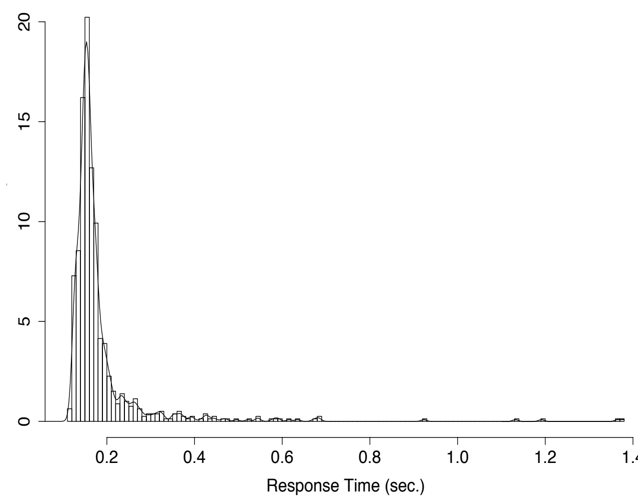
Heavy-tailed Distribution of Parallel I/O System Response Time

Bin Dong, Surendra Byna, and Kesheng Wu
Scientific Data Management group
Lawrence Berkeley National Laboratory, Berkeley, CA



PDSW2015: 10TH Parallel Data Storage Workshop, Austin, TX, November 16, 2015

Outline

- Motivation
- Response time sampling method
- Analysis results of response time

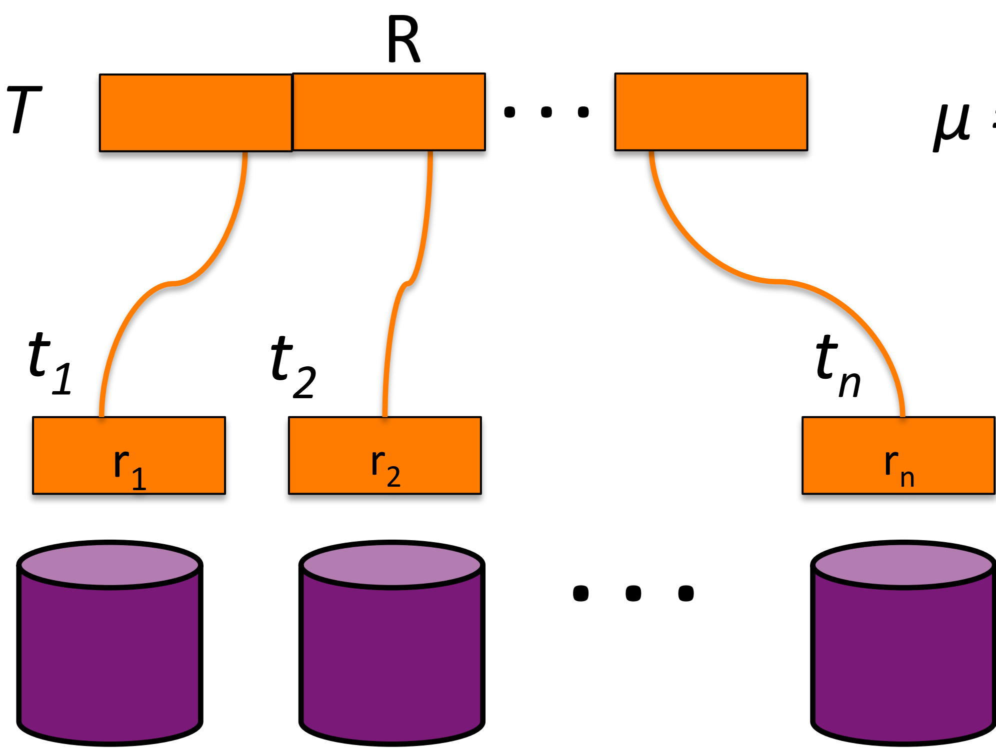
Estimating Response Time of I/O is Essential Element

- Data analysis query plan optimizing
  - Choose index or data organization with minimum read time
  - Scientific Data Services (SDS) framework, PostgresSQL, SciDB
- Data writing performance tuning
  - Select striping size, striping account, and other parameters to reduce write time
  - ExaHDF5, I/O Scheduler
- Simulator, Job Scheduler , Quality of service (QoS), etc.

Modeling Response Time for Parallel I/O

Response time of a single big file request R:

$$T = max\,(t_1, t_2, \; . \; . \; . \; , \; t_n) + \mu$$



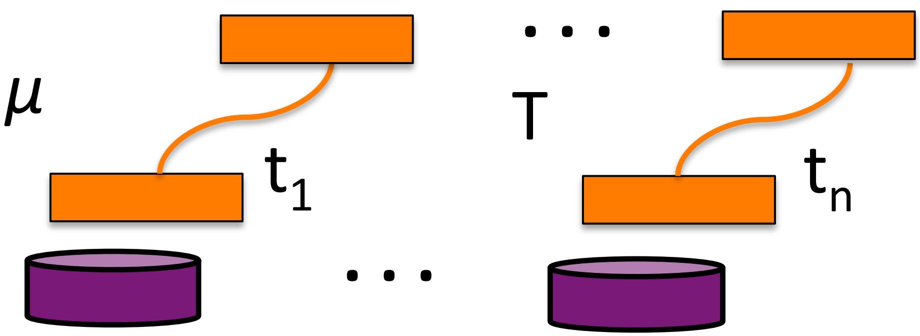
$\mu = \begin{cases} \text{Split overhead,} & \text{write} \\ \text{Merge overhead,} & \text{read} \end{cases}$

$t_1, t_2, \; . \; . \; ., \; t_n$: response times of n small requests

I/O Servers in PFS (e.g., OST in Lustre)

Simplifying Response Time Model

$T = max\,(t_1, t_2, \ . \ . \ . \ , \ t_n) + \mu$



- Split/merge overhead μ is constant
- *n* small requests ≈ *n* sampling (**i.i.d.**) of *n* IO Servers
- $t_1, \ldots, t_n$ ≈ *n* **i.i.d.** statistical variables
- Focus study on one (denoted by *t*) among $t_1, \ldots, t_n$
  - t : continuously distributed variable on $(0, +\infty)$

Applying Order Statistics to Estimate *T*

$$T = maximum\,(t_1, \ldots, t_n) + \mu$$

t : continuously distributed variable on $(0, +\infty)$

$F_t(x)$: distribution function of t

$f_t(x) = F_t{'}(x)$: density function of t

- *Step 1* : Compute density function $f_{Yi}(y)$ *with* $F_t(x)$ *and* $f_t(x)$
  - Y_i : the ***i*-th** largest value $(t_1, t_2, \ldots, t_n)$
  - $f_{Yi}(y) = F(y)^{n-i}(1-F(y))^{n-i} f_t(y)\, n!/[(i-1)!(n-i)!]$

  } Order Statistics
- *Step 2* : Compute response time $T = Y_n$

Problem Statement

- What is the distribution function *F(t)* for the response time of each small file request?
  - Existing researches assume
    - Uniform Distribution
    - Normal Distribution
  - Are these assumptions true ?
  - If not, are there other distributions fitting better ?

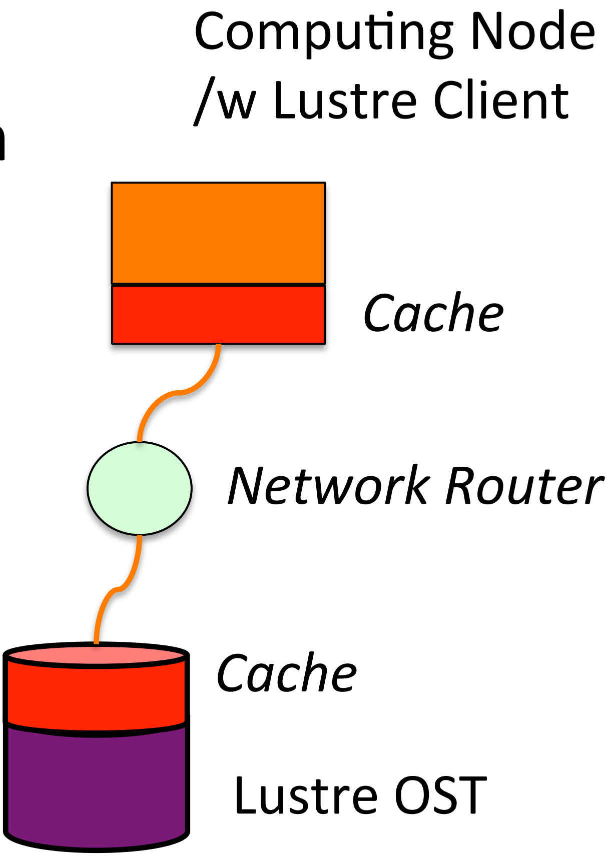
Our Method

- Sample the response time of two production storage systems

- Analyze statistical properties of response time

Response Time Sampling Environments

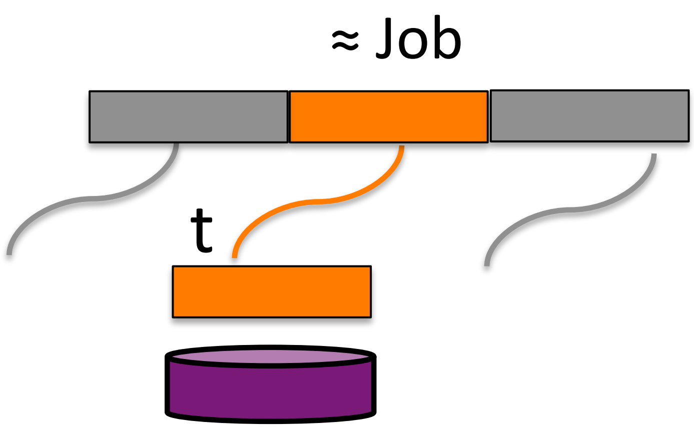
- Hopper and Edison at NERSC[1]
  - *153K and 130K* CPU cores, *1.28* PF and *2.39*PF
  - *5000* registered users
  - *300* online active users on Edison
  - I/O Intensive jobs use Lustre
- Lustre file system
  - Cache on client and I/O server
  - Network latency
  - 1 ~ 143 OSTes



[1]*National Energy Research Scientific Computing Center*
https://www.nersc.gov/

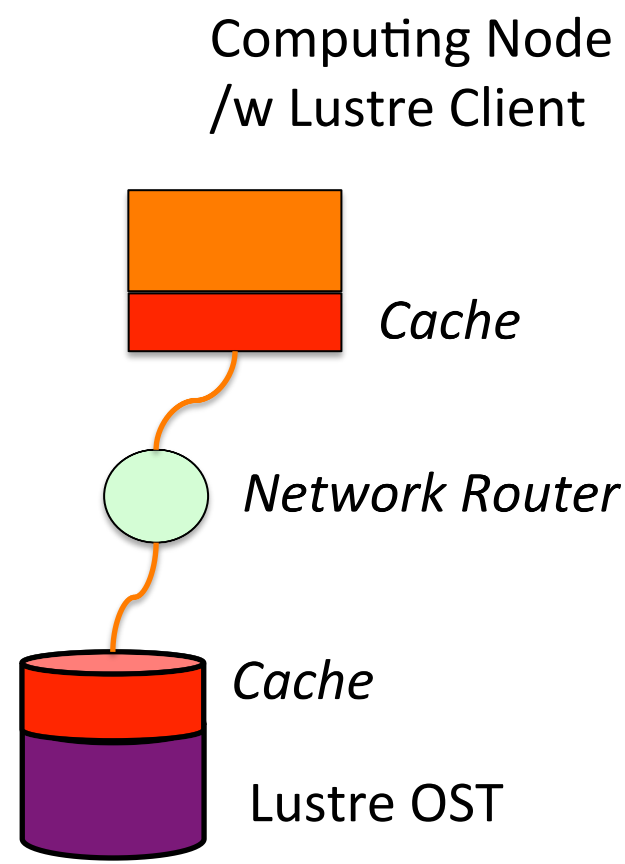
Sampling Method

- One job sampling one OST
  - A job ≈ A small file request
  - Measure time of reading and writing separately
  - Test different reading/writing sizes
    - 12 different sizes: 512KB, 1MB, 2MB, … , 1024MB
  - Match request size and striping size



Sampling Method

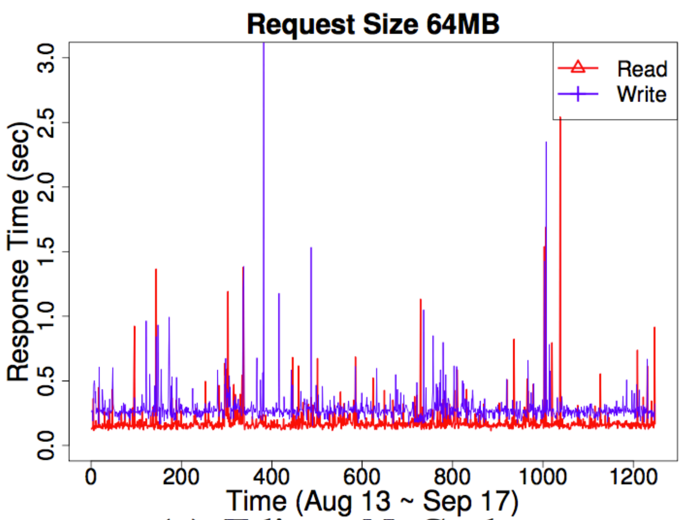
- Measure response time on computing node
  - network, disk, cache
- Cache Consideration
  - No Cache
    - clear cache by accessing memory sized data before sampling
    - call fsync() after write
  - Cache
    - High frequently sampling



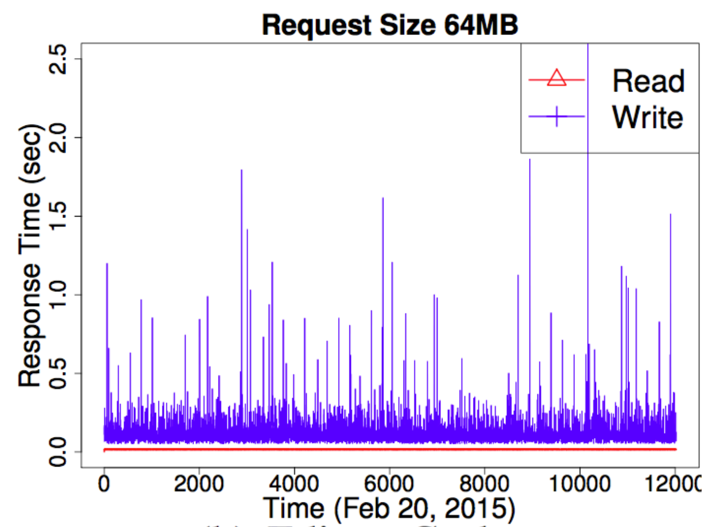
Sampling Results Statistics Overview

	Start Time	End Time	Days	# of Sampling	# of OSTs
Edison-No-Cache	08/13/2014	09/17/2014	35	14,977	12
Edison-Cache	02/20/2015	02/20/2015	1	927,691	12
Hopper-No-Cache	10/01/2014	01/13/2015	104	13,868	12
Hopper-Cache	02/20/2015	02/20/2015	1	1,581,364	12
		Summary	141	2,537,900	48

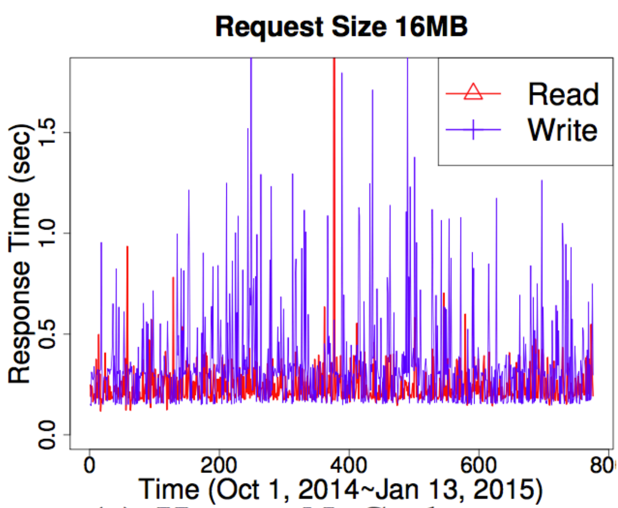
Variability of Raw Response Time for Edison and Hopper, Cache and No-Cache



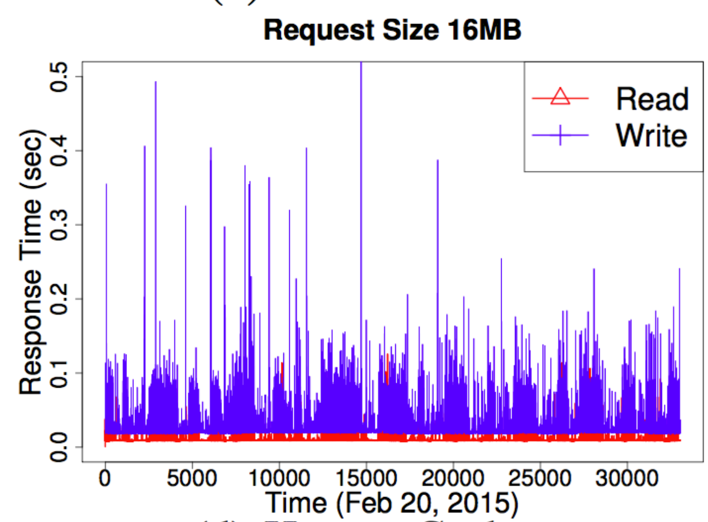
(a) Edison-NoCache



(b) Edison-Cache

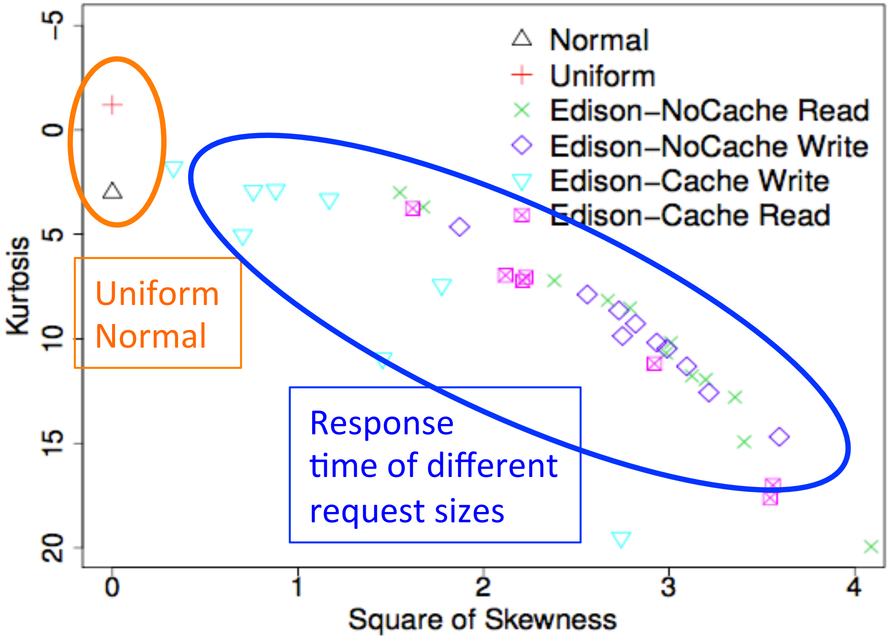


(c) Hopper-NoCache

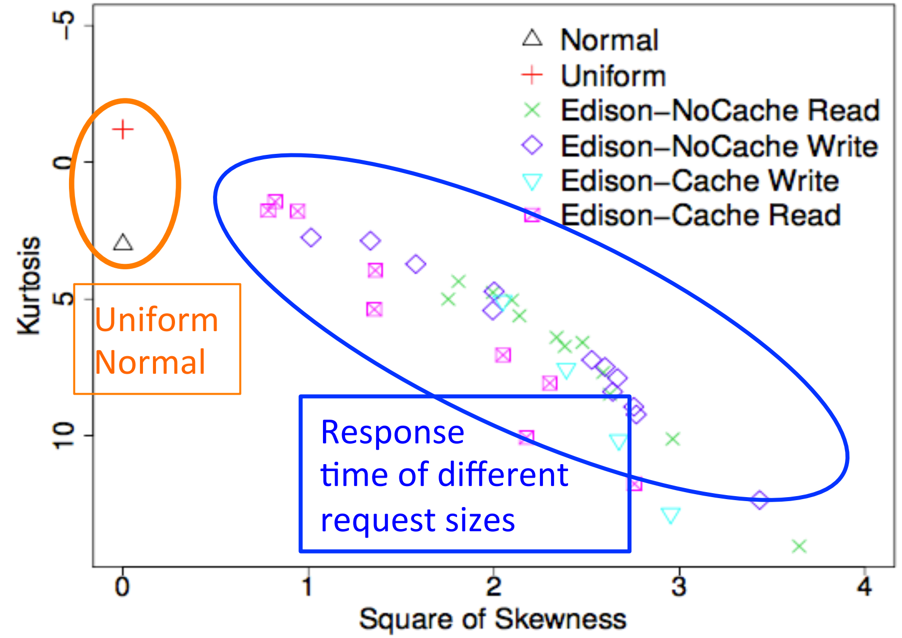


(d) Hopper-Cache

Ill-fit of Uniform or Normal Distribution



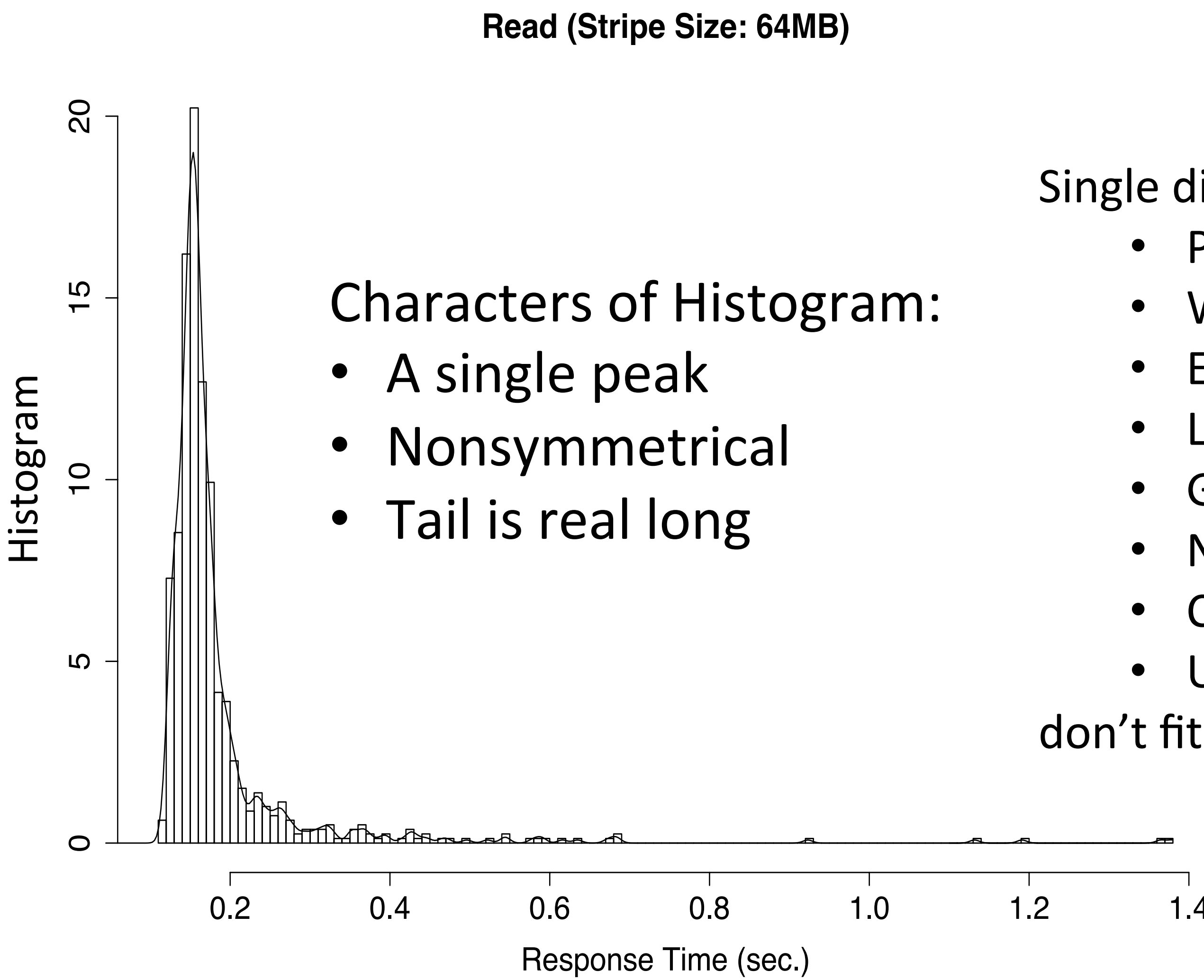
(a) Edison



(b) Hopper

Metrics	Uniform	Normal
Kurtosis	- 1.2	3
Skewness	0	0

Ill-fit of Uniform, Normal, and Other Single Distribution Function



Characters of Histogram:

- A single peak
- Nonsymmetrical
- Tail is real long

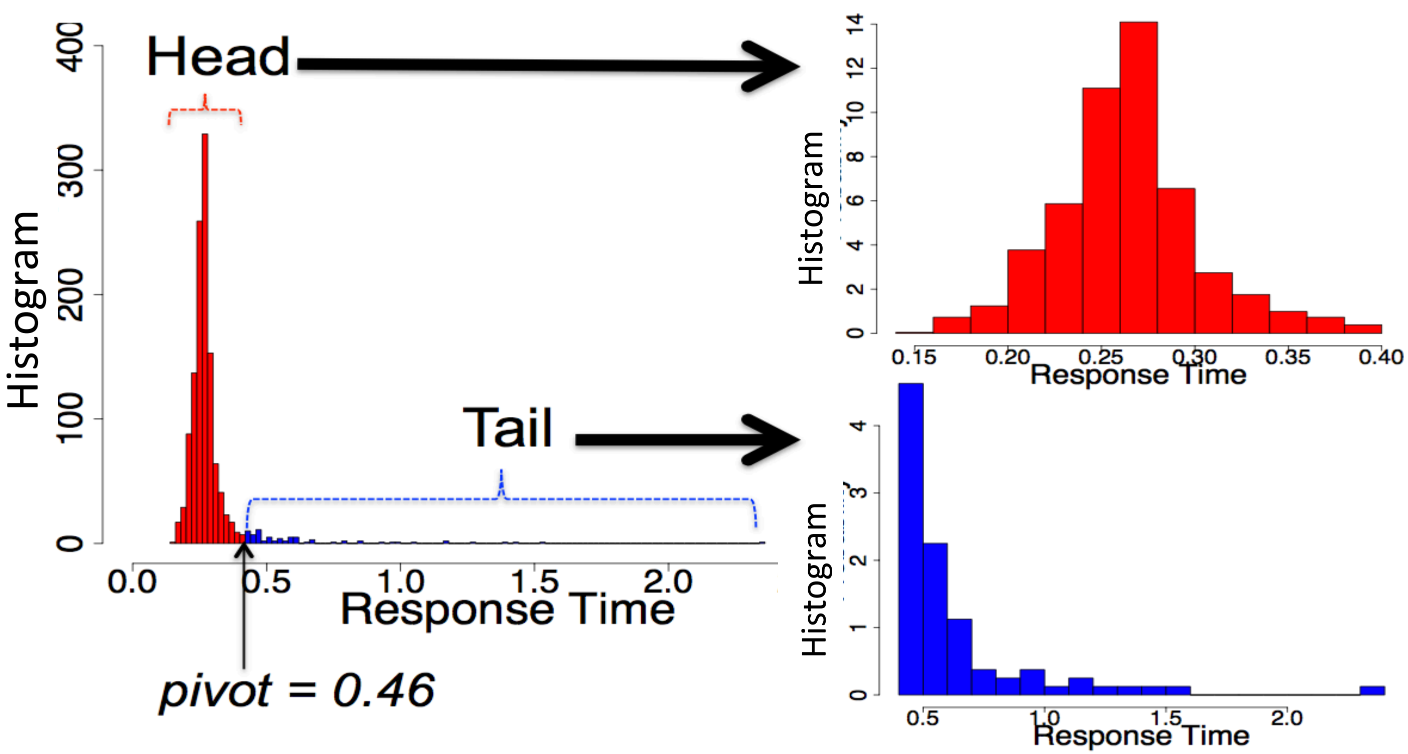
Single distribution functions

- Power Law
- Weibull
- Exponential
- Log Normal
- Gamma
- Normal
- Cauchy
- Uniform

don't fit very well !

Exploring New Distributions

- Partition response time into Head and Tail
- Find the pivot
  - minimizing KS (Kolmogorov-Smirnov) distances

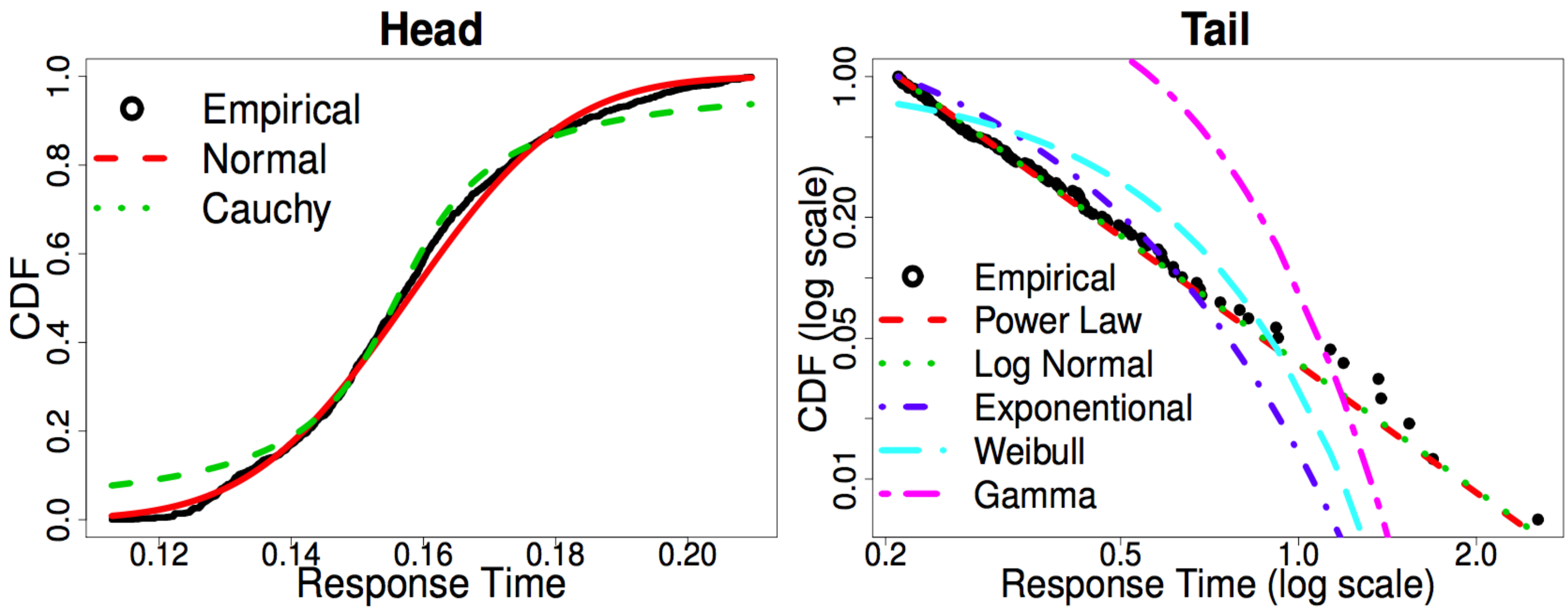


- Normal
- Cauchy

- Power Law
- Weibull
- Exponential
- Log Normal
- Gamma

Fitting Results

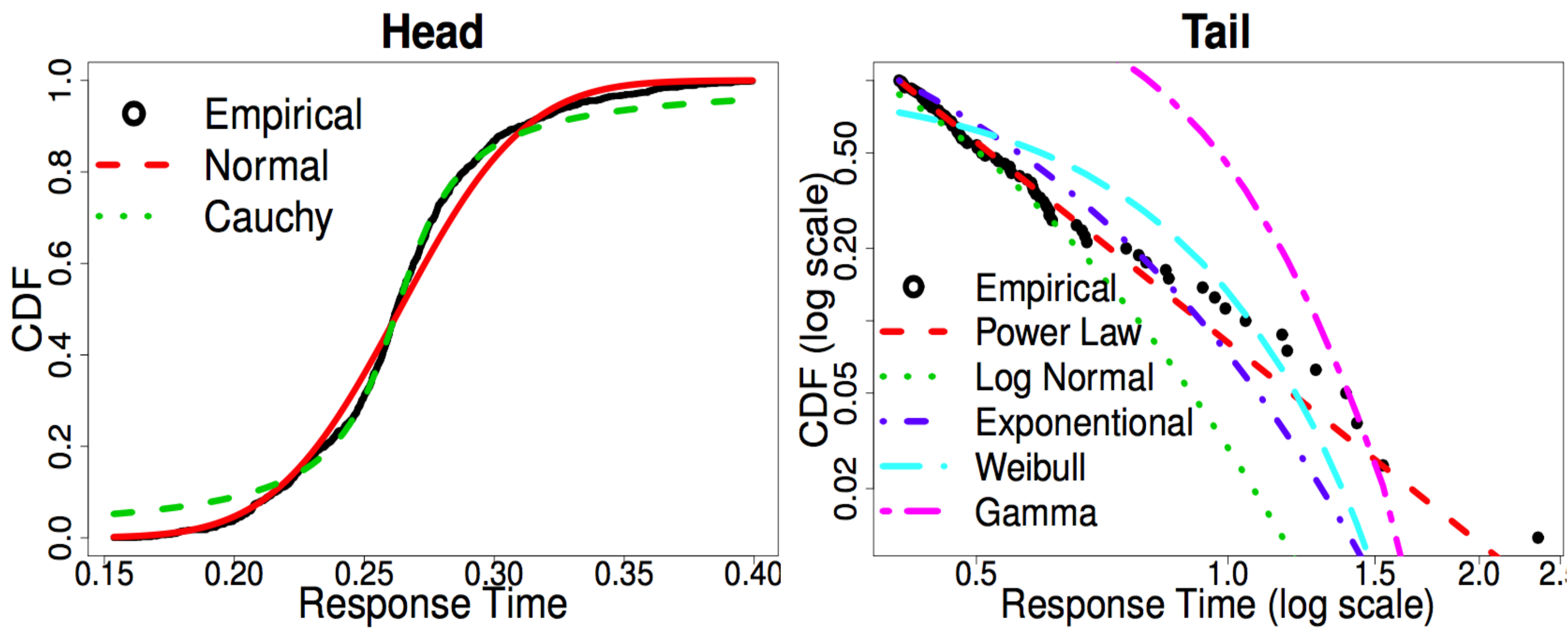
- Edison–NoCache, Read Response Time, 64MB



	Accuracy
Head Group	Normal > Cauchy
Tail Group	Power Law > Log Normal > Exponential > Weibull > Gamma

Fitting Results

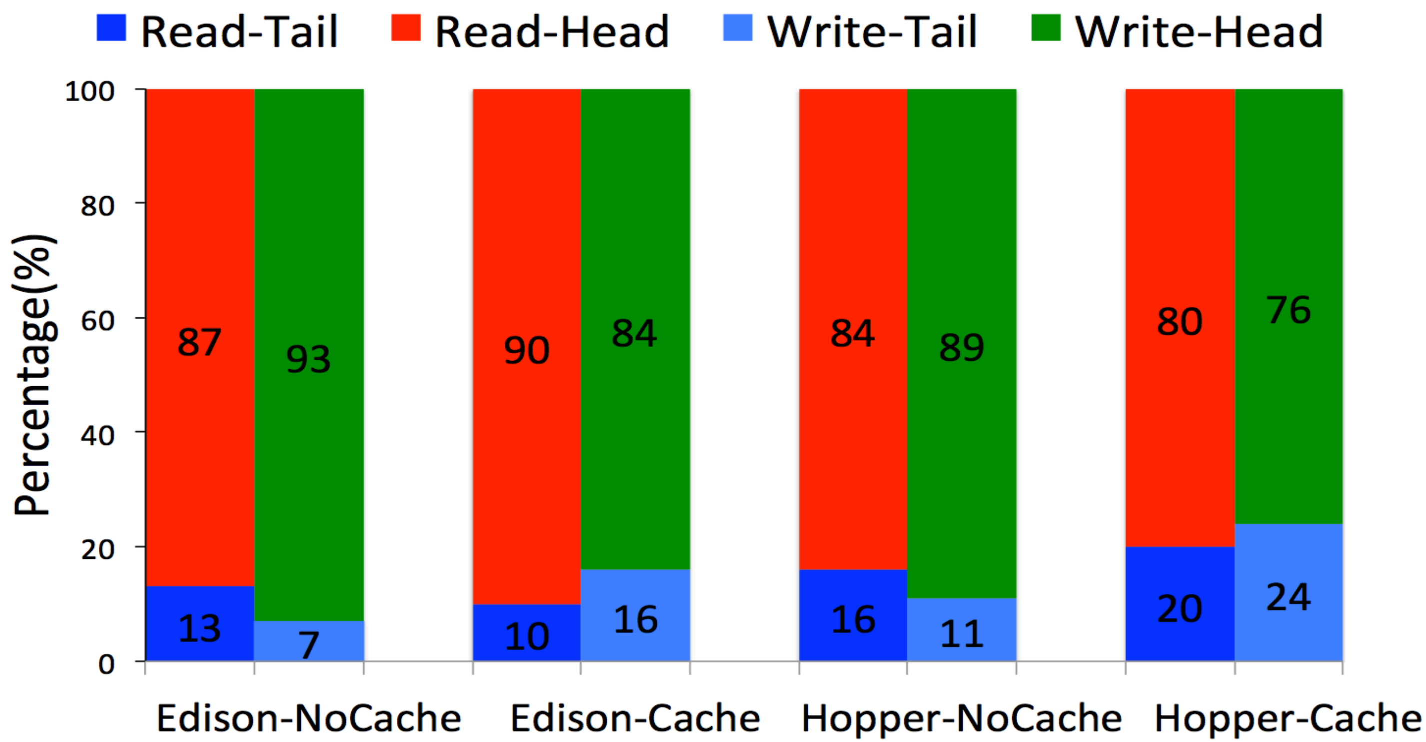
- Edison–NoCache, Write Response Time, 64MB



	Accuracy
Head Group	Normal > Cauchy
Tail Group	Power Law > Weibull > Exponential > Log Normal > Gamma

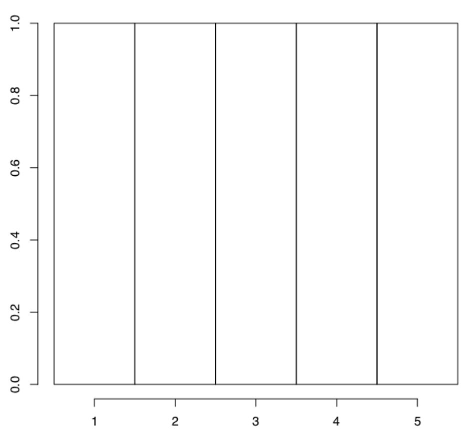
Percentage of Head group and Tail group

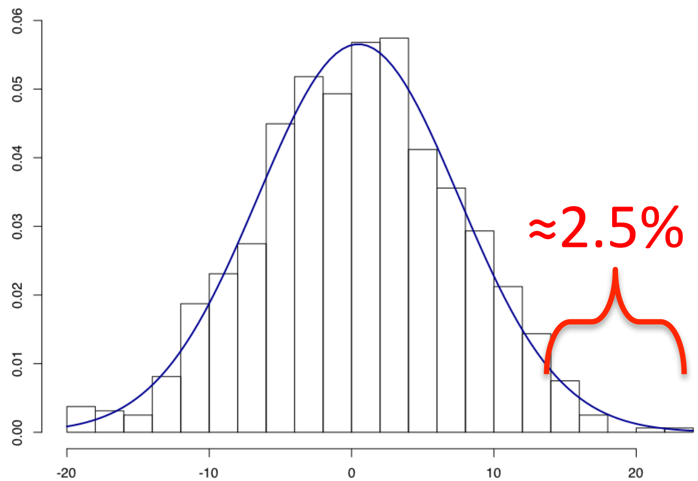
- 85% in Head group (i.e., small response time)
- 15% in Tail group (i.e., long response time)

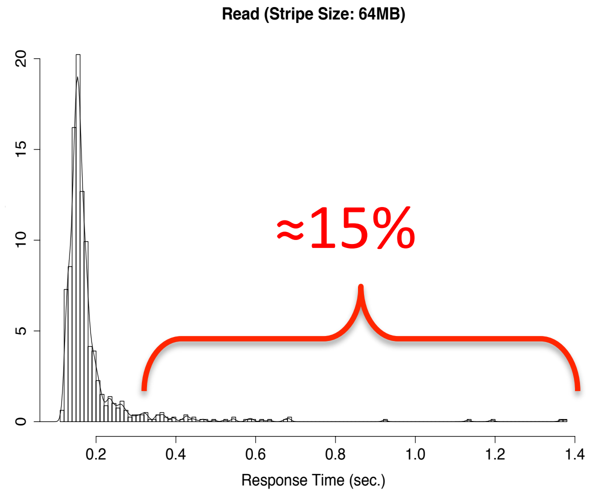


What is Wrong with Using Normal or Uniform?

	Long Response Time (Rare Event)
Uniform Distribution	All equal
Normal Distribution	2.5%
Real Storage Systems (Edison and Hopper)	15%







Summary

- Distribution function of response time of storage system is essential in estimating I/O performance
- We collected *2,537,900* response time sampling from 48 OSTes of *2* petascale storage systems across *141* days
- We found that single Normal or single Power law does not fit the response time
- We found that "Normal + Power law" fits response time better
- Future work
  - sample other storage systems
  - build accurate performance model
  - apply model to applications

Acknowledgments

- Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy (Program manager: Lucy Nowell), support for the SDS project under contract number DE-AC02-05CH11231

- National Energy Research Scientific Computing Center

Thanks, Questions ?

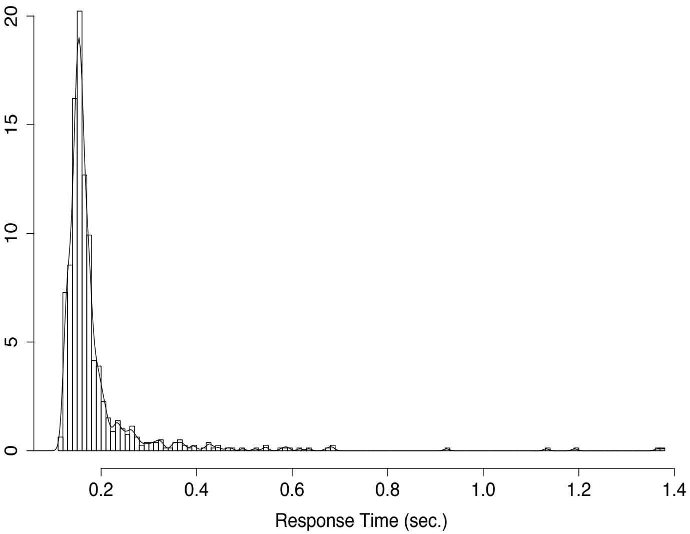
- other questions, please email to: ***dbin@lbl.gov***

Heavy-tailed Distribution of Parallel I/O System Response Time

Bin Dong, Surendra Byna, and Kesheng Wu

Scientific Data Management group

Lawrence Berkeley National Laboratory, Berkeley



PDSW2015: 10TH Parallel Data Storage Workshop, Austin, TX, November 16, 2015