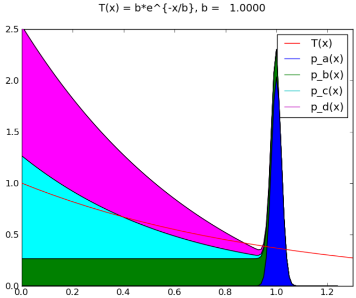
The Statistical Properties of Lustre Server-side I/O

A work in progress

Lustre User Group April 12, 2011



Andrew Uselton
National Energy Research Scientific Computing Center
Lawrence Berkeley National Lab

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

Contents

1. LMT: The Lustre Monitoring Tool
2. LMT Use Cases
3. I/O System Balance
4. Occurence Histograms
5. A Simple Model
6. Conclusions

Outline

1. LMT: The Lustre Monitoring Tool
2. LMT Use Cases
3. I/O System Balance
4. Occurence Histograms
5. A Simple Model
6. Conclusions

LMT: The Lustre Monitoring Tool

LMT Use Cases

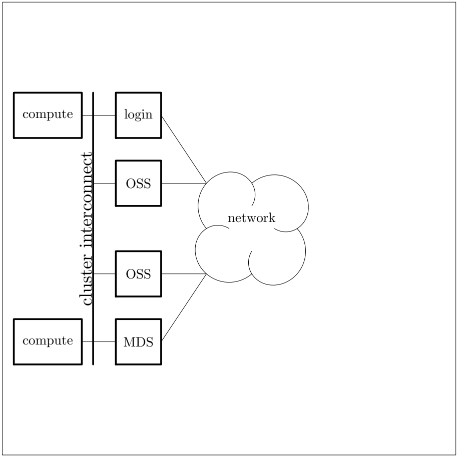
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

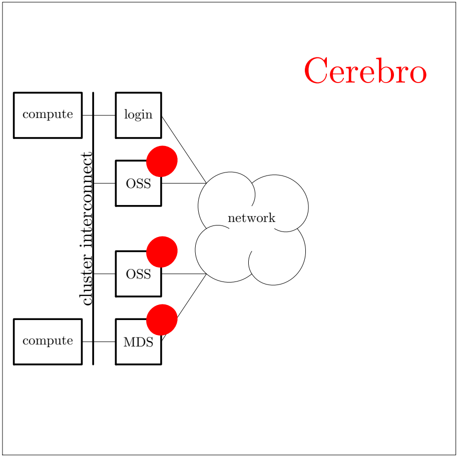
System layout



- milage may vary

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

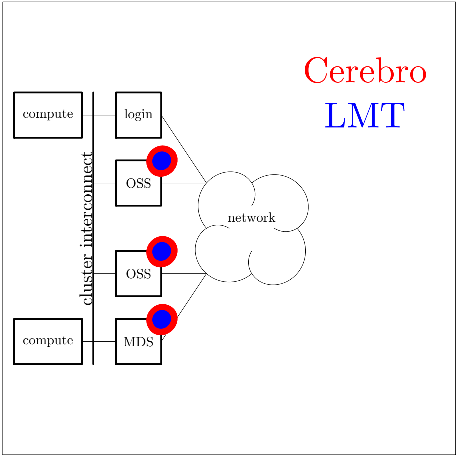
Cerebro



- lightweight
- extensible
- handles data transfer

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

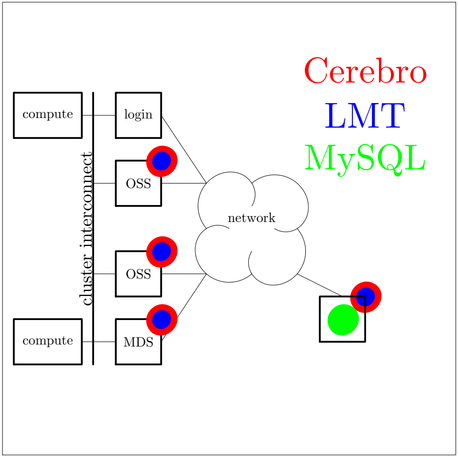
LMT



- compiled libraries
- one per sever
- harvests `/proc` values

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

MySQL



- daemon receives packets (UDP)
- library processes contents
- db stores values
- cron job summarizes (optionally ages)
- misc. tools for querying db

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

Data

- Bytes read
- Bytes written
- Inodes available
- Queue depths
- Operations (eg. `open()`) per second
- *many more*

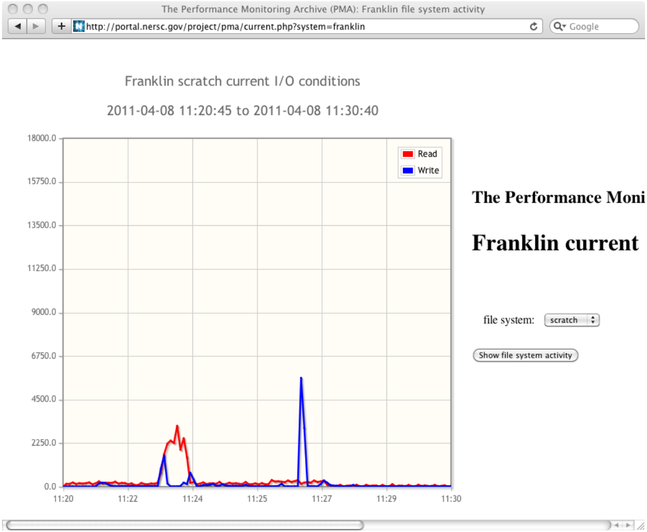
Data

- Bytes read
- Bytes written
- Inodes available
- Queue depths
- Operations (eg. `open()`) per second
- *many more*

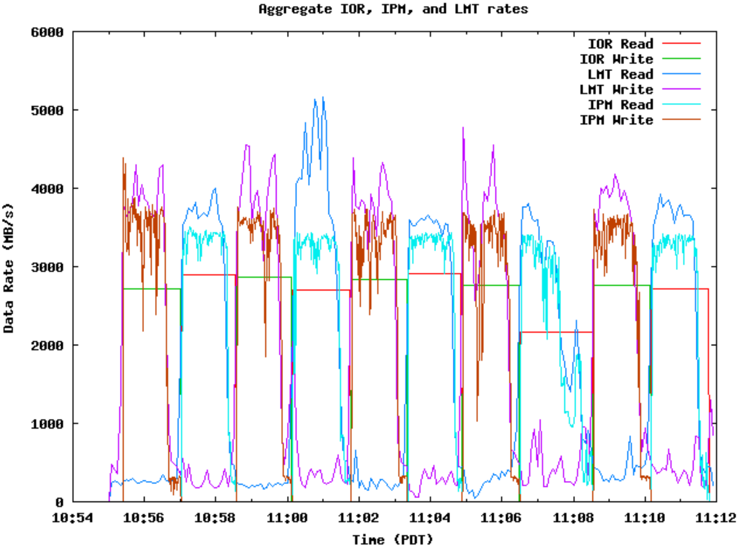
Outline

1. LMT: The Lustre Monitoring Tool
2. **LMT Use Cases**
3. I/O System Balance
4. Occurence Histograms
5. A Simple Model
6. Conclusions

Monitoring Activity in Real Time



Detailed Performance Analysis



LMT: The Lustre Monitoring Tool

LMT Use Cases

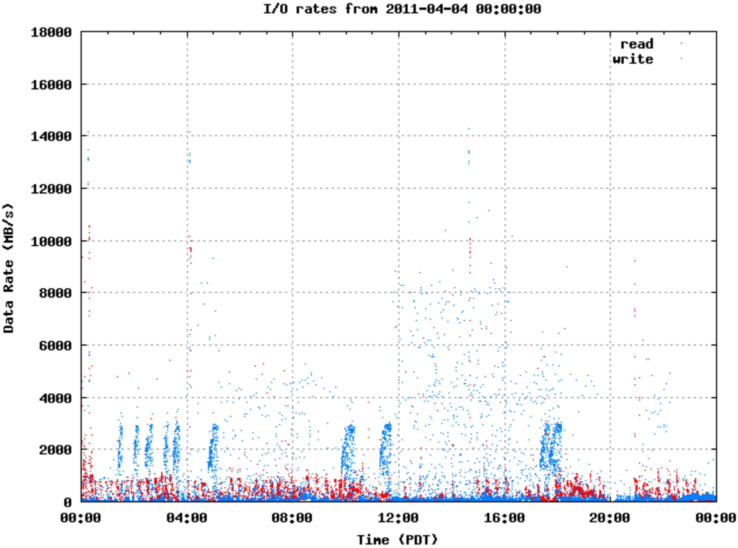
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Gestalt of a Full Day of Activity



LMT: The Lustre Monitoring Tool

LMT Use Cases

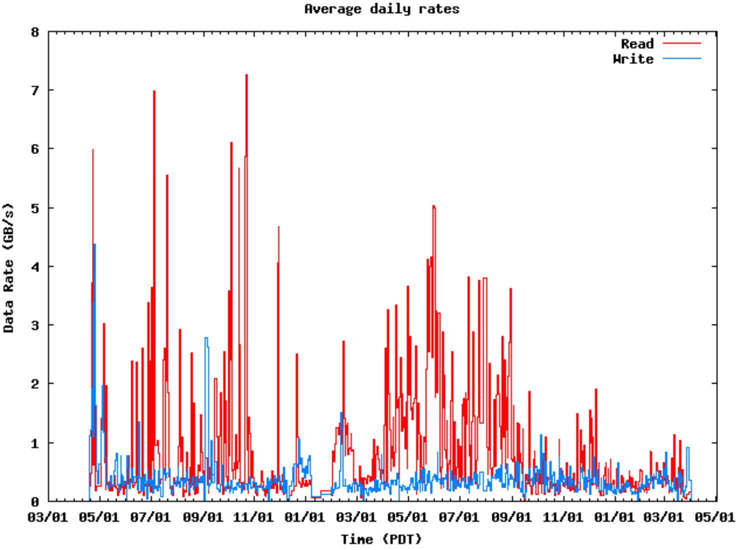
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Monitoring Long Term Trends



LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Outline

1 LMT: The Lustre Monitoring Tool

2 LMT Use Cases

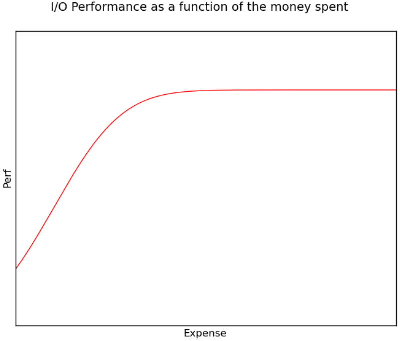
3 I/O System Balance

4 Occurence Histograms

5 A Simple Model

6 Conclusions

I/O System Balance - between cost and performance



- More money spent means (we hope) better performance

LMT: The Lustre Monitoring Tool

LMT Use Cases

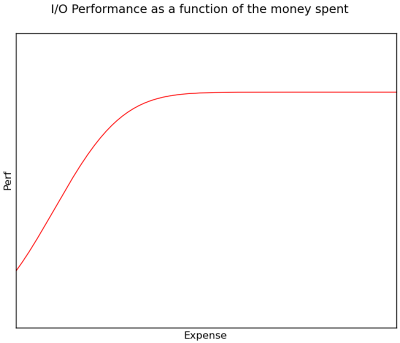
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

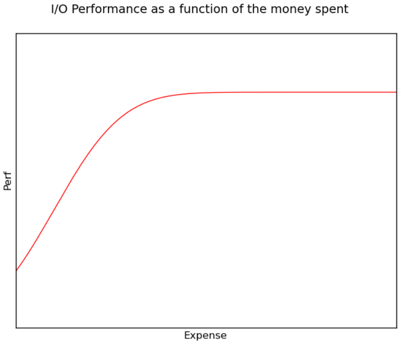
I/O System Balance - between cost and performance



- More money spent means (we hope) better performance
- Upto a point

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

I/O System Balance - between cost and performance

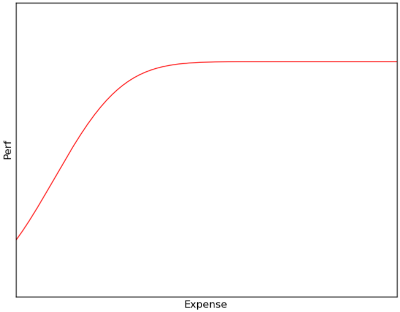


- More money spent means (we hope) better performance
- Upto a point
- How can you tell where that point is?

I/O System Balance - between cost and performance

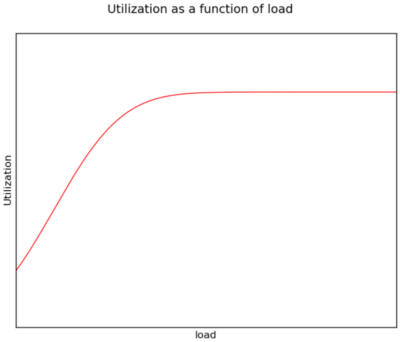
LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

I/O Performance as a function of the money spent



- More money spent means (we hope) better performance
- Upto a point
- How can you tell where that point is?
- The answer depends on both the I/O system and the workload

I/O System Balance - between I/O and compute capacity



- We want to keep the compute resource near 100% utilized

LMT: The Lustre Monitoring Tool

LMT Use Cases

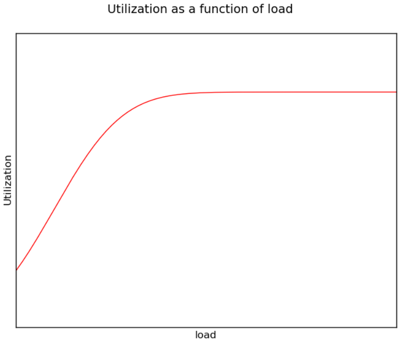
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

I/O System Balance - between I/O and compute capacity



- We want to keep the compute resource near 100% utilized
- Job schedulers are designed to make this happen

LMT: The Lustre Monitoring Tool

LMT Use Cases

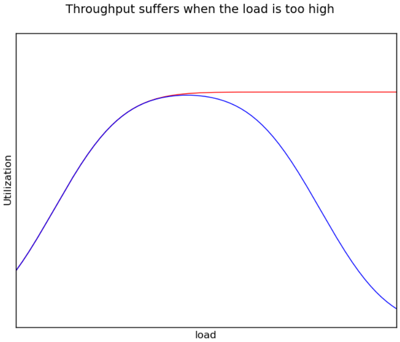
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

I/O System Balance - between I/O and compute capacity



- We want to keep the compute resource near 100% utilized
- Job schedulers are designed to make this happen
- A clogged I/O system creates a hidden penalty

LMT: The Lustre Monitoring Tool

LMT Use Cases

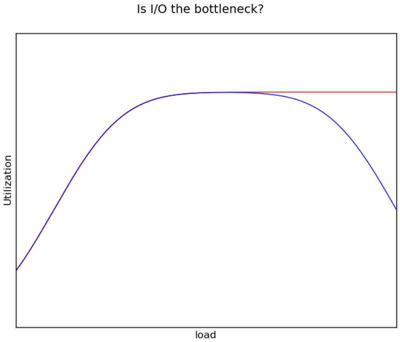
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

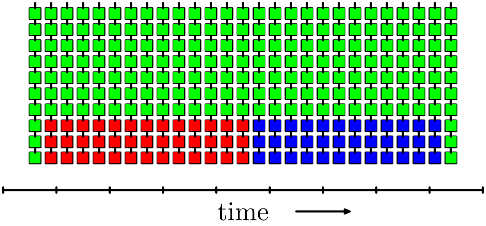
I/O System Balance - between I/O and compute capacity



- We want to keep the compute resource near 100% utilized
- Job schedulers are designed to make this happen
- A clogged I/O system creates a hidden penalty
- Can we “buy” compute resource (cheaper) by buying I/O?

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

I/O Contention



LMT: The Lustre Monitoring Tool

LMT Use Cases

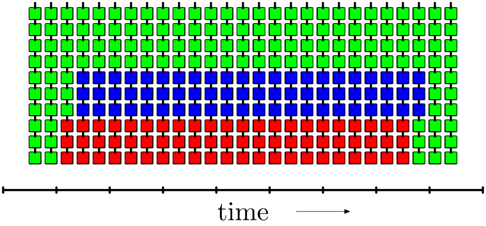
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

I/O Contention



LMT: The Lustre Monitoring Tool

LMT Use Cases

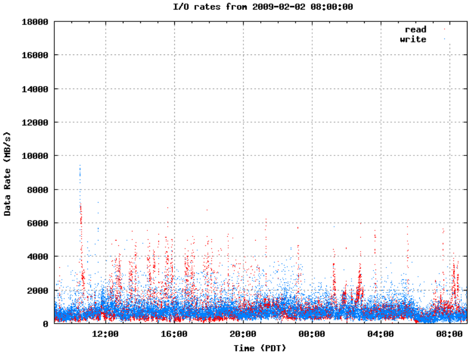
I/O System Balance

Occurence Histograms

A Simple Model

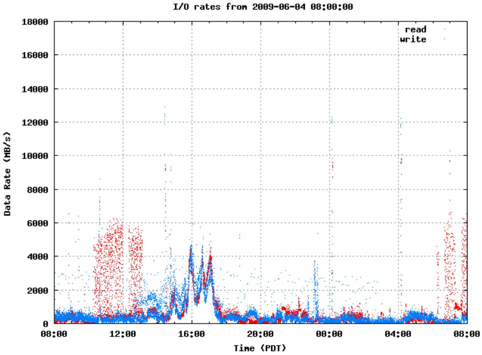
Conclusions

Case Study: April 2009 I/O Upgrade



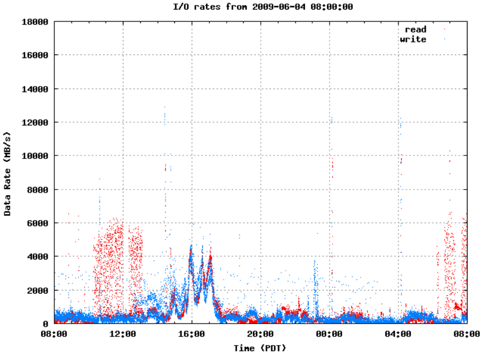
- I/O upgrade in April 2009 significantly improved performance

Case Study: April 2009 I/O Upgrade



- I/O upgrade in April 2009 significantly improved performance
- It is hard to see that fact in the before and after rate graphs

Case Study: April 2009 I/O Upgrade



- I/O upgrade in April 2009 significantly improved performance
- It is hard to see that fact in the before and after rate graphs
- Were the workloads on the two days even comparable?

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Outline

1. LMT: The Lustre Monitoring Tool
2. LMT Use Cases
3. I/O System Balance
4. **Occurence Histograms**
5. A Simple Model
6. Conclusions

LMT: The Lustre Monitoring Tool

LMT Use Cases

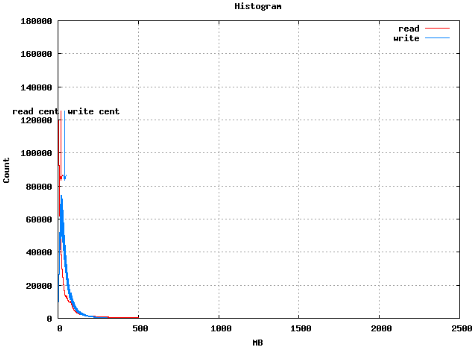
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Hisogram: Before April 2009



- A histogram shows the frequency that I/O of a particular size occurred

LMT: The Lustre Monitoring Tool

LMT Use Cases

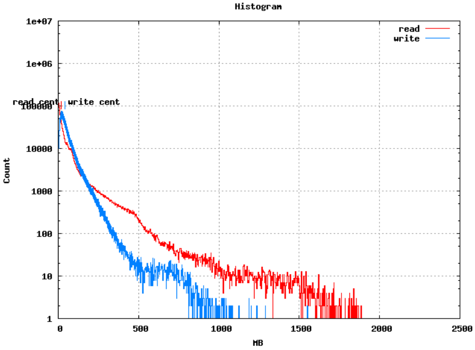
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Hisogram: Before April 2009



- A histogram shows the frequency that I/O of a particular size occurred
- A log scale makes it easier to see the shape of the distribution

LMT: The Lustre Monitoring Tool

LMT Use Cases

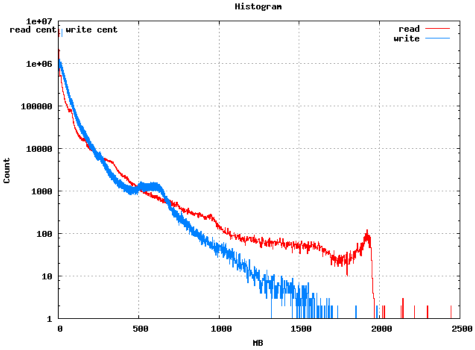
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Hisogram: Before April 2009



- A histogram shows the frequency that I/O of a particular size occurred
- A log scale makes it easier to see the shape of the distribution
- A histogram can compile data over an arbitrary time scale

LMT: The Lustre Monitoring Tool

LMT Use Cases

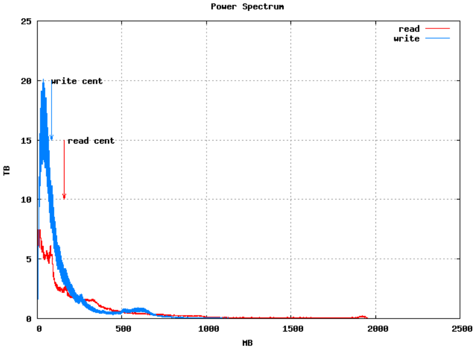
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Power Spectrum: Before and After



- A power spectrum mutiplies the histogram by the size of the observations

LMT: The Lustre Monitoring Tool

LMT Use Cases

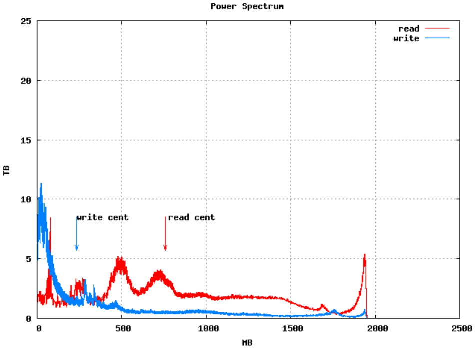
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Power Spectrum: Before and After



- A power spectrum mutiplies the histogram by the size of the observations
- before and after data (without log scale)

LMT: The Lustre Monitoring Tool

LMT Use Cases

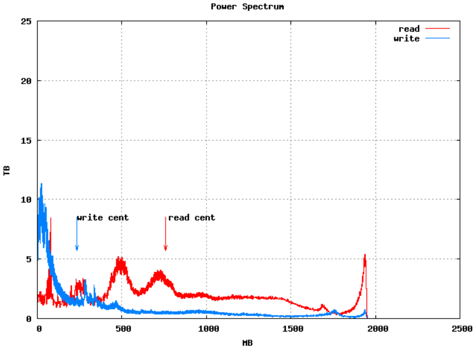
I/O System Balance

Occurence Histograms

A Simple Model

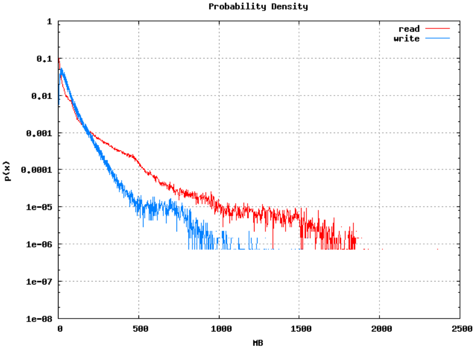
Conclusions

Power Spectrum: Before and After



- A power spectrum mutiplies the histogram by the size of the observations
- before and after data (without log scale)
- This emphasizes the significance of the larger transactions

Probability Density: Before and After Comparrison



- one day of data before a major upgrade

LMT: The Lustre Monitoring Tool

LMT Use Cases

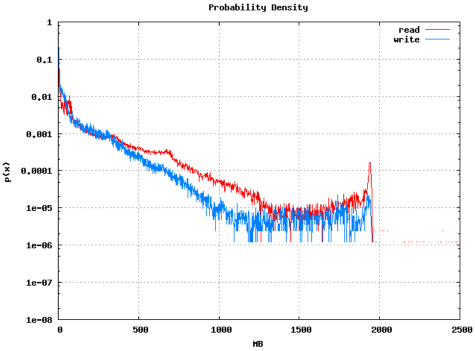
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Probability Density: Before and After Comparrison



- one day of data before a major upgrade and after
- but are the two days really comparable?

LMT: The Lustre Monitoring Tool

LMT Use Cases

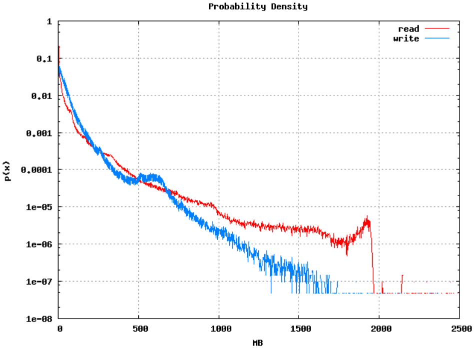
I/O System Balance

Occurence Histograms

A Simple Model

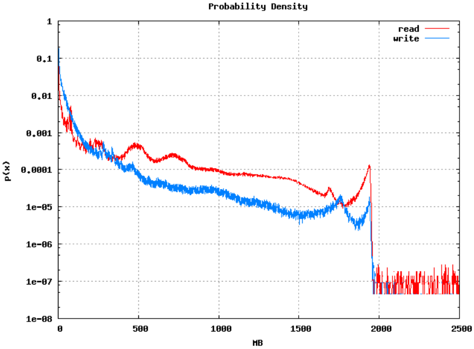
Conclusions

Probability Density: Before and After Comparrison



- one day of data before a major upgrade and after
- but are the two days really comparable?
- We can expect (hope) that the workload over long timescales is constant

Probability Density: Before and After Comparrison



- one day of data before a major upgrade and after
- but are the two days really comparable?
- We can expect (hope) that the workload over long timescales is constant

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Outline

1 LMT: The Lustre Monitoring Tool

2 LMT Use Cases

3 I/O System Balance

4 Occurence Histograms

5 A Simple Model

6 Conclusions

LMT: The Lustre Monitoring Tool

LMT Use Cases

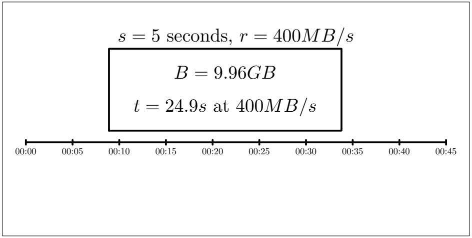
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

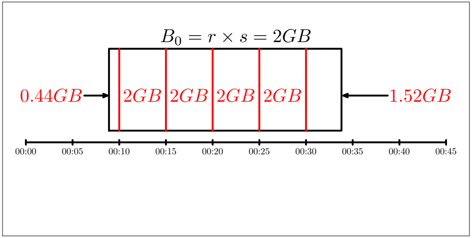
A Simple Model



- A transaction arrives at some arbitrary point

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

A Simple Model



- A transaction arrives at some arbitrary point
- It is split accross multiple obseration intervals

LMT: The Lustre Monitoring Tool

LMT Use Cases

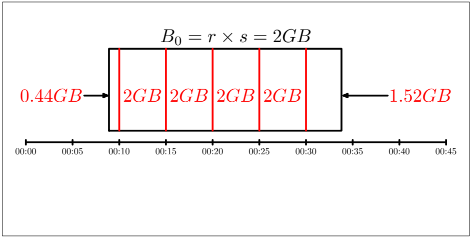
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Simple Model



- A transaction arrives at some arbitrary point
- It is split accross multiple obseration intervals
- Assumptions:
  - All the I/O in the transaction comes in together

LMT: The Lustre Monitoring Tool

LMT Use Cases

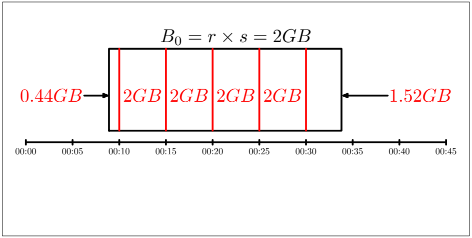
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Simple Model



- A transaction arrives at some arbitrary point
- It is split accross multiple obseration intervals
- Assumptions:
  - All the I/O in the transaction comes in together
  - The I/O proceeds at its maximum rate until complete

LMT: The Lustre Monitoring Tool

LMT Use Cases

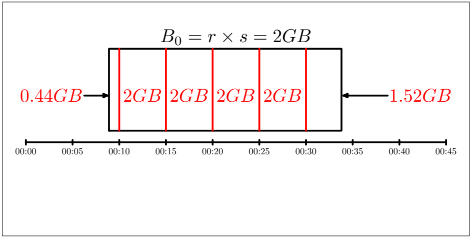
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Simple Model



- A transaction arrives at some arbitrary point
- It is split accross multiple obseration intervals
- Assumptions:
  - All the I/O in the transaction comes in together
  - The I/O proceeds at its maximum rate until complete
  - One transaction at a time

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

LMT: The Lustre Monitoring Tool

LMT Use Cases

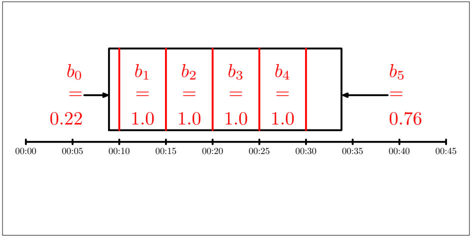
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Big Transactions $T > 1$, $(B > B_0)$



- It simplifies things to express the transaction size as a multiple of the maximum observation size: $T = B/B_0$
- $b_0 + b_5 < 1$ and $n = \lfloor T \rfloor + 2$

LMT: The Lustre Monitoring Tool

LMT Use Cases

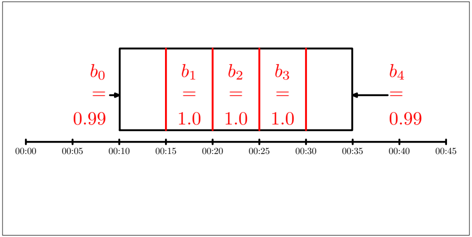
I/O System Balance

Occurence Histograms

A Simple Model

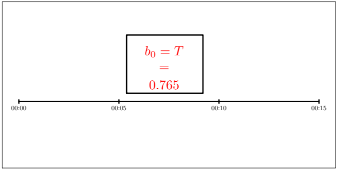
Conclusions

Big Transactions $T > 1$, $(B > B_0)$



- It simplifies things to express the transaction size as a multiple of the maximum observation size: $T = B/B_0$
- $b_0 + b_5 < 1$ and $n = \lfloor T \rfloor + 2$
- $b_0 + b_4 > 1$ and $n = \lfloor T \rfloor + 1$

Small Transactions $T < 1$, ($B < B_0$)



- Fits within one observation interval

LMT: The Lustre Monitoring Tool

LMT Use Cases

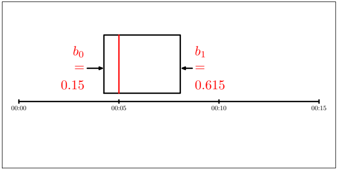
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Small Transactions $T < 1$, ($B < B_0$)



- Fits within one observation interval
- Split accross two

LMT: The Lustre Monitoring Tool

LMT Use Cases

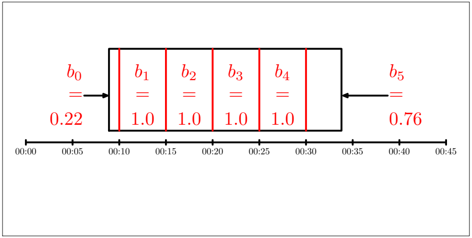
I/O System Balance

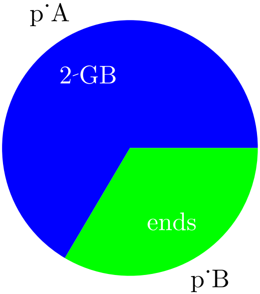
Occurence Histograms

A Simple Model

Conclusions

Distribution of Observations for Large Transactions, $T > 1$





- $p_A = \frac{T-1}{T+1}$ chance that an observation is at $x = 1$

LMT: The Lustre Monitoring Tool

LMT Use Cases

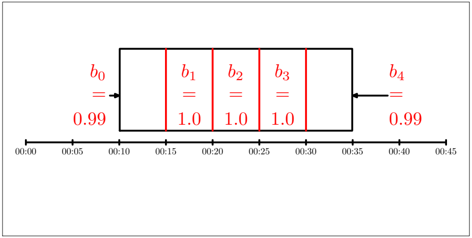
I/O System Balance

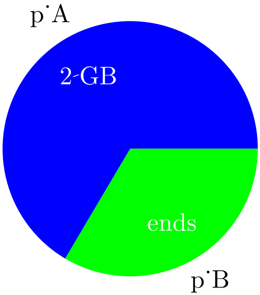
Occurence Histograms

A Simple Model

Conclusions

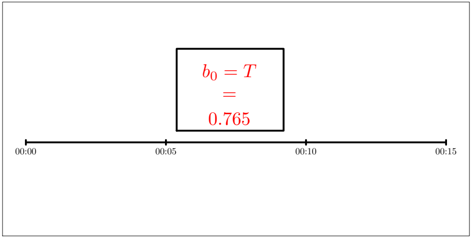
Distribution of Observations for Large Transactions, $T > 1$

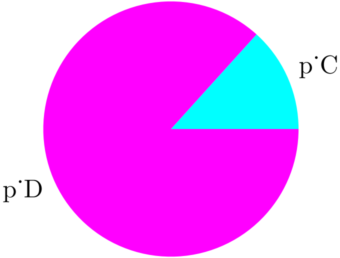




- $p_A = \frac{T-1}{T+1}$ chance that an observation is at $x = 1$
- $p_B = \frac{2}{T+1}$ chance that an observation is an *end*

Distribution of Observations for Small Transactions, $T < 1$

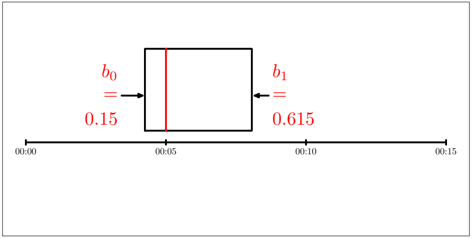


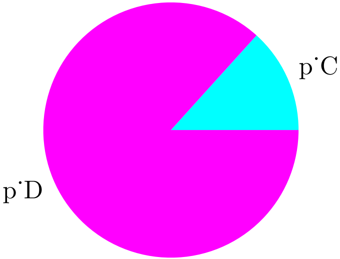


- $p_C = \frac{1-T}{1+T}$ chance that an observation is at $x = T$

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

Distribution of Observations for Small Transactions, $T < 1$





- $p_C = \frac{1-T}{1+T}$ chance that an observation is at $x = T$
- $p_D = \frac{2T}{1+T}$ chance that it is as piece of a *split* transaction

LMT: The Lustre Monitoring Tool

LMT Use Cases

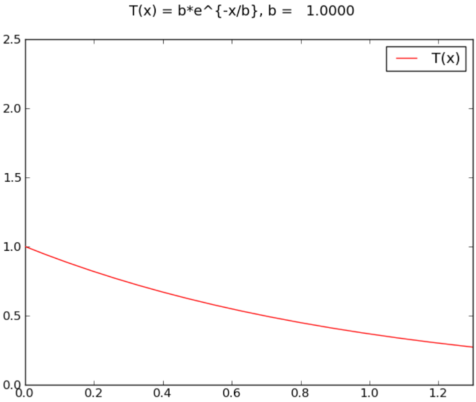
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x > 1$)



- Suppose $T(x) = \beta \exp(-x/\beta)$

LMT: The Lustre Monitoring Tool

LMT Use Cases

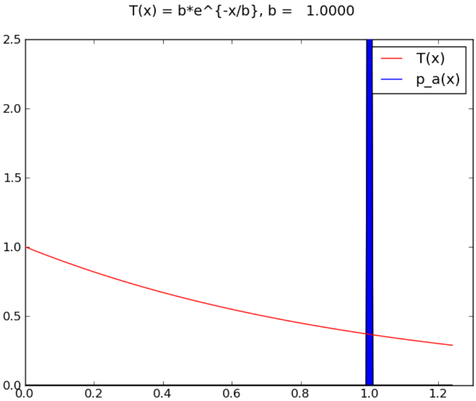
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x > 1$)



- Suppose $T(x) = \beta \exp(-x/\beta)$
- $p_a(x)$ is a delta function with $p_A = \int_1^\infty \frac{x-1}{x+1} T(x) dx$

LMT: The Lustre Monitoring Tool

LMT Use Cases

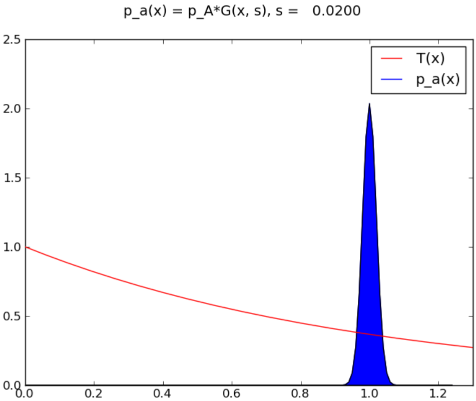
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x > 1$)



- Suppose $T(x) = \beta \exp(-x/\beta)$
- $p_a(x)$ is a delta function with $p_A = \int_1^{\infty} \frac{x-1}{x+1} T(x) dx$
- But suppose there is a variability $p_a(x) = p_A G(x, \sigma)$ (Gaussian)

LMT: The Lustre Monitoring Tool

LMT Use Cases

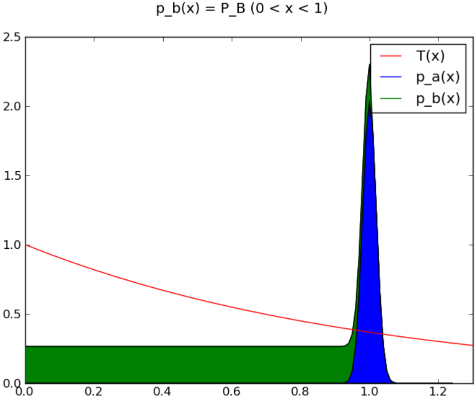
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x > 1$)



- Suppose $T(x) = \beta \exp(-x/\beta)$
- $p_a(x)$ is a delta function with $p_A = \int_1^\infty \frac{x-1}{x+1} T(x)dx$
- But suppose there is a variability $p_a(x) = p_A G(x, \sigma)$ (Gaussian)
- $p_b(x) = \int_1^\infty \frac{2}{x+1} T(x)dx$

LMT: The Lustre Monitoring Tool

LMT Use Cases

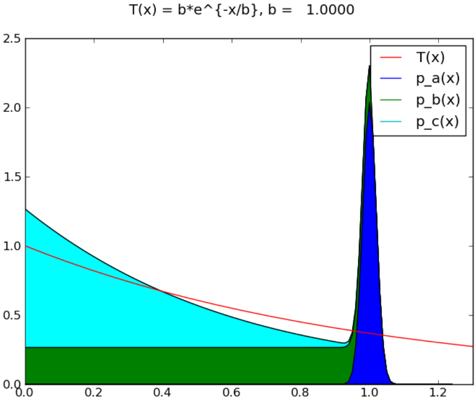
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x < 1$)



- $p_c(x) = (1 - x)T(x)$

LMT: The Lustre Monitoring Tool

LMT Use Cases

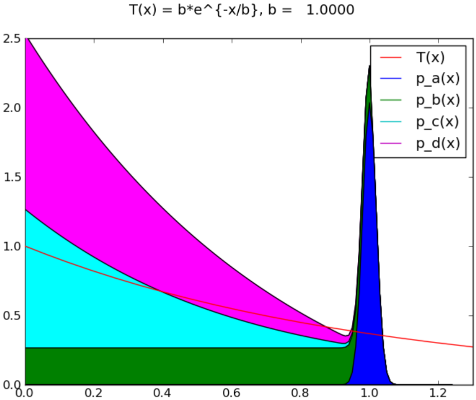
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x < 1$)



- $p_c(x) = (1 - x)T(x)$
- $p_d(x) = 2 * \int_x^1 T(x')dx'$

LMT: The Lustre Monitoring Tool

LMT Use Cases

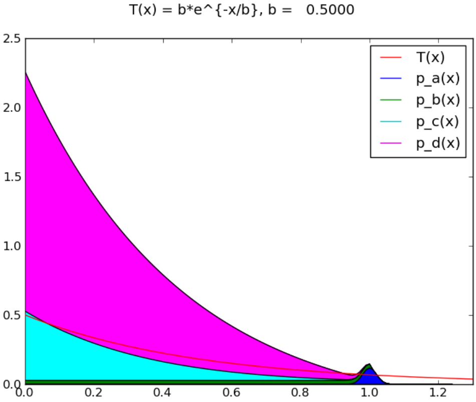
I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

A Distribution $T(x)$ Over Transaction Sizes ($x < 1$)



- $p_c(x) = (1 - x)T(x)$
- $p_d(x) = 2 * \int_x^1 T(x')dx'$

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Outline

1 LMT: The Lustre Monitoring Tool

2 LMT Use Cases

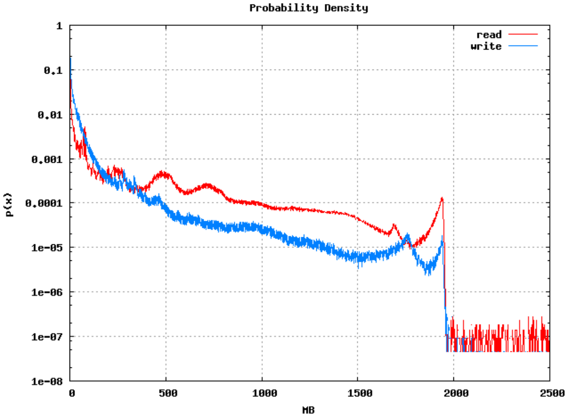
3 I/O System Balance

4 Occurence Histograms

5 A Simple Model

6 Conclusions

Challenges



- curve shape
- curve fit estimates
- variability
- modes

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

Conclusions

- LMT is a source of data on file system performance
- I/O contention can result from a system imbalance
- System balance depends on the workload
- A statistical view can illuminate the workload pattern
- A very simple model helps relate the workload to the observations
- Using the observations to infer the workload is hard

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Conclusions

- LMT is a source of data on file system performance
- I/O contention can result from a system imbalance
- System balance depends on the workload
- A statistical view can illuminate the workload pattern
- A very simple model helps relate the workload to the observations
- Using the observations to infer the workload is hard
- But not impossible

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Questions?

- `https://computing.llnl.gov/linux/cerebro.html`
  Al Chu
- `http://code.google.com/p/lmt/`
  Herb Wartens, Jim Garlick

LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions

Additional Slides

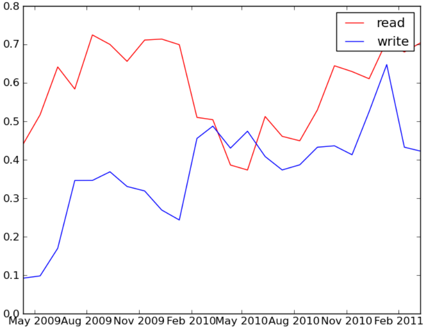
(counts in millions)

	read	*write*	*both*	*read* × *rwrite*
count	450			
zero	205	124	70	
bin_0	267	178	141	
$zero/count$	0.46	0.28	0.15	0.07
$bin_0/count$	0.60	0.40	0.31	0.23

A large fraction of all observations are 0.0 and even more are in the first bin close to 0.0. If the *read* and *write* I/O streams were truely independent the occurence of both *read* and *write* observations simultaneously would be about the product of their separate probabilities.

LMT: The Lustre Monitoring Tool
LMT Use Cases
I/O System Balance
Occurence Histograms
A Simple Model
Conclusions

Zeros, Month by Month



LMT: The Lustre Monitoring Tool

LMT Use Cases

I/O System Balance

Occurence Histograms

A Simple Model

Conclusions