SNIA

Education

Energy Conservation Strategies for Storage in HPC Environments

Alan G. Yoder, Ph.D. , NetApp

Abstract

- Energy Conservation Strategies for Storage in HPC Environments
  - With some large-scale HPC data center owners admitting that storage and computing resources can cost more to power over their lifetimes than to purchase, attention to energy management in HPC is timely. This talk will focus on

    storage; it will survey various storage schemes for HPC, ranging from Hadoop clusters to enterprise storage arrays, and compare energy usage and management in the various schemata. The potential contribution of point technologies such as data deduplication will also be covered.

Outline

- The problem(s)
- Facilities technologies for energy savings
- Storage technologies for energy saving
- Wrap up

Problem: making heat just to cool it

- Servers, storage and switches are HEATERS
  - 100% efficient energy-to-heat conversion
  - Rotating media uses 85% of max power *at idle*!
- A/C is a big "undo" mechanism for overheating
  - But less than 100% efficient (typically 70%)

> 60% of the power in a traditional data center does no IT work

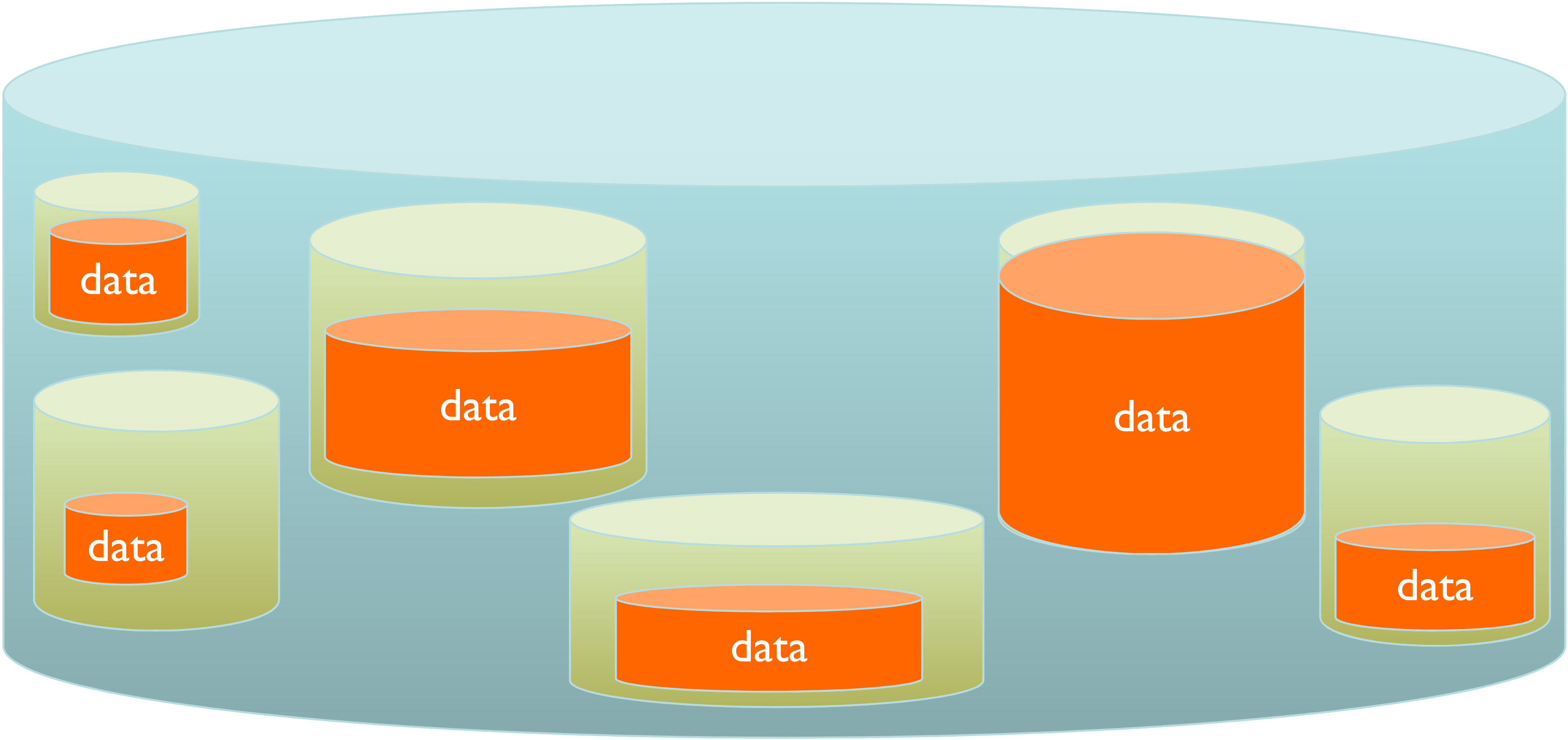
(PUE* ~ 2.5)

HEAT
BUILDING
UNDO

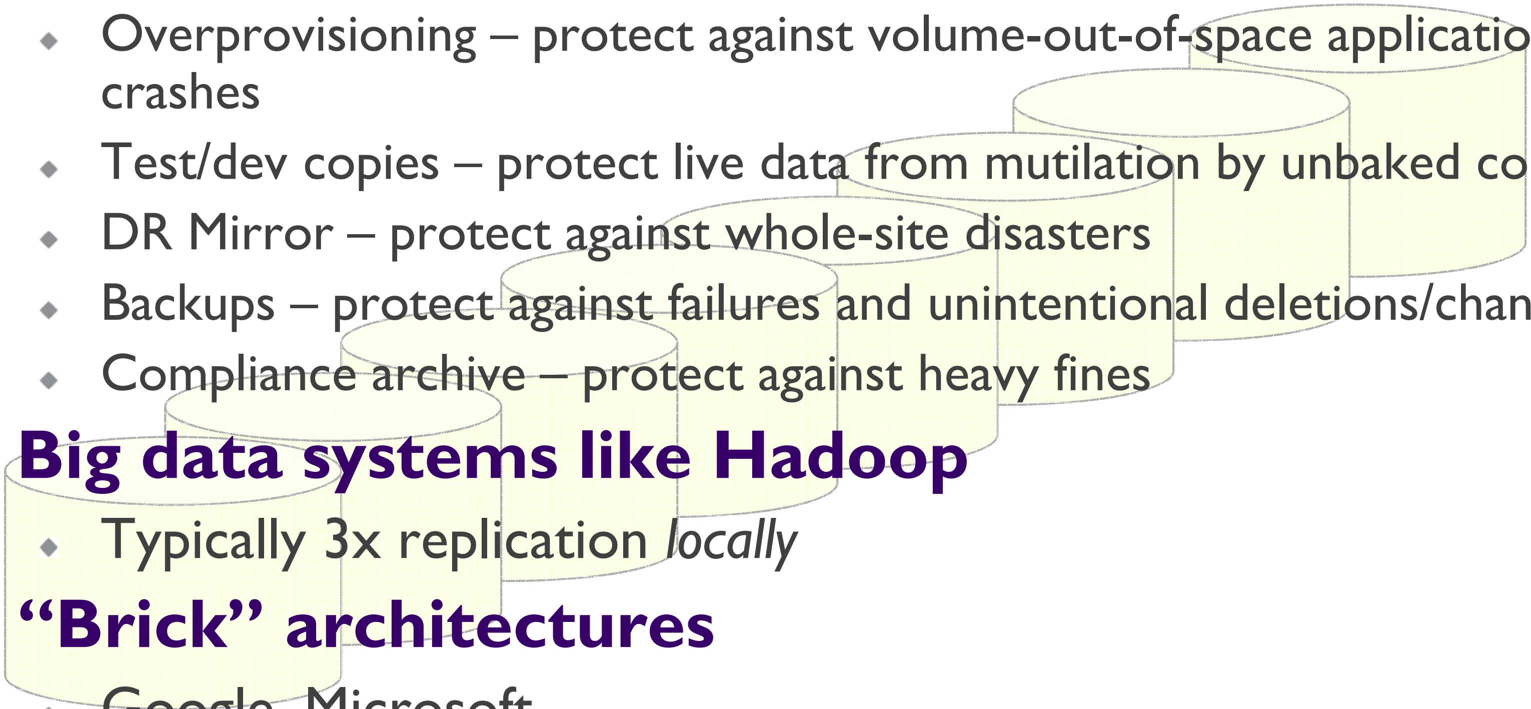
* PUE defined later

Problem: unused space

- Overprovisioning of systems
- Overprovisioning of containers

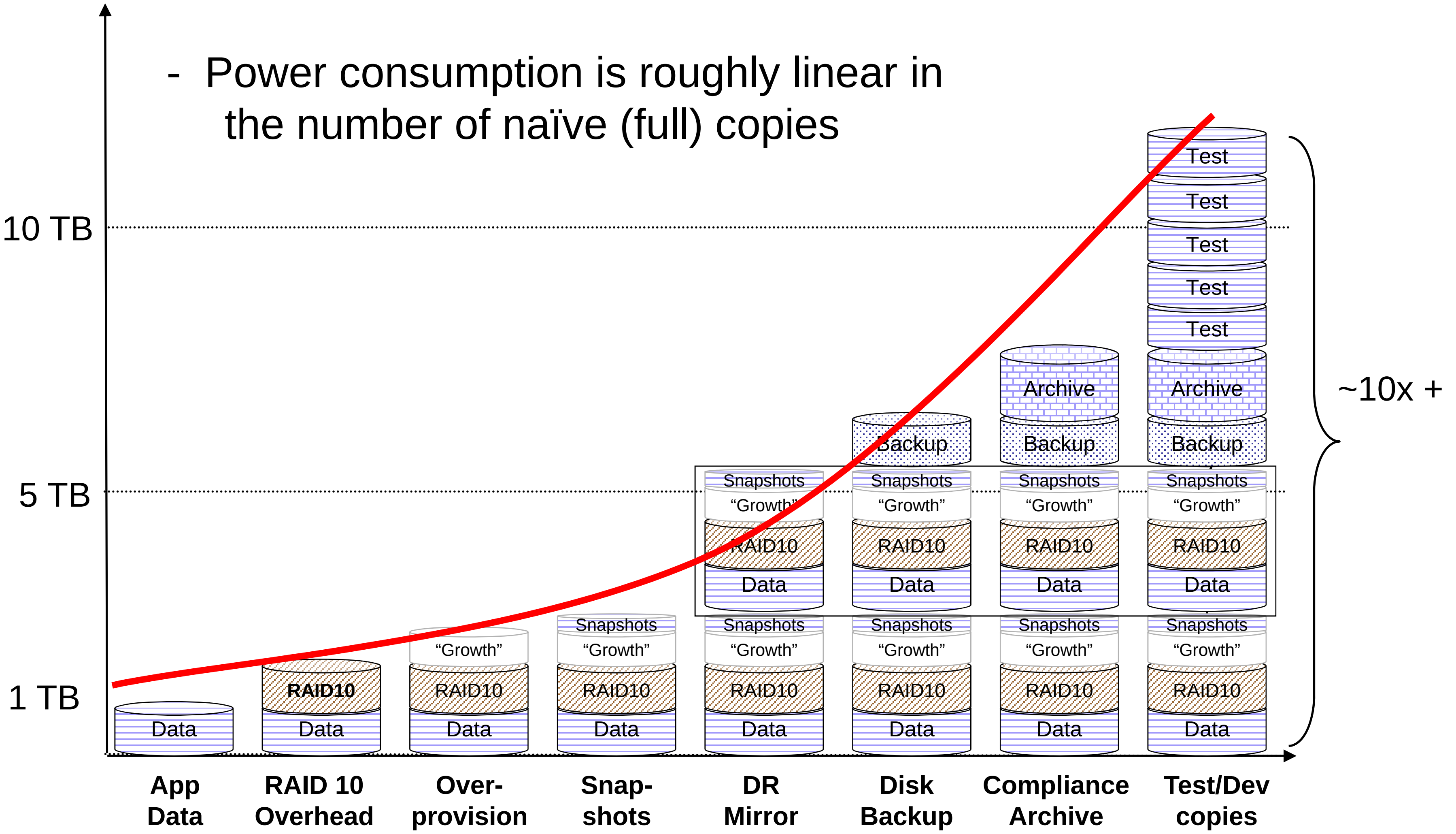


Problem: replication



- Traditional data center system redundancy
  - Overprovisioning – protect against volume-out-of-space application crashes
  - Test/dev copies – protect live data from mutilation by unbaked code
  - DR Mirror – protect against whole-site disasters
  - Backups – protect against failures and unintentional deletions/changes
  - Compliance archive – protect against heavy fines
- **Big data systems like Hadoop**
  - Typically 3x replication *locally*
- **“Brick” architectures**
  - Google, Microsoft
  - At least 2x

Result of redundancy



Green data center technology overview

- “Green” facility placement
- Water and natural cooling
- **Hot aisle technologies**
- Flywheel UPSes
- PUE monitoring
- **Thin provisioning**
- **Compression**
- **Delta snapshots**
- **Parity RAID**
- **Deduplication and SIS**
- **Capacity vs. high performance drives**
- ILM / HSM / Tiering
- MAID
- **SSDs / “Flash and stash”**
- Power supply and fan efficiencies

Green facilities

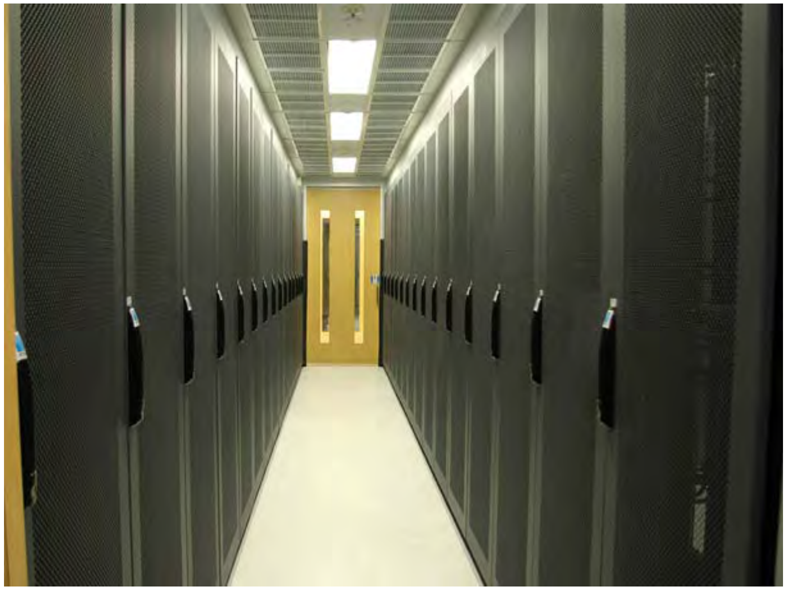
- PUE – Power Use Efficiency

$$PUE = \frac{Total\_Facility\_Power}{IT\_Power}$$

- Weighted upward by
  - UPS and power conditioning ineffieciencies
  - **Inefficient cooling**
- Traditionally 2.5, modern best practice = 1.25
- Can be gamed
  - Use of equipment fans to drive hot air exhaust

Hot aisle / cold aisle technologies

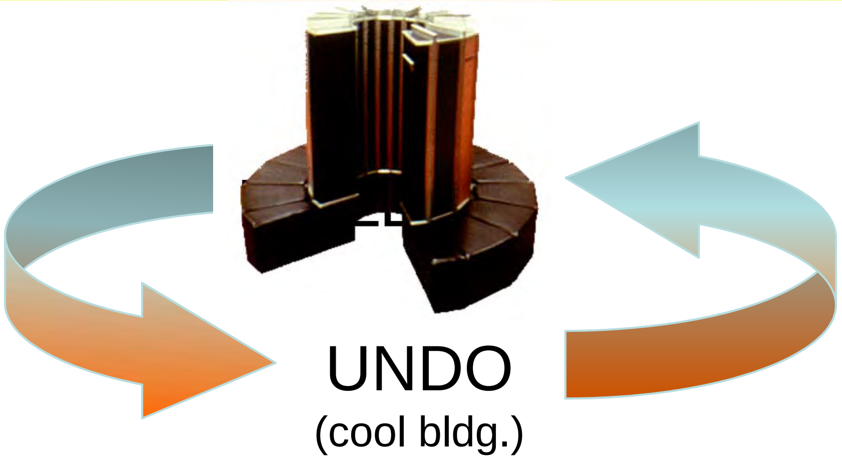
- Segregate airflows into hot and/or cold aisles (backs and fronts of servers)
  - More precise control
  - Allows higher temperature differentials (more efficient)
  - Current trend toward hot aisle containment with cold air plenum
  - Must-have: blanking plates
    - Very important
  - Normally deployed in comb. w/ air economizers



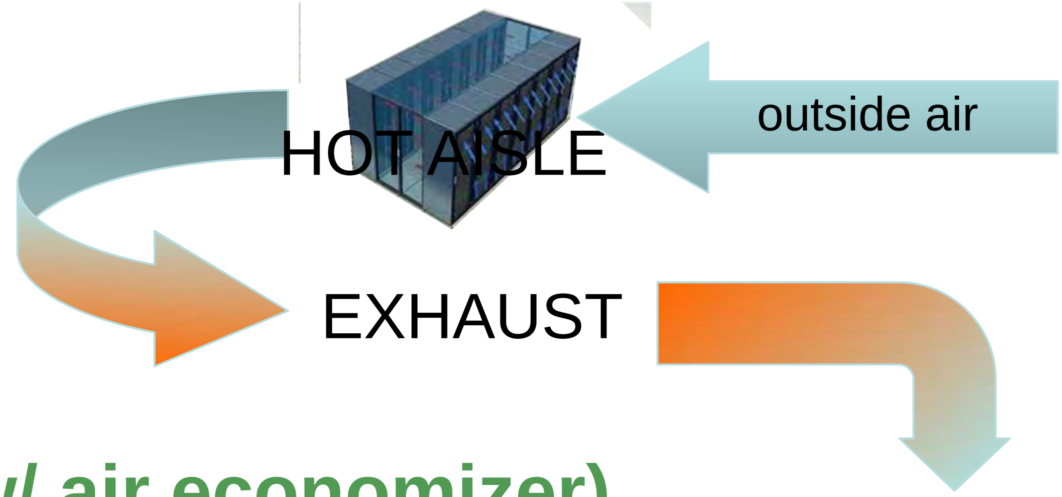
Green facilities

e.g.

BEFORE



AFTER

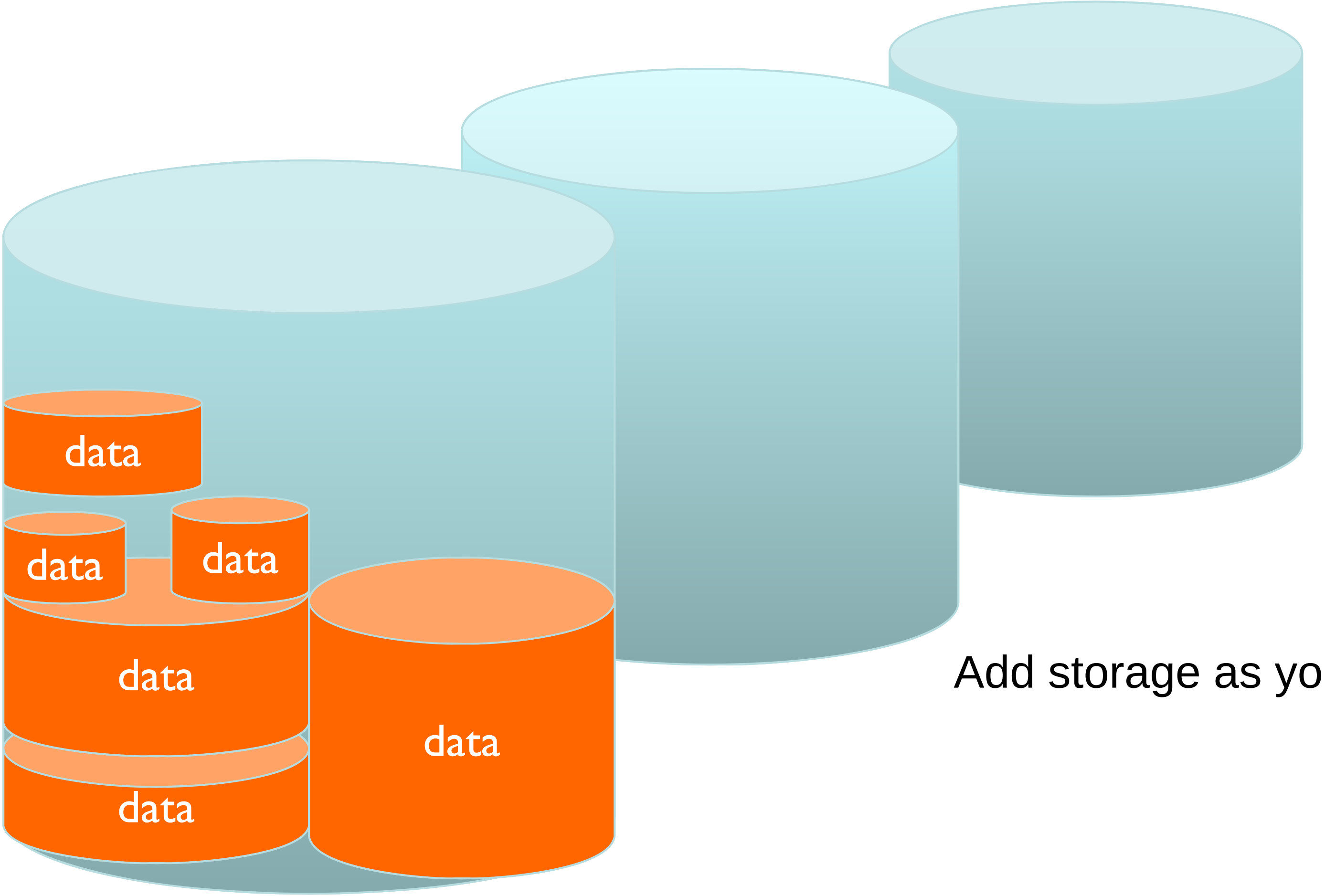


40% SAVINGS (w/ air economizer)
(PUE: 2.5 → 1.5)

Contributing to a green facility

- Hot aisle containment
  - Need temperature monitoring
    - Recomment rack level PDUs
  - Rigorous use of blanking plates required
  - With appropriate air economizers etc., the single biggest contribution to energy conservation you can make
- Note re “green power”
  - It’s only green if you generate it yourself (TGG)

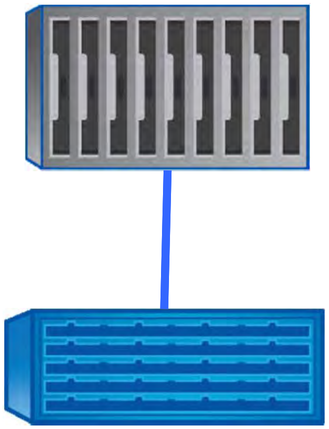
Thin provisioning



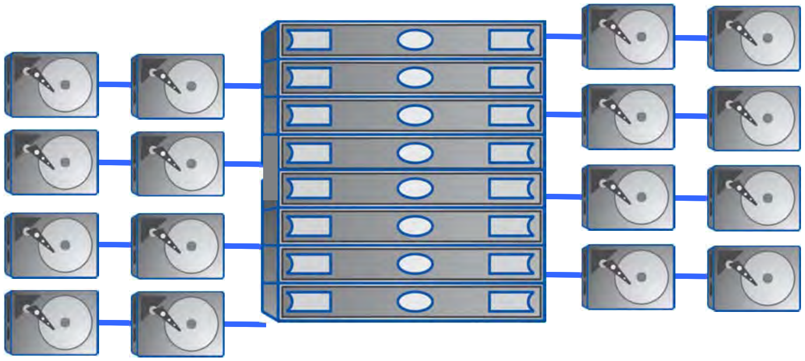
Add storage as you need it

Thin provisioning notes

- Most useful w/ variable ratios of compute and storage requirement
  - Especially true in cloud PaaS and SaaS environments
  - TP biases toward centralized storage

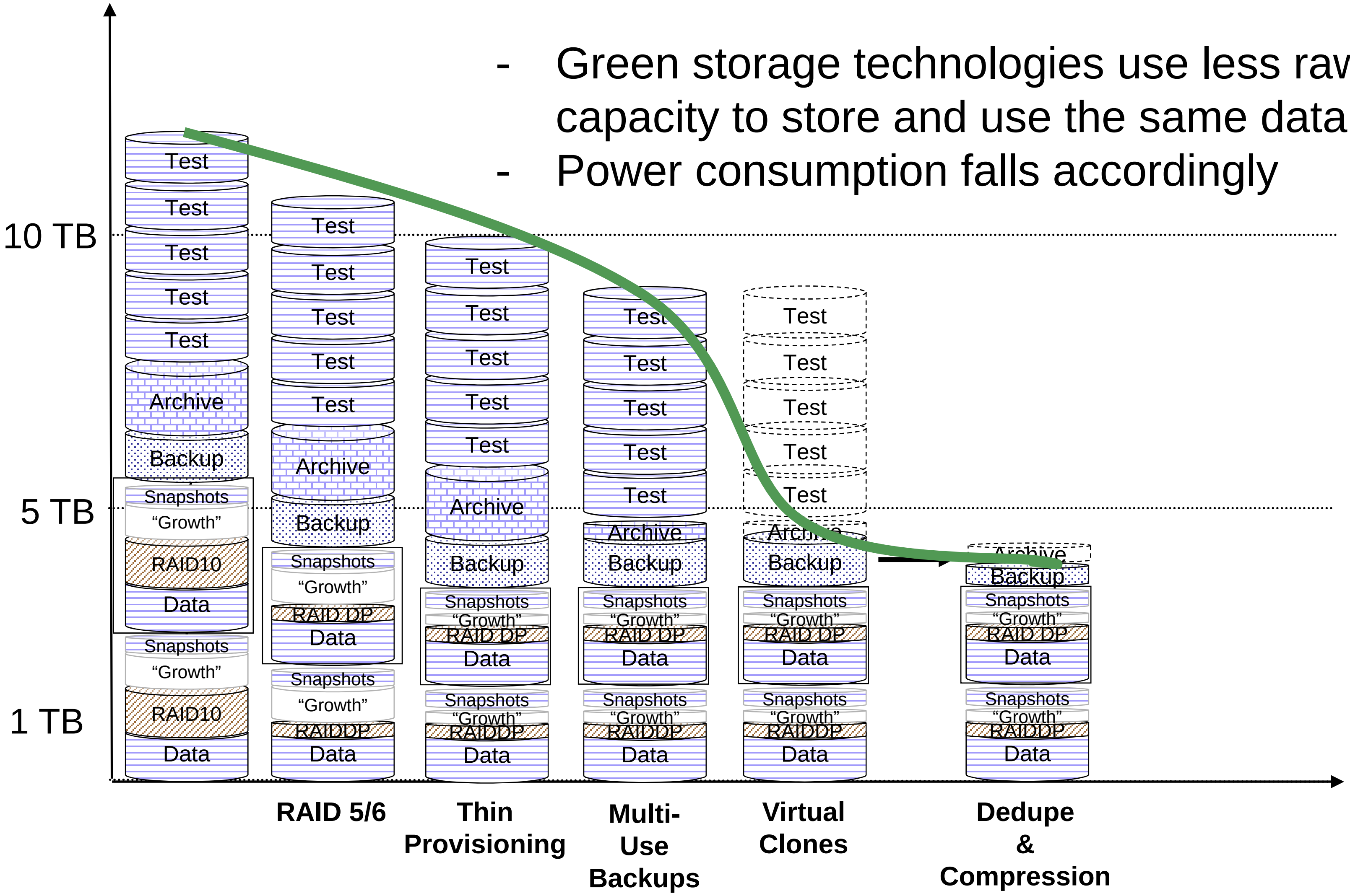


vs.



Solution: Capacity optimization technologies

- Green storage technologies use less raw capacity to store and use the same data set
- Power consumption falls accordingly



RAID

- May seem odd to call using extra disk “capacity optimization
- Requirement is to tolerate *n* failures
- Challenge is to do this with least amount of raw capacity
- RAID 1 – 50% overhead
- RAID 5 – 15% to 20% overhead
  - larger disks → unacceptable RG reconstruct times
- RAID 6 – 15% to 25% overhead (8 + 2 common)

Oak Ridge Labs “Spider”

Aggregate Bandwidth	240GB/s
Storage Systems	48 x DDN S2A9900 Storage Arrays
Hard Drives	13,440 1TB SATA Hard Drives
Aggregate Capacity	13.44 Petabytes (Raw), 10.7 Petabytes (Usable – 8+2 RAID 6)
Lustre Storage Servers	192 Lustre OSS Servers
Cabling	Over 1,000 20Gb InfiniBand Cables
Data Center Cabinets	32 Data Center Racks, 572 ft

Power savings on disk at ORNL

- Assume ~5.2W per SATA drive
- Assume 10,752 data drives

- RAID 6 (8+2)
  - 13,440 disks = 69.9kW, 6.5 W/TB (usable)
- Full duplication
  - 21,504 disks = 111.8kW, 10.4 W/TB
- Triplicate data
  - 32,256 disks = 167.7kW, 15.6 W/TB

LLNL “Sequoia”

Aggregate Bandwidth	> 1 TB/s
Storage Systems	480 x E5460 Storage “RBODs” (350K IOPs)
Hard Drives	23,200 3TB 6Gb/s SAS Hard Drives
Aggregate Capacity	76 Petabytes Raw, 56 Petabytes Usable – 8+2 RAID 6
Lustre Storage Servers	960 Lustre OSS Servers
Cabling	Over 1,000 20Gb InfiniBand Cables
Data Center Cabinets	48 Data Center Racks

Power savings on disk at LLNL

- Assume ~11W per drive
- Assume 23,200 data drives, 76PB

- RAID 6 (8+2)
  - 23,300 disks = 255kW, 4.64 W/TB (usable)
- Full duplication
  - 37,120 disks = 408kW, 7.4 W/TB
- Triplicate data
  - 55,680 disks = 612kW, 11.1 W/TB

> 1/3 of a megawatt difference between RAID 6 and triplicate

Green software technologies

- Compression
- Delta snapshots
- Thin provisioning
- Parity RAID
- Deduplication and SIS

Compression

- Compression
  - Old and venerable
  - Almost a “gimme”
  - Configuration matters
    - Compress before encrypting, decrypt before decompressing

Delta snapshots

- Data sharing
  - Form of deduplication
  - Data in snapshot shared with live data until one of them is written
  - Two fundamental techniques
    - Copy Out on Write
    - Write to new live location
- Mainly useful in HPC for VM booting
  - One storage system can support a couple hundred diskless servers

Deduplication and SIS

- Find duplicates at some level, substitute pointers to a single shared copy
- Block or sub-file based (dedup)
- Content or name based (SIS *, “file folding”)
- Inline (streaming) and post-process techniques
- Savings increase with number of copies found
- Performance hit may be too expensive for HPC

* SIS = Single Instance Store

Check out the SNIA Dictionary !

www.snia.org/dictionary

SSDs (Solid State Disks)

- Pros
  - Great READ performance
  - At rest power consumption = 0
  - No access time penalty when idle (cf. MAID)
  - No need to keep some disks spinning (cf. MAID)
- Cons
  - WRITE performance usually < mechanical disks
  - Cost >> mechanical disks except at very high perf points
  - Wear leveling requires a high space overhead
- Note: these dynamics changing rapidly with time

“Flash and stash”

- Large arrays of SATA-based disks fronted by large flash caches
  - > 1TB flash
  - Great for high I/O, esp. w/ contained working sets
  - Reduced power (SATA vs. SAS)
  - Not useful for write-intensive workloads

Power supply and fan efficiencies

- Efficiency of power supply an up front waste
  - Formerly 60-70%
  - Nowadays 80-95%
    - Climate Savers
    - 80plus group
- Variable speed fans
  - Common nowadays
  - Software (OS) control

Workload trends

- Many HPC workloads favor centralized storage and management
  - nonproliferation, counter-terrorism, energy security, etc.
  - health care analytics (SOX, HIPAA)
  - pharma research (FDA)
  - weapons
- Security and retention/compliance
  - both easier and better with the enforcement point in the storage layer
  - physical security still necessary

Other points

- Centralized storage also offers
  - MUCH better data management
  - Very efficient DR and remote replication
  - Strong vendor support
  - Great virtualization support
- But
  - capabilities may be wasted in extreme HPC environments
  - increased up-front cost often a psychological barrier

Key takeaways

- Hot aisle containment
- RAID 6 is current best practice
- Headless server configurations are possible
- Use of software capacity optimizations highly recommended in cloud-style environments