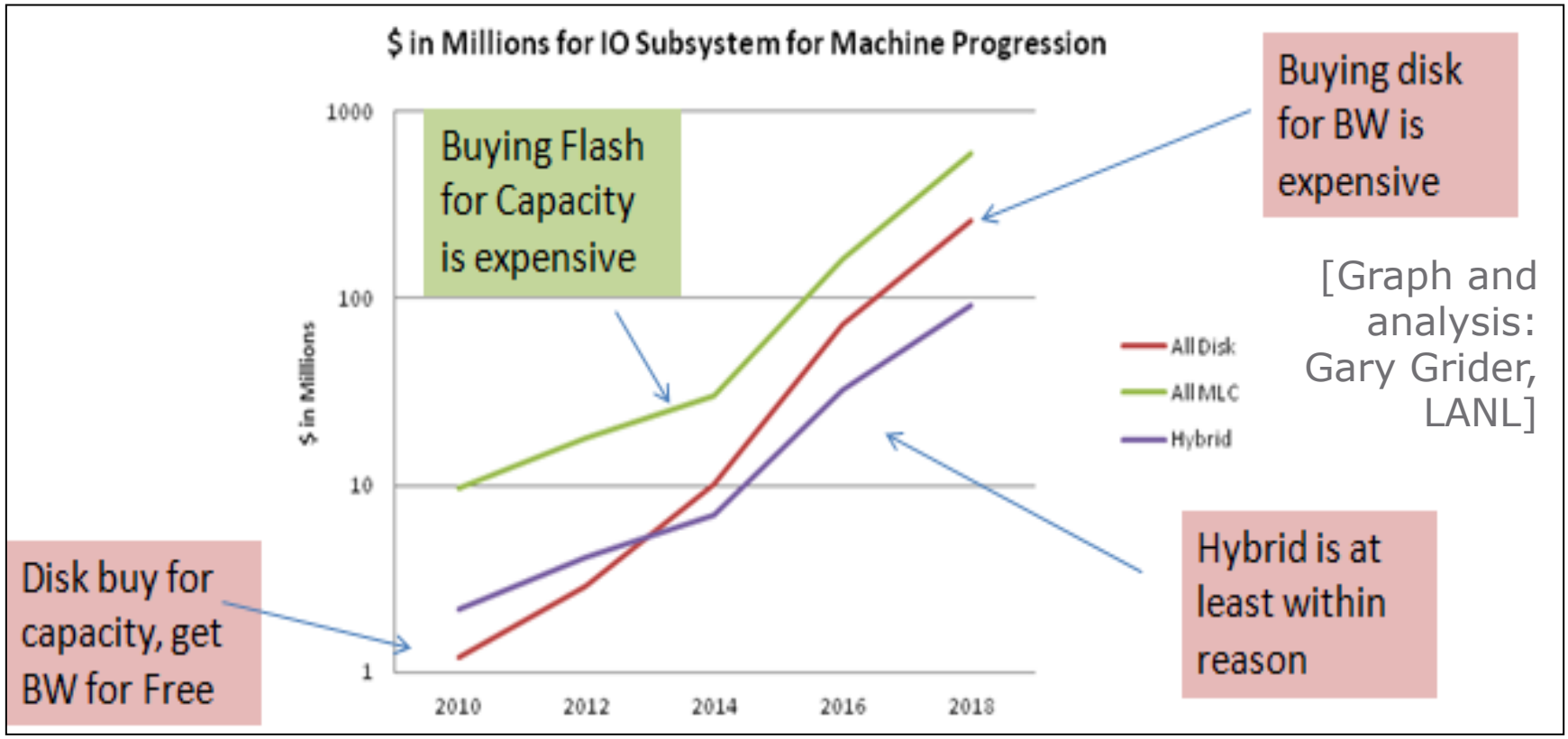
© Copyright 2015 EMC Corporation. All ric

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		~~20 MW~~	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	~~1 TF~~	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

Congressionally dictated efficiency → 100 TB/s

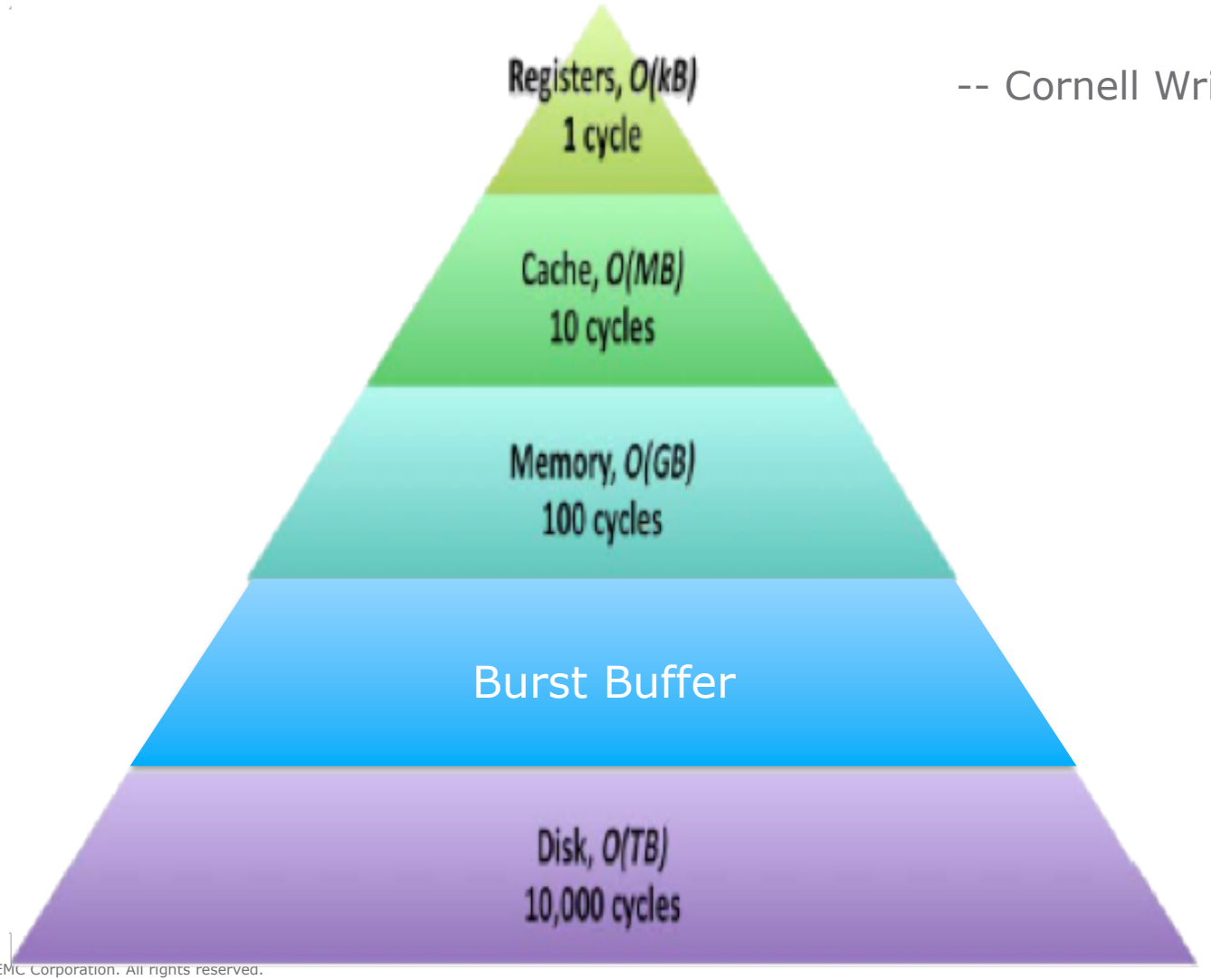
© Copyright 2015 EMC Corporation. All rights reserved.

HPC HAS A HARDWARE PROBLEM



Projected Storage Requirement in 2018: 1 EB Capacity and 100 TB/sec

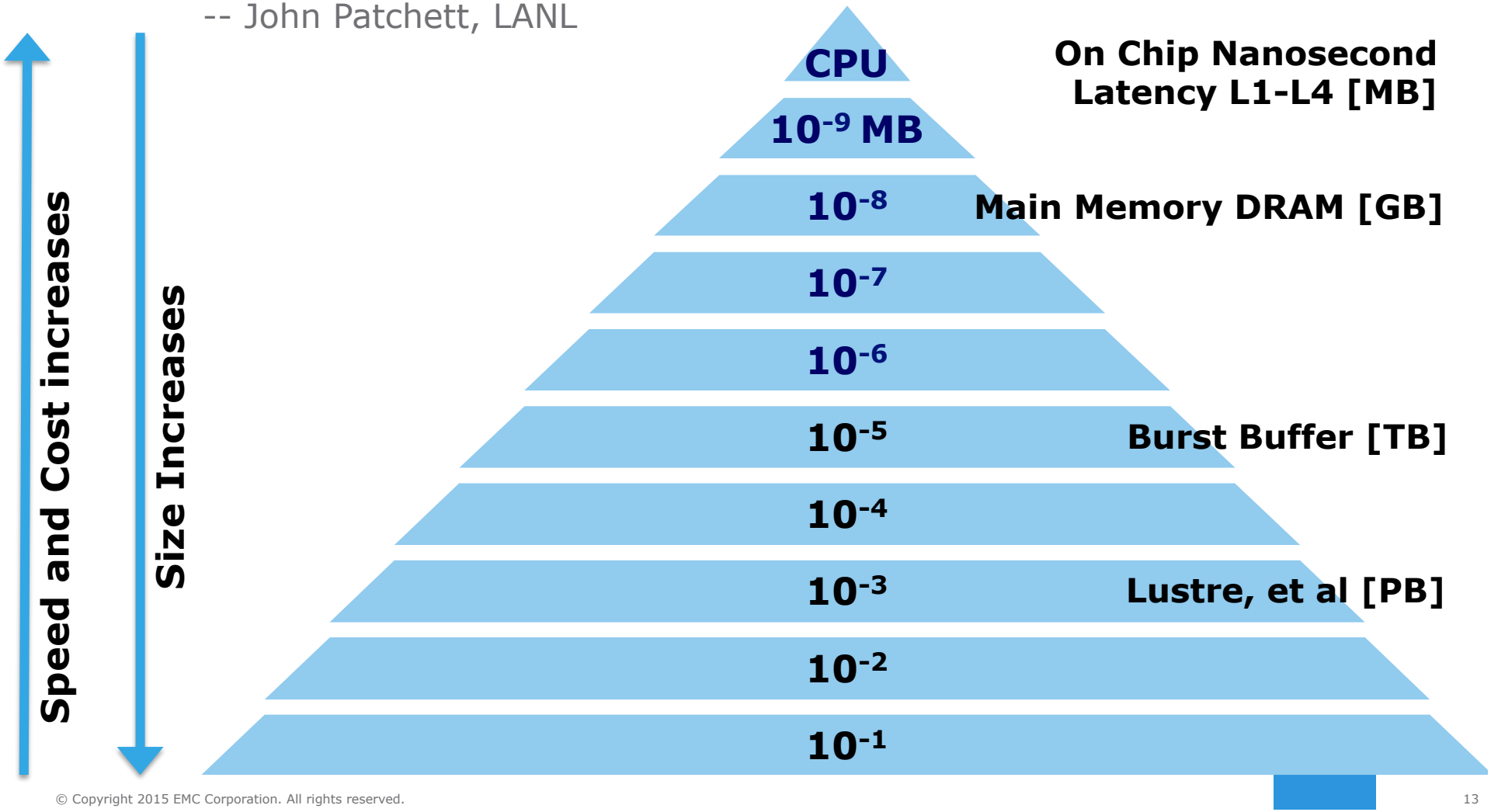
© Copyright 2015 EMC Corporation. All rights reserved.



-- Cornell Wright, LANL

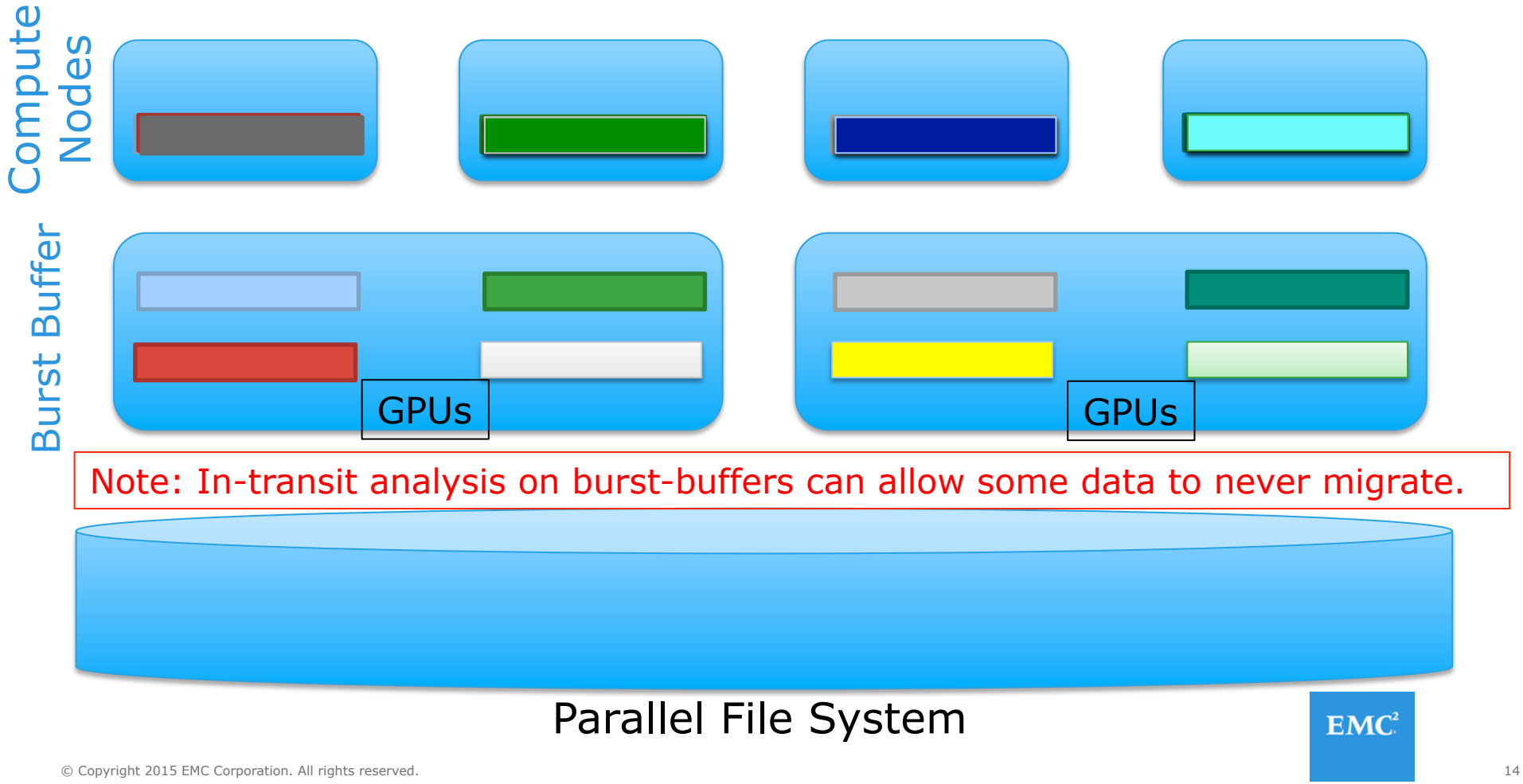
© Copyright 2015 EMC Corporation. All rights reserved.

-- John Patchett, LANL



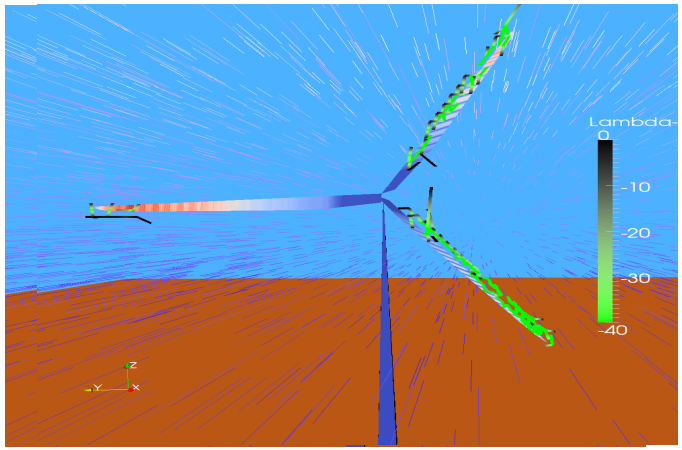
© Copyright 2015 EMC Corporation. All rights reserved.

CHECKPOINT WITH BURST BUFFERS

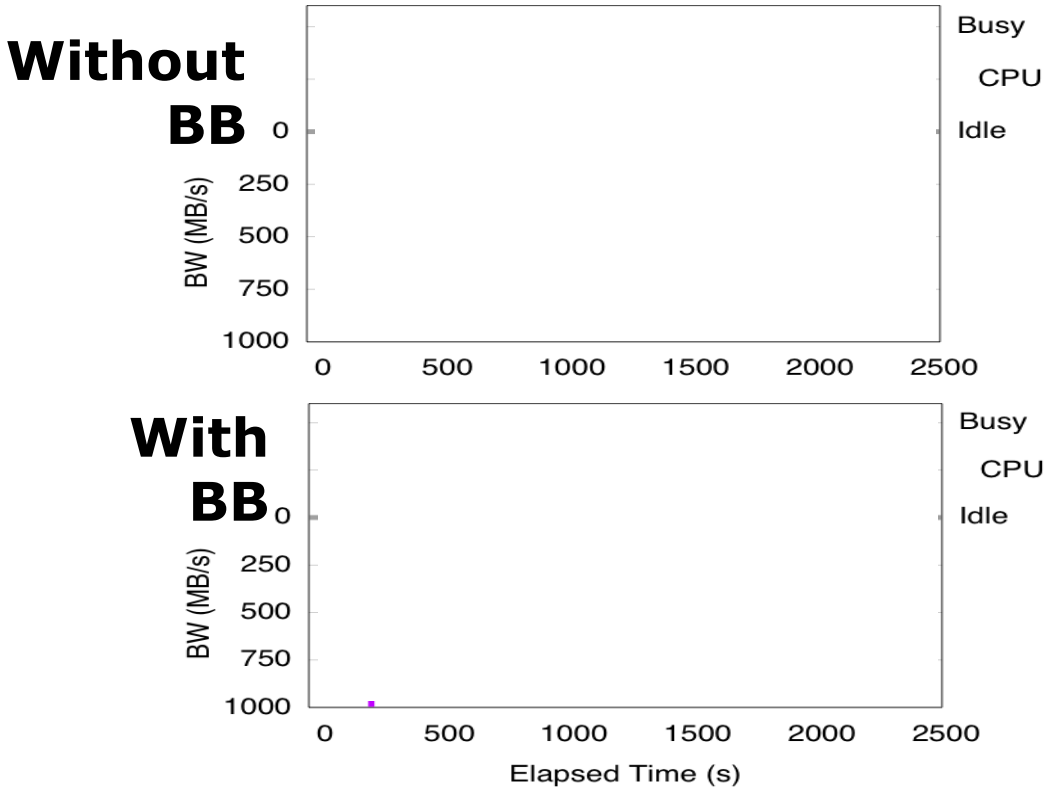


© Copyright 2015 EMC Corporation. All rights reserved.

PROTOTYPE BURST BUFFER [LANL/EMC]



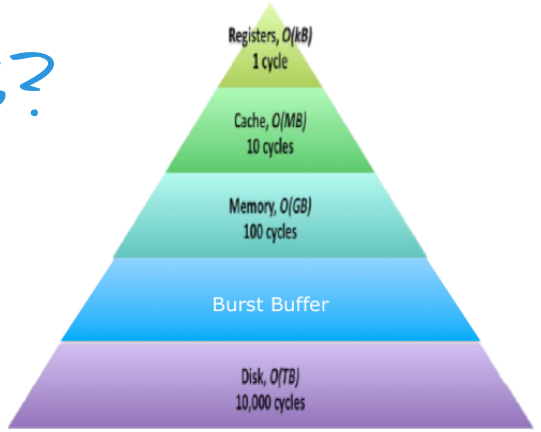
LANL wind turbine workload showing 8 phases of compute, checkpoint, and vizualization running on EMC prototype aBBa.



[http://johnbent.com/Pubs/bent_msst12.pdf]

© Copyright 2015 EMC Corporation. All rights reserved.

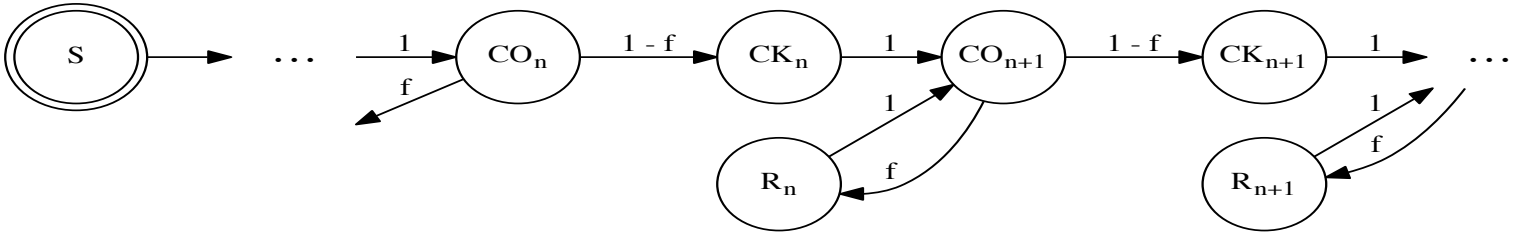
WHAT ARE BURST BUFFERS?



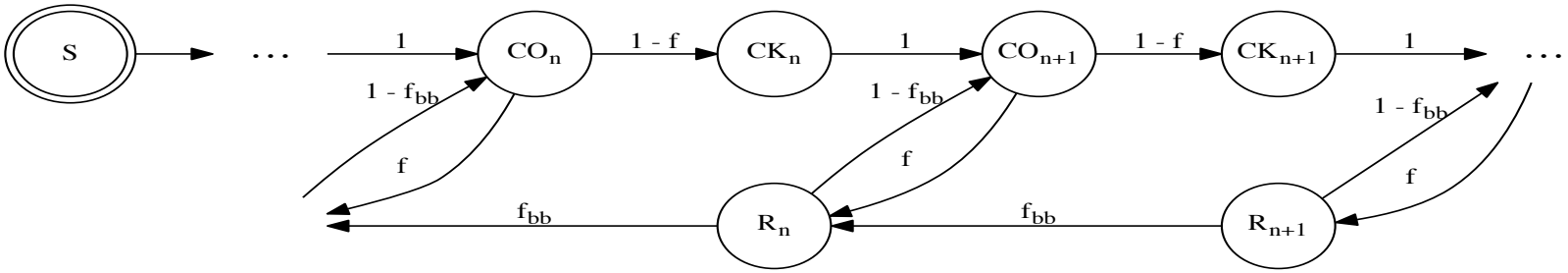
- An extension of memory?
  - Data can be lost due to failure
- An extension of storage?
  - Data must be protected and recoverable following failure
- Built with NVRAM so they look like storage
- However, BAD-FS and Tachvon remind us that
  - "Lost data can *also* be recovered via recomputation"

© Copyright 2015 EMC Corporation. All rights reserved.

Burst Buffer Workflows



Workflow w/ reliable burst buffer

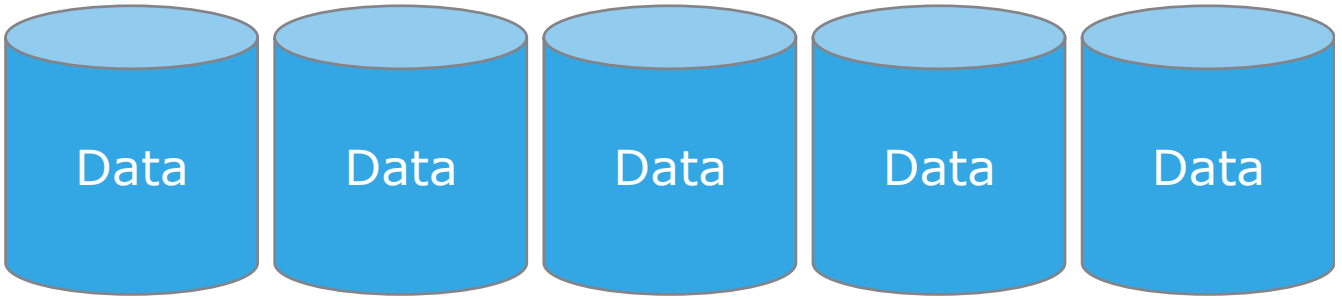


Workflow w/ unreliable burst buffer

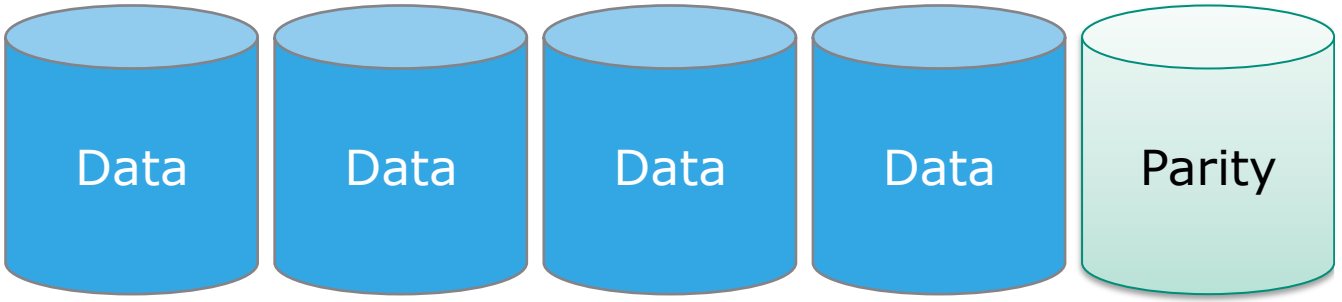
© Copyright 2015 EMC Corporation. All rights reserved.

WHY DOES RELIABILITY ADD OVERHEAD?

Nonreliable BB



Reliable BB



© Copyright 2015 EMC Corporation. All rights reserved.

OUR MODEL

Compute Nodes	100K
Compute Cores	1 billion
Checkpoint Data	1EB / day
Compute Node : Burst Buffer Node	100:1
Burst Time	5 minutes
Drain Time	60 minutes
Compute Phase	60 minutes
MTBI	24 hours

Important variables

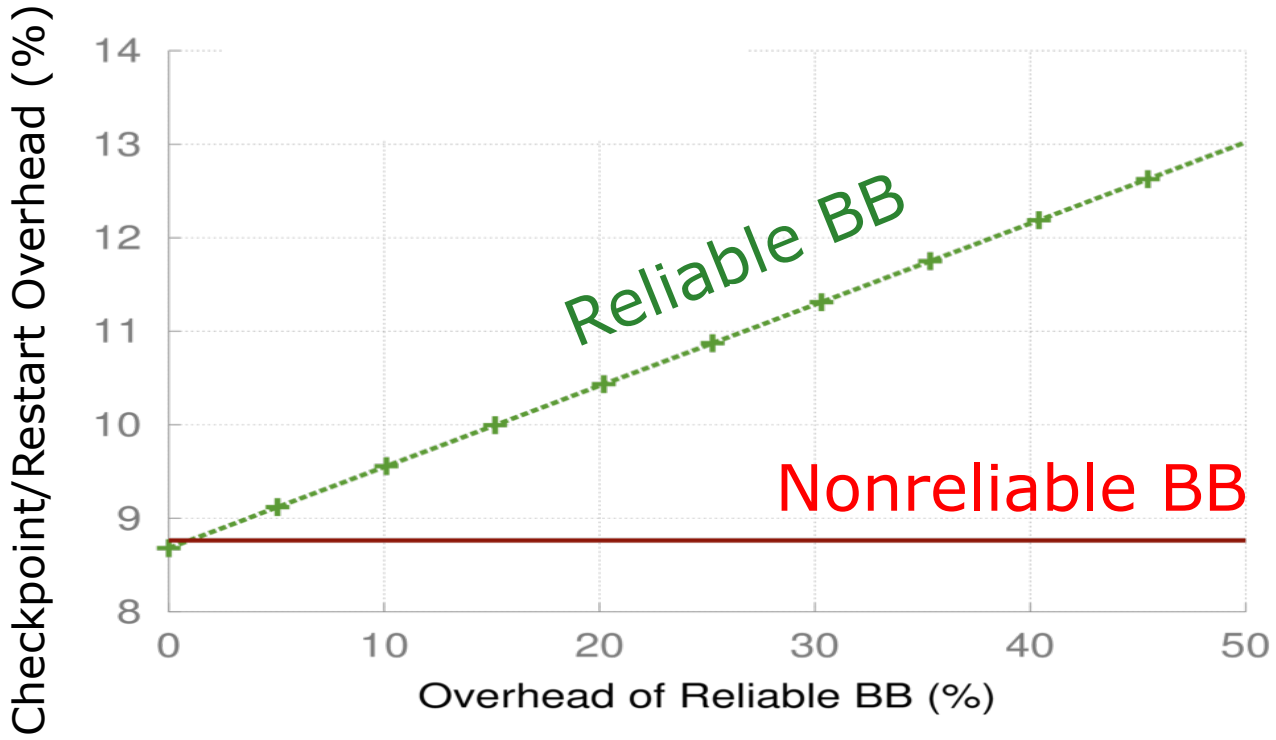
- CN : BBN ratio
- Burst time
- Drain time
- Compute phase
- MTBI

Important assumptions

- Decoupled failure domains
- Deterministic computation
- Coherent failure detection

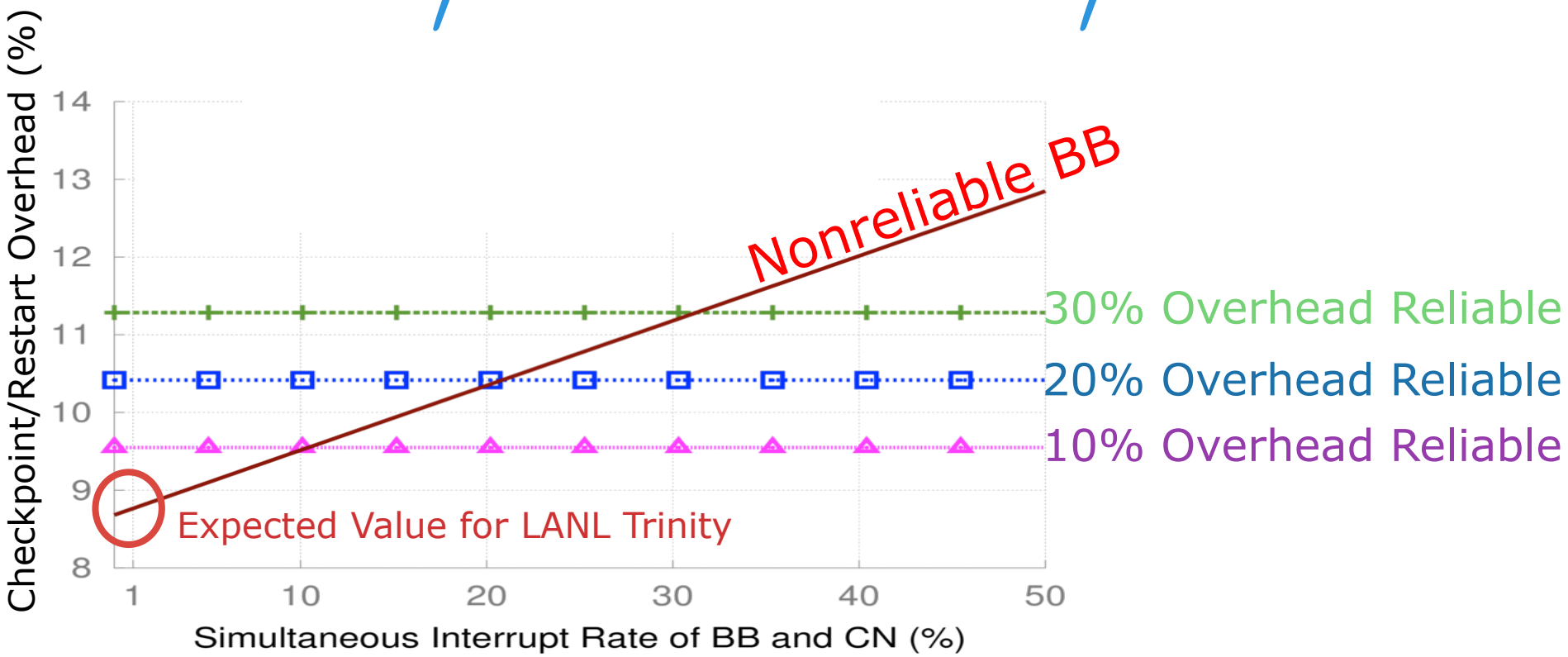
© Copyright 2015 EMC Corporation. All rights reserved.

Reliability adds Overhead



© Copyright 2015 EMC Corporation. All rights reserved.

Reliability is Not Normally Needed



© Copyright 2015 EMC Corporation. All rights reserved.

CONCLUSIONS

- General
  - Don’t optimize the infrequent at the expense of the frequent
- Specific
  - Don’t be misled by the ‘N’ in ‘NVRAM’

- Thanks!
  - On to MDHIM!

john.bent@emc.com

© Copyright 2015 EMC Corporation. All rights reserved.


EMC²
john.bent@emc.com