Power Use of Disk Subsystems in Supercomputers

Matthew L. Curry

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

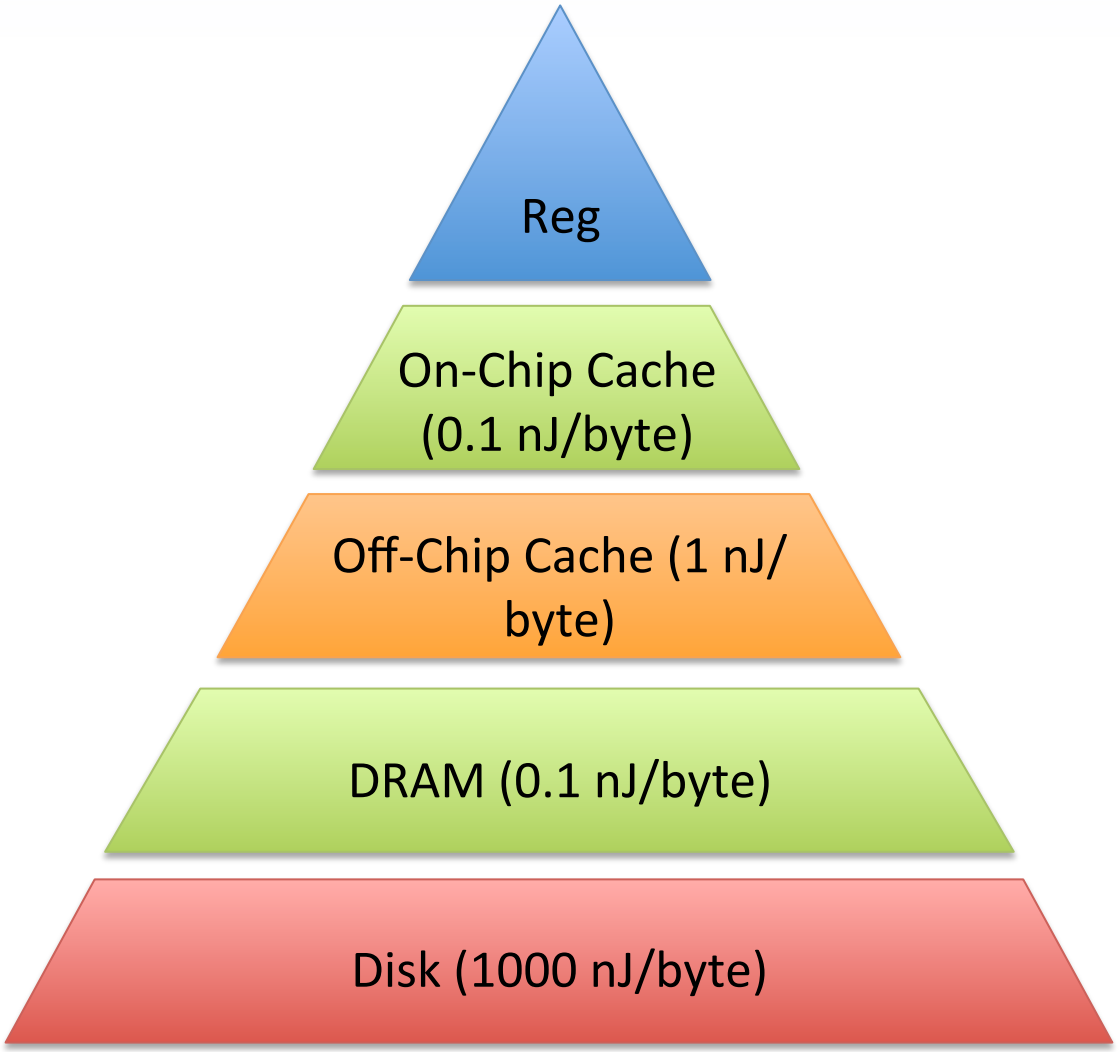
Supercomputer Power Use and Exascale

- DOE Plan: First exascale machine will consume up to 20 MW, or 50 GF/W
- The June 2011 Green500 list has a BG/Q prototype as the most efficient machine
  - 2 GF/W
  - In the next decade, machines need to be 25x more power efficient!
- Where can we find more power efficiency?

Sandia National Laboratories

Memory Hierarchy and Power

- The first reaction is often to look at which operations require the most power
- Disks are far away and (most) have moving parts
- How much power does storage really use for real application behavior?



Sandia National Laboratories

A Study of Power in Supercomputing

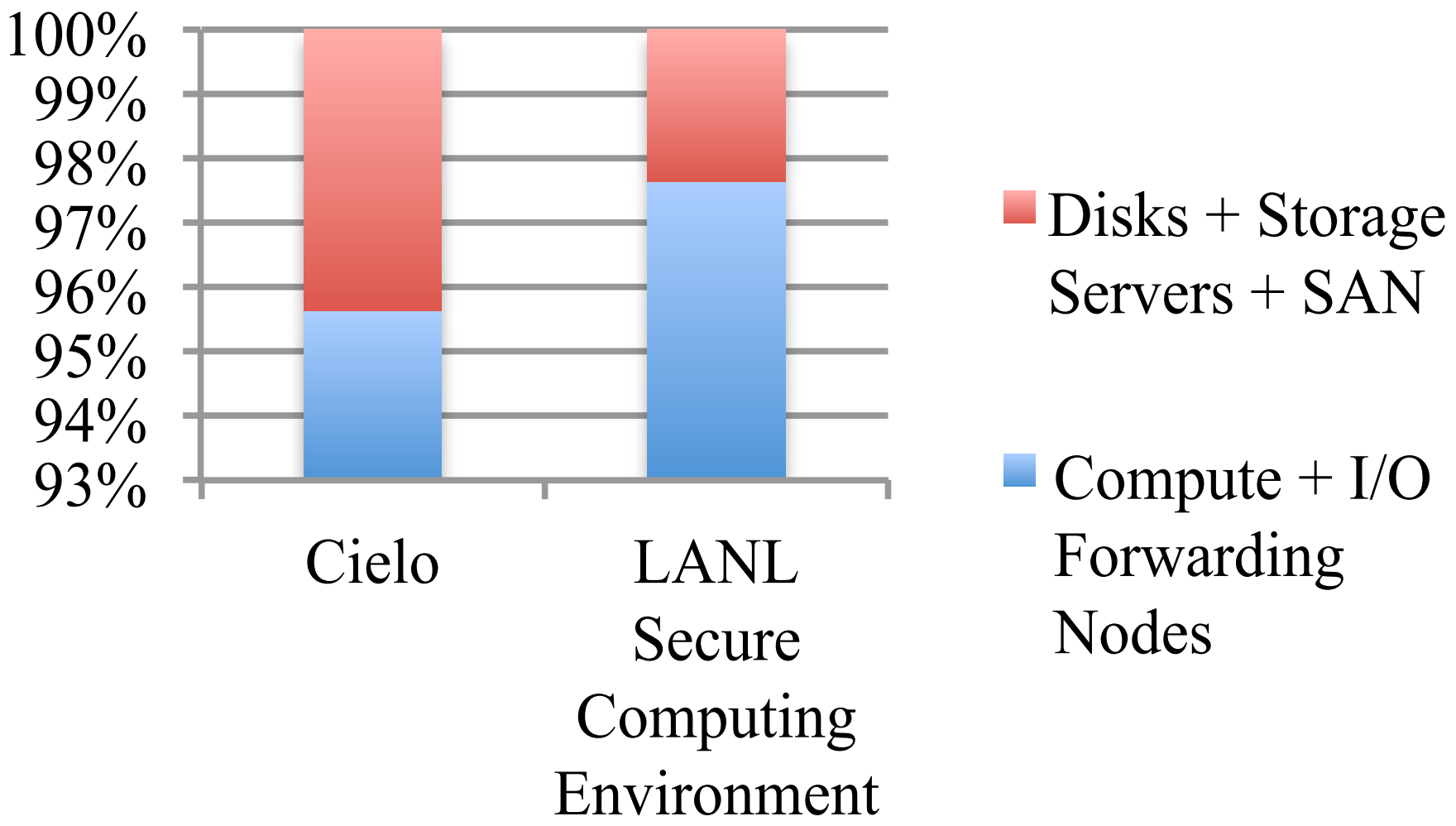
- Survey three sites with large machines
  - Los Alamos: Roadrunner, #10, and others
  - Los Alamos/Sandia ACES: Cielo, #6
  - Sandia: Red Sky, #16
  - Clemson University's Palmetto, #96
- Asked for power data from compute and I/O infrastructure separately
  - No cooling, external infrastructure, etc. Just compute, I/O servers, disks.

Sandia National Laboratories

Los Alamos Description

- Two separate methods of sampling
  - Cielo individually
    - 4.7-6.7 MW
    - 1.1 PF (~143k cores)
    - 10PB of dedicated Panasas storage
  - Secure Computing Environment, which includes Cielo, Roadrunner, capacity clusters, etc.
    - 16.5 MW typical
    - 3.5 PF
    - 20 PB of Panasas storage, with 10PB served to all machines except Cielo via a 10GigE fabric

Los Alamos Results



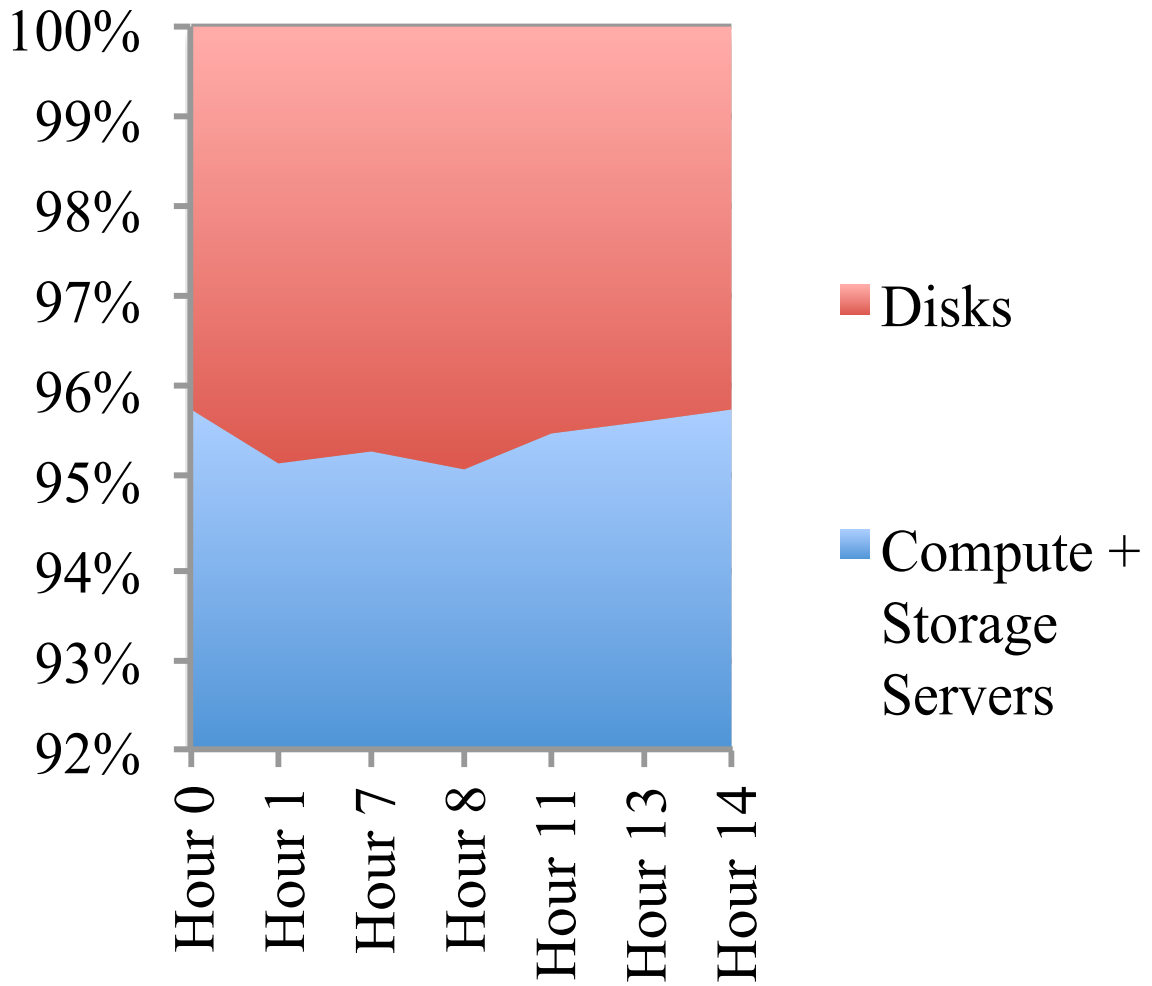
Sandia National Laboratories

Sandia Description

- Red Sky/Red Mesa is the premier capacity platform for Sandia and NREL
  - 3 PB
  - 433.5 PF (~42k cores)
- One rack of storage and compute measured throughout a single day
- Extrapolated to unclassified section of Red Sky, which is approximately 56% of the Red Sky/Red Mesa machine

Sandia National Laboratories

Sandia Results



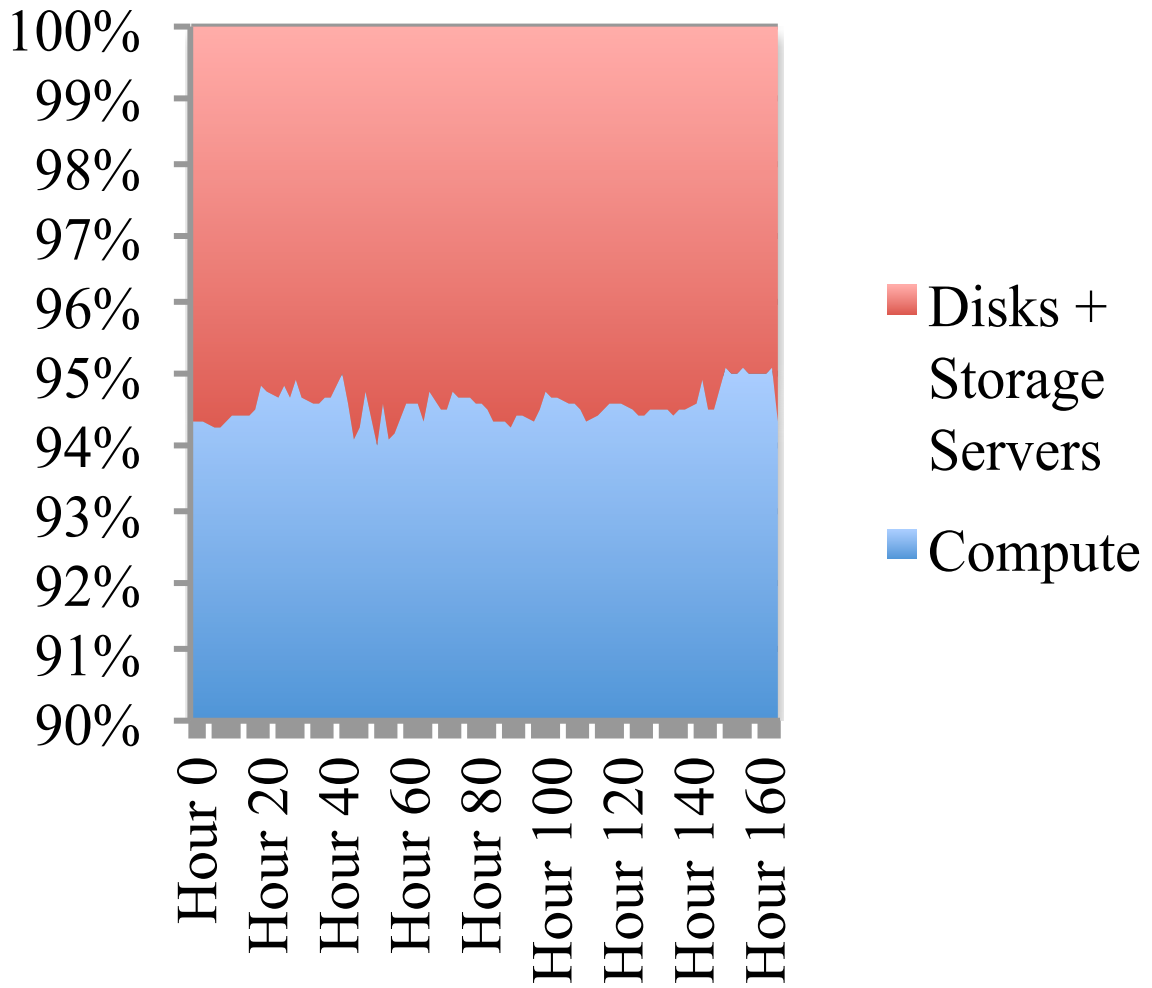
Sandia National Laboratories

Clemson Description

- Capacity, condominium cluster at Clemson University
  - 92TF, ~14k cores
  - 616TB
- Data collection at two-hour intervals over two weeks
  - Storage infrastructure used mostly constant power throughout

Sandia National Laboratories

Clemson Results



Sandia National Laboratories

Extrapolating to Exascale

- Exascale storage systems will require 320PB-1EB of storage at 106.7 TB/s
  - 32PB main memory
  - Checkpoint every hour
  - 95% (57/60 minutes) must be spent computing
- Predictions for future disks (~30TB capacity, ~380 MB/s bandwidth) dictate 277k disks!
  - 66% of power budget if power per disk remains constant

Sandia National Laboratories

Burst Buffer

- Grider has detailed in many presentations a "burst buffer" idea for checkpointing
  - Quickly accept a checkpoint in smaller flash store
  - Bleed flash to slower disk-based storage between checkpoints
- It has been shown that this will work from a purchase price standpoint
  - Power?

Sandia National Laboratories

Flash Characteristics

- Current flash (e.g., Intel 320 series) can accept 1MB/s per gigabyte of capacity
  - Even today, 90PB of flash (to hold three checkpoints) is sufficient to sustain 90TB/s of bandwidth
- Use 10TB/s disk-based store
  - Requires 25k disks, which may hold 738 PB
  - Extrapolating from today's disk power, this is 6% of the power budget
  - Flash uses a comparable amount of power, yielding 6.6% of 20MW for disk and flash

Sandia National Laboratories

Conclusion

- I/O consumes a low proportion of power within the machine
  - 4.4-5.5%
- One exascale storage model, the burst-buffer scheme, can be done with 6.6% of the power budget
- Inefficiencies in the power feed systems of the data center can be a larger consumer of power!
- We should always be on the lookout for ways to be more efficient
  - Especially for workloads that *aren't* checkpointing

Sandia National Laboratories

Acknowledgements

- Authors
  - Lee Ward, Sandia
  - Gary Grider, Los Alamos
  - Jill Gemmill, Clemson
  - Jay Harris, Clemson
  - Dave Martinez, Sandia

Sandia National Laboratories