OPEN Compute Engineering Workshop

March 9, 2015
San Jose

OPEN
Compute Engineering Workshop

Architecting Highly Efficient Web-scale Cold Storage for Unstructured Data

Every bit and watt counts

Mark Goros
Caringo
CEO and Co-founder

Jie Yu, PhD
WD, a Western Digital Company
Sr. Director Alliances and Analytics, WD Labs

Caringo Overview

Highly efficient object storage software

- Founded in 2005, complete focus on software
- Pure object storage; store metadata with data & eliminate complexity
- Field proven. Shipping v7, clusters running 7x24 without incidents for 5+ years
- Consistent innovation in efficiency, simplicity, performance, data protection

Caringo
Swarm™
Storage Software

Recognized leader

STORAGE MAGAZINE PRODUCTS OF THE YEAR 2014 FINALIST

CRN TOP 100 Coolest Cloud Vendors 2014

IDC Analyze the Future

Gartner

Customers with Different Use Cases

....that all have similar goals

- Radically simplify infrastructure
- Cost effective scalability for unstructured data (i.e. web-scale)
- Data must be searchable, accessible and protected

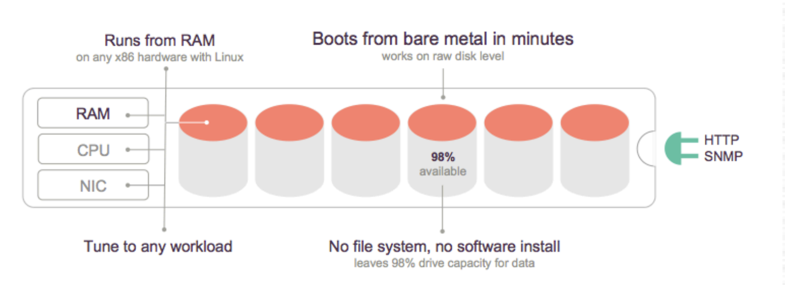
U.S.ARMY
AMERICA'S NAVY
NETFLIX
verizon
Telefonica
BT

DEPARTMENT OF JUSTICE FEDERAL BUREAU OF INVESTIGATION
Ask
JOHNS HOPKINS UNIVERSITY
CITY OF AUSTIN FOUNDED 1839
match.com
St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

… plus hundreds more

Technology that enables Flexibility at Scale

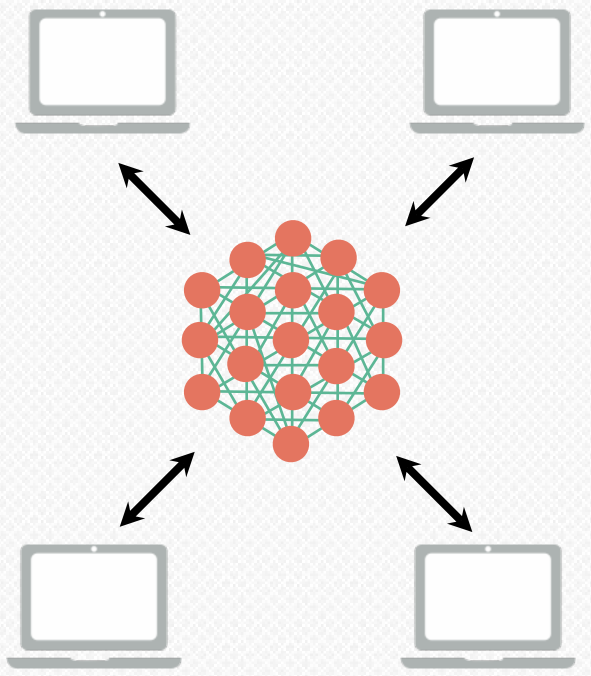
1 All-inclusive software


Runs from RAM
on any x86 hardware with Linux
Boots from bare metal in minutes
works on raw disk level
RAM
CPU
NIC
98%
available
HTTP
SNMP
Tune to any workload
No file system, no software install
leaves 98% drive capacity for data

Node 1
Node 2
Node 3
Node …

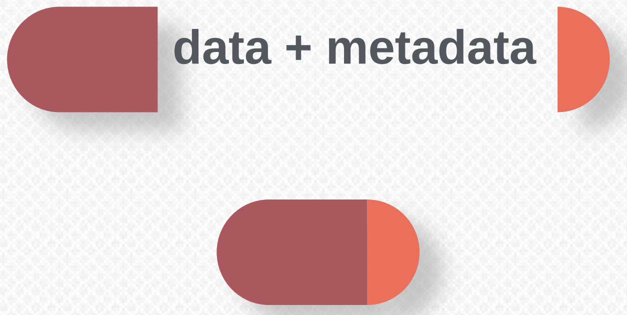
- Boots from bare metal
- **Simplified approach** – no file system, no RAID, no single points of failure

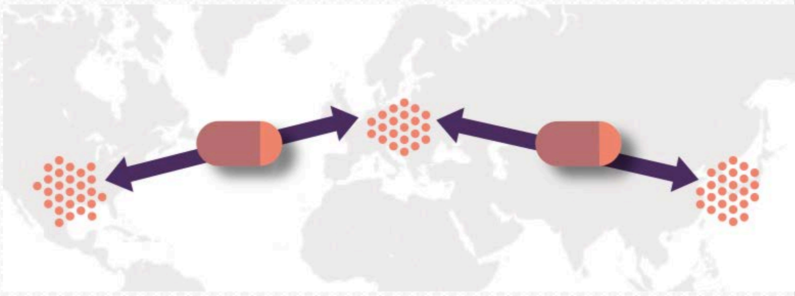
2 Swarm architecture



- All nodes cooperate to perform all functions
- Gets faster as cluster grows
- All nodes run the same code

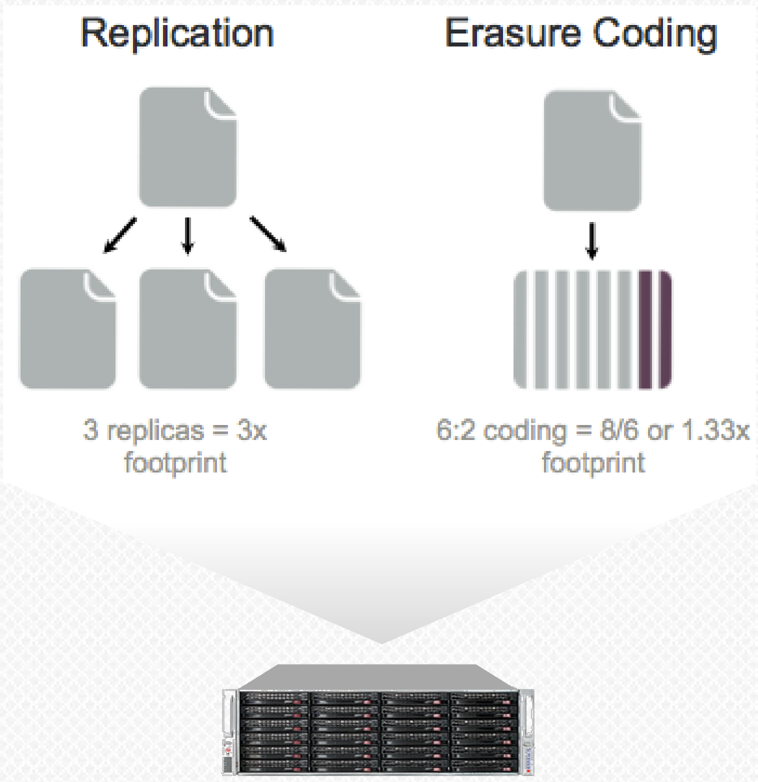
3 Encapsulated data





- Object contains all system, custom metadata, Lifepoints
- **Unlocks data** from application and location
- No metadata database

4 Elastic Content Protection

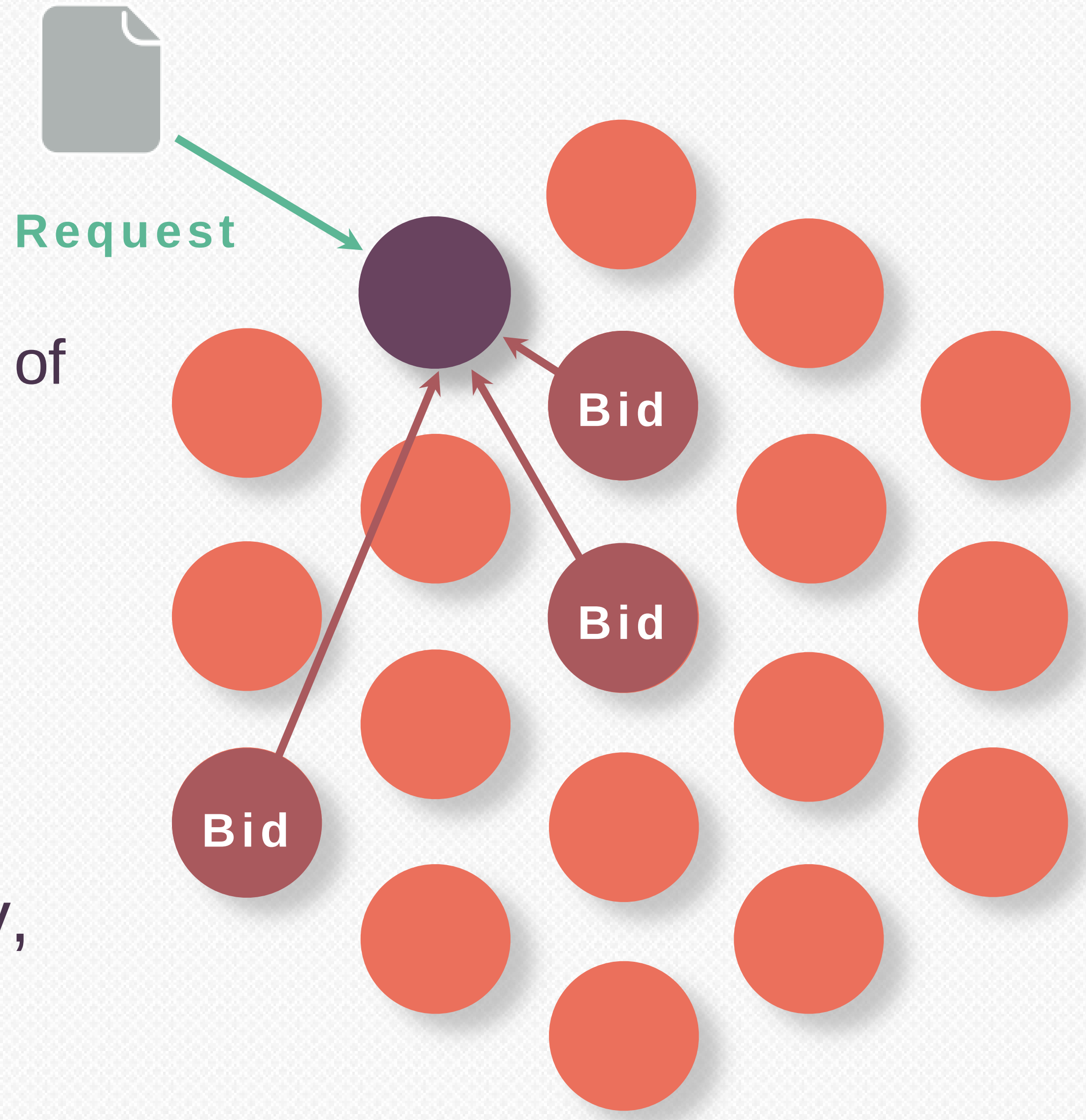


- **Patented:** store erasure coded and replicated objects on the same node
- Shift between replication and erasure coding to reflect the value of the data

The Swarm Data Market

Use any node to access cluster

- Swarm considers market conditions for execution
  - “Bids” are willingness to write, update or retrieve objects
  - Bid affected by: volume state, cached objects, load, storage used, state of volume, how busy, i/o queues, etc.
- HTTP redirect to lowest cost node for WRITE/READ
  - With caravanning protection



Why is this important?

- Enables <u>automatic load balancing</u>, Super-fast volume recovery, adaptive power conservation (Darkive) and more
- Content caching and sub-clusters taken into consideration
- Mechanism is expandable, e.g. tiering within a cluster

Darkive™ - The Key to Cold Storage

Patented adaptive power conservation enabled by bidding

- Spin down drives and reduce CPU utilization
- Saves power and cooling
- Two types:
  - Adapts automatically to system behavior based on configurable period of inactivity
  - Admin can designate Darkive nodes or even Darkive sub-clusters
- Designated archive nodes bid more aggressively until full
- Bidding directs writes towards the few fuller nodes while others remain spun down

Cold Storage with Darkive Results

In a recent 30 PB Cold Archive

- Power savings reduced monthly TCO by a full 70%
  - Bidding seeks to fill volumes and let nodes go quiescent
  - Saved in overall power costs due to spun down nodes
- Since object indexes are in RAM, Swarm can answer queries without spinning up drives and can choose active drives to serve data
- Admin can also select a wake/sleep cycle for nodes health checking

Why is this important?

Estimated data growth by 2020

7.3 to 40 Zettabytes

90% unstructured data

The Path to Zettabytes

Driving the need for Cold Storage

- Companies of all sizes amassing large data sets
- Data is the basic unit for build value; it must remain accessible
  - The value of data changes over time
  - The access patterns also change as the data ages
- Value diminishes but doesn’t vanish
- A perfect solution is Responsive Cold Storage

Pioneers in Cold Storage

Amazon Glacier

- 11 nines durability
- Runs in high capacity low cost discs

Facebook “Sub-zero”

- 2 dedicated datacenters
- Storage servers power off when not in use
- Two backups of data

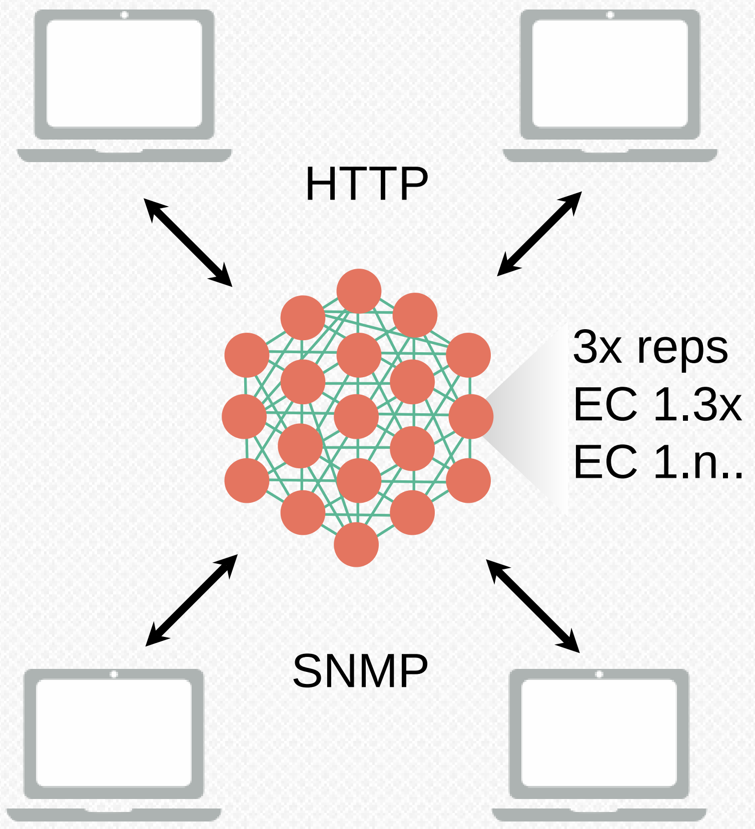
Your name here

Cold Storage Requirements

- A minimum of 11 nines 99.999999999% (like Glacier)
- Equivalent of 2 backups (like Facebook)
- Power off server when not in use (like Facebook)
- Optimized for
  - Efficient use of data center space (fewer racks)
  - $ per GB
  - Watts per GB

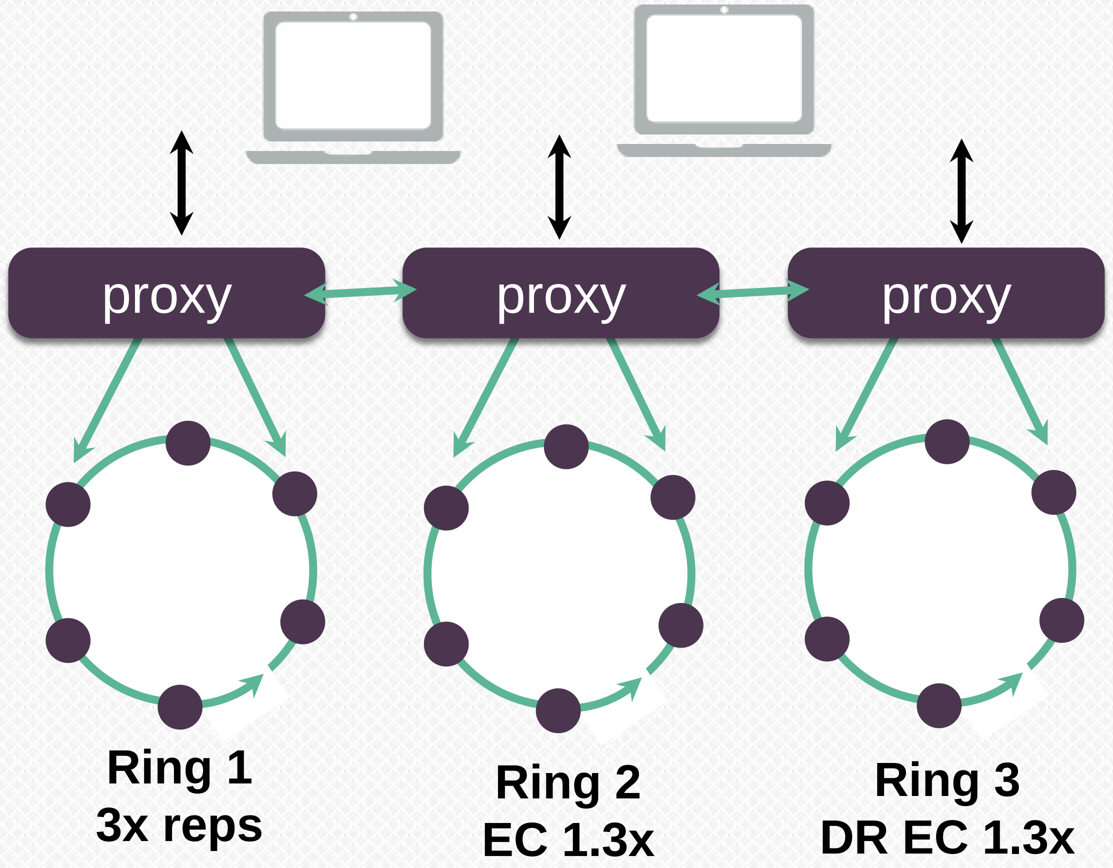
Software Need – Efficiency at Scale

Caringo Swarm



With thousands of servers which do you want?

Ring Architecture



- Single pool of resources
- Plug into network, boot and up and running in 90 secs
- Storage automatically deploys and balances
- Swarm manages metadata, OS, access and storage

- Complex setup, different drives in node for OS, metadata, and storage
- RAID controllers, RAID 1 for OS and metadata drives
- Separate RINGS for each protection scheme
  - **Rigid scale**, must add an equal amount of servers across all RINGS
  - Migration managed by supervisor/proxy
  - **Single point of failure**, EC reconstituted on connector servers
- Heavy reliance on script automation to manage complex deployment

3 PB Cold Archive Example: Swift vs. Swarm

	Cold Archive Requirements
Reliability	11 9’s or greater
Storage need	3,086 PB useable
Chassis need	72x6TB HDD per
Racks required	10 chassis per rack
Drives required	

Same Example: Standard Enterprise HDD vs WD Ae HDD

openstack Swift

		Standard enterprise HDD	WD Ae HDD
100% operation	Drives (watts)	10,863	6,830
	Chassis (watts)	2,640	2,640
	Total	**13,503**	**9,470**

Progressive Capacity

"Delivering Value Across the Capacity Continuum" through

- More capacity-per-cabinet/floor tile,
- Thus, reduced compute infrastructure,
- Thus, better infrastructure utilization

Application value

- Render more capacity-per-drive and capacity-per-volumetric space
- Improved TCO

Guiding Principles

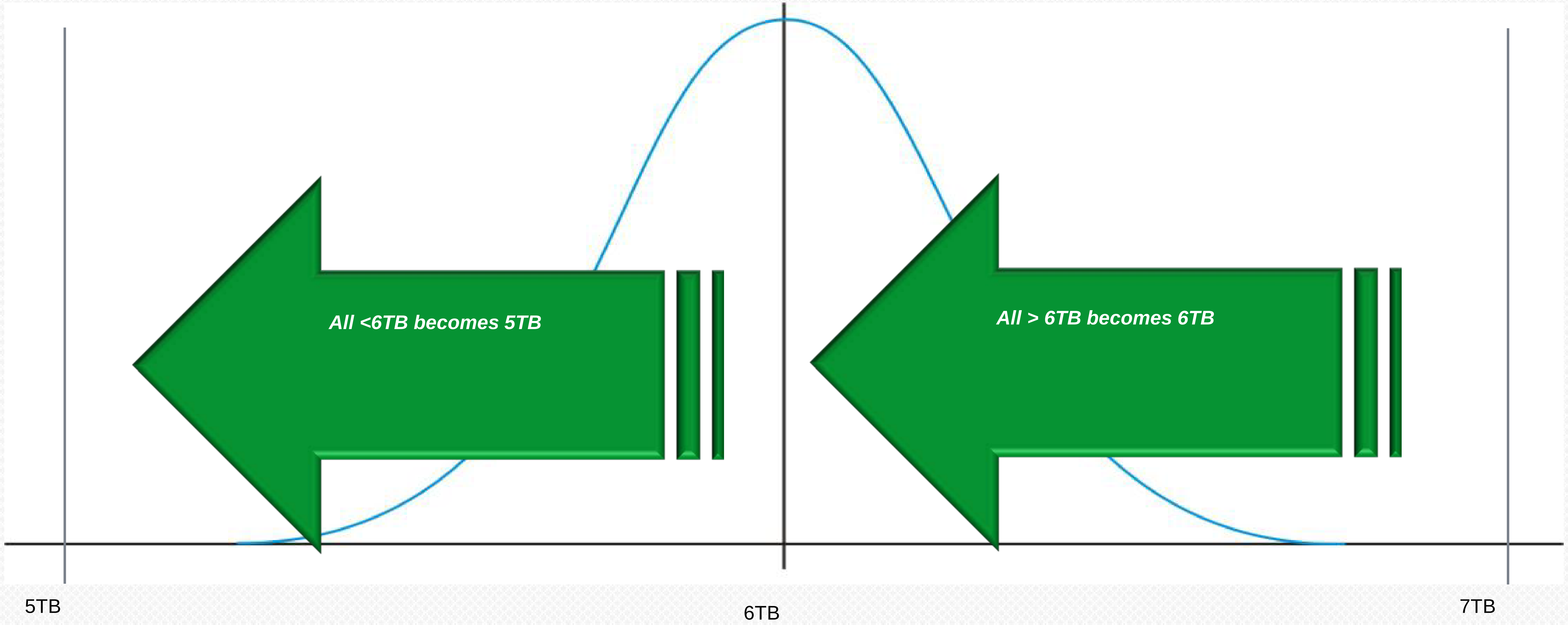
- Common platform product strategy to minimize cost of qualification + components:
    - **Single Platform** for ~18 months with incremental feature improvements (minimizes qualification costs)
    - **Scale Economics** / high volume components and architectures (establish cost leverage thru volume across multiple market segments)
- Efficient and extendible investments for both supplier and customer
    - Any work done to enable features will span multiple generations

Application considerations

- Handling a range of capacities (In 100GB increments: e.g. from 5.9 TB to 6.6 TB)

Conventional Approach

Suboptimal Capacity Utilization



- **Traditional Fixed Capacity Points**
  - Requires majority of manufacturing distribution above target
  - Large capacity penalty for population below target

Progressive Capacity Renders Increased Capacity Gains & Cost Savings through Natural Technology & Factory Maturation

PQ1

- Avg Capacity: 6.0 TB
- 200K units = 1.200 EBs

PQ2

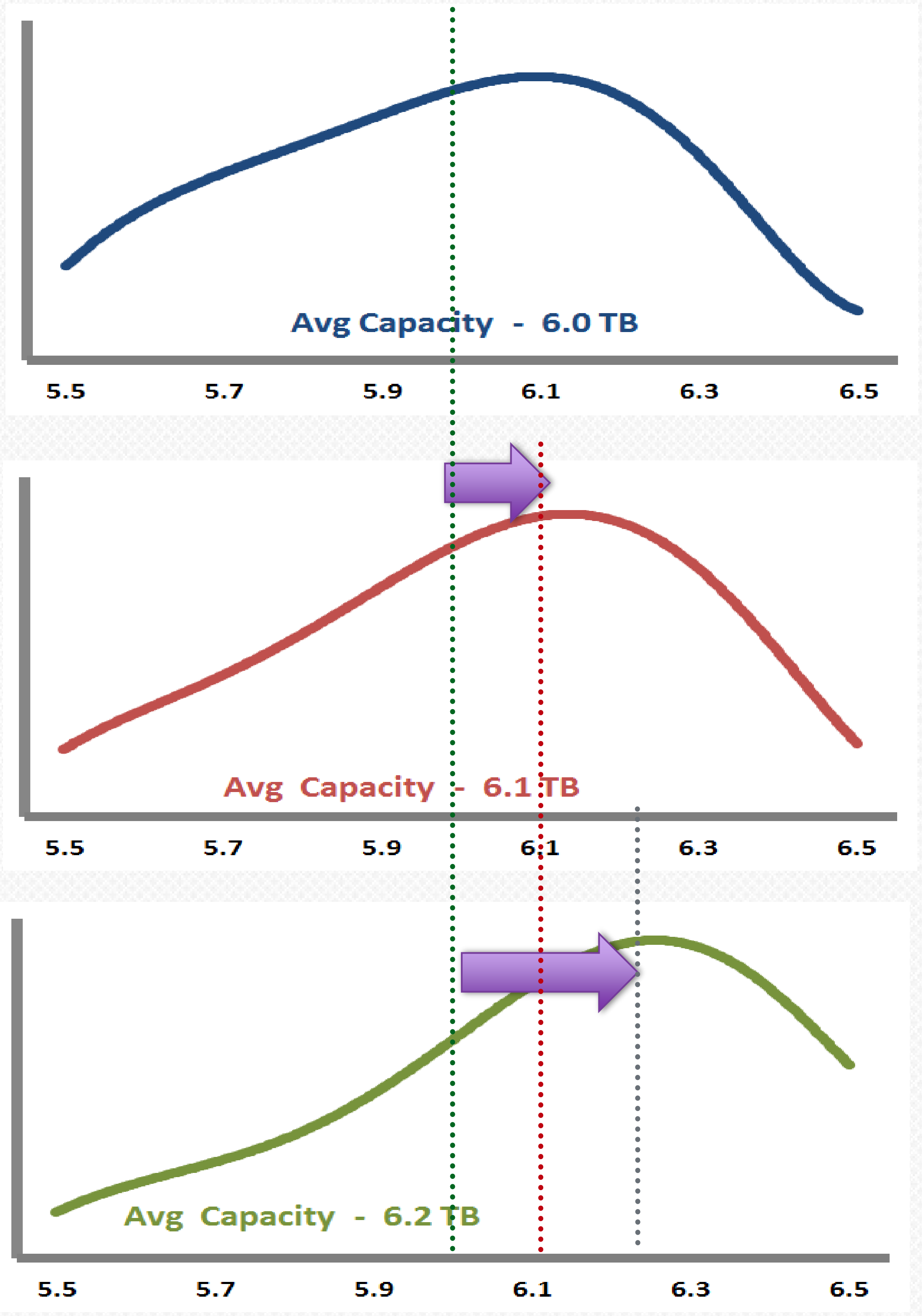
- Avg Capacity: 6.1 TB
- 200K units = 1.220 EBs

Adding 20,000 TBs

PQ3

- Avg Capacity: 6.2 TB
- 200K units = 1.240 EBs

Adding 40,000 more TBs



Summary – Caringo Swarm + WD Ae HDDs

Responsive and responsible Cold Storage

- Responsive in real time to application needs
- Responsible in the way it burns power, cooling and data center footprint
- Darkive the key to Cold Storage power savings
- Use of WD Ae drives combines the best of both worlds

Saving Bits

50%+
less hardware DC footprint

Saving Watts

54% - 95%
Less power

OPEN

Compute Engineering Workshop

March 9, 2015

San Jose