LNet and LND Tuning Explained

Chris Horn, Cray Inc.
hornc@cray.com

Agenda

- Why I'm Giving This Talk
- LNet's Credit System
- LNet Credits and 4MB RPCs
- o2iblnd parameters
- Quick Testing Tutorial

Google optimal lnet lnd tuning configuration

Did you mean: optimal lnet ***lund*** tuning configuration

Tuning NGINX for Performance - NGINX

https://www.nginx.com/blog/**tuning**-nginx/ ▾ NGINX, Inc. ▾

Oct 10, 2014 - You can **tune** almost any **setting**, but this post concentrates on the few **settings** for ... **net**.core.somaxconn – The maximum number of connections that can be .... Prices · nginx.conf 2015 Schedule, **Best** Prices, Sneak Peek, and More! ... United States, Afghanistan, Ã…**land** Islands, Albania, Algeria, Andorra ...

CleanFlight Setup Tuning Guide for Naze32 / CC3D ...

blog.oscarliang.**net**/cleanflight-naze32-**setup**/ ▾

Jan 16, 2015 - This tutorial will show you how to **setup** Cleanflight firmware flash on Naze32 Flight Controller. ... OscarLiang.**net** .... carry out a series of **tuning** cycles, perform all sorts of movements to determine the "**best**" PID **settings**. ... Before that, to adjust PID values, we usually have to **land**, disarm, and connect your ...

MinimOSD Micro Setup Tutorial - Naze32 PID Tuning via ...

blog.oscarliang.**net**/minimosd-micro-**setup**-naze32-pid-rssi/ ▾

Apr 30, 2015 - OscarLiang.**net** ... Although It's more complicated to **setup** than those standalone .... It also allows you to **tune** PID with OSD menu. .... I found it's **best** to get RSSI from a spare PPM channel, that way you ..... So when I takeoff, battery voltage reads 12.6 and stays at 12.6, until I **land** and disarm, then it changes ...

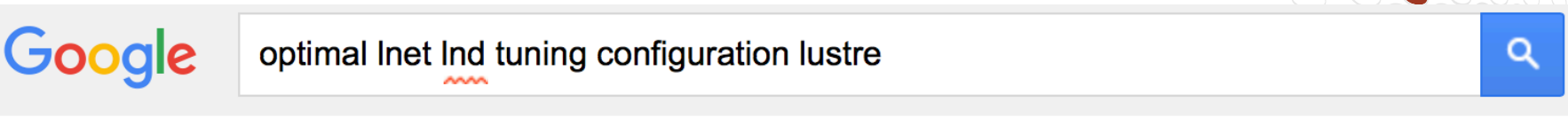
[PDF] An optimally tuned ensemble of the "eb_go_gs ...

www.geosci-model-dev.**net**/6/1729/2013/gmd-6-1729-2013.pdf ▾

by R Marsh - 2013 - Related articles

Oct 21, 2013 - www.geosci-model-dev.**net**/6/1729/2013/ doi:10.5194/gmd-6-1729-2013

An **optimally tuned** ensemble of the "eb_go_gs" **configuration** of ... off between



Lustre Tuning - Obsolete Lustre Wiki

wiki.old.**lustre**.org/manual/**Lustre**Manual18_HTML/**LustreTuning**.html

This chapter contains information to **tune Lustre** for better performance and ... is a process of trial and error, and varies for each particular **configuration**. ... At this time, no testing has been done to determine the **optimal** number of MDS ... This section describes **LNET** tunables. .... We are making changes to the **ptllnd** module.

LustreProc - Obsolete Lustre Wiki

wiki.old.**lustre**.org/manual/**Lustre**Manual20_HTML/**Lustre**Proc.html

LND timeouts that ensure point-to-point communications complete in finite time in the ... If **Lustre** timeouts are not accompanied by **LNET** timeouts, then you need to ... One of the goals of adaptive timeouts is to relieve users from having to **tune** the .... The **Lustre** engine always attempts to pack an **optimal** amount of data into ...

Book Index - Obsolete Lustre Wiki

wiki.old.**lustre**.org/manual/**Lustre**Manual20_HTML/ix.html

LNET self-test. commands, 1 ... administration, regenerating **Lustre configuration** logs, 1. administration ... reliability **best** practices, 1. selecting storage for ... **SOCKLND** kernel TCP/IP **LND**, 1. starting. **LNET**, 1. statahead, **tuning**, 1. stopping .

[PDF] Tips and Tricks for Diagnosing Lustre Problems on Cray ...

https://cug.org/5-publications/proceedings.../12A-Spitz-Paper.pdf

by C Spitz - Cited by 1 - Related articles

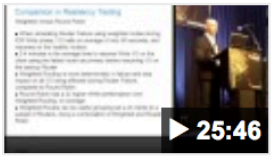
LNET, and **LND** messages are recorded in the syslog messages file on the SDB ... Since the tool does not require the **configuration** from the. **Lustre** MGS or the ...

Google

lustre lnet tuning configuration

[PDF] LNET Router Resiliency and Tuning - OpenSFS
cdn.opensfs.org/wp.../04/**Lustre-Net**work-Router-**Config**_Fragalla.pdf
Apr 15, 2015 - **LNET** Router **Configuration**/Testing. □ Fine **Tuning** and testing 4MB I/O ... –1 IB **Lustre** File System, IB Clients for performance, Ethernet.

Video: Lustre Network (LNET) Router Configuration and ...
insidehpc.com/.../video-**lustre-net**work-**lnet**-router-**configuration**-and-**tun**...
Apr 22, 2015 - In this video from LUG 2015 in Denver, John Fragalla from Seagate presents: **Lustre** Network (**LNET**) Router **Configuration** and **Tuning**.

Lustre Network (LNET) Router Configuration and Tuning ...


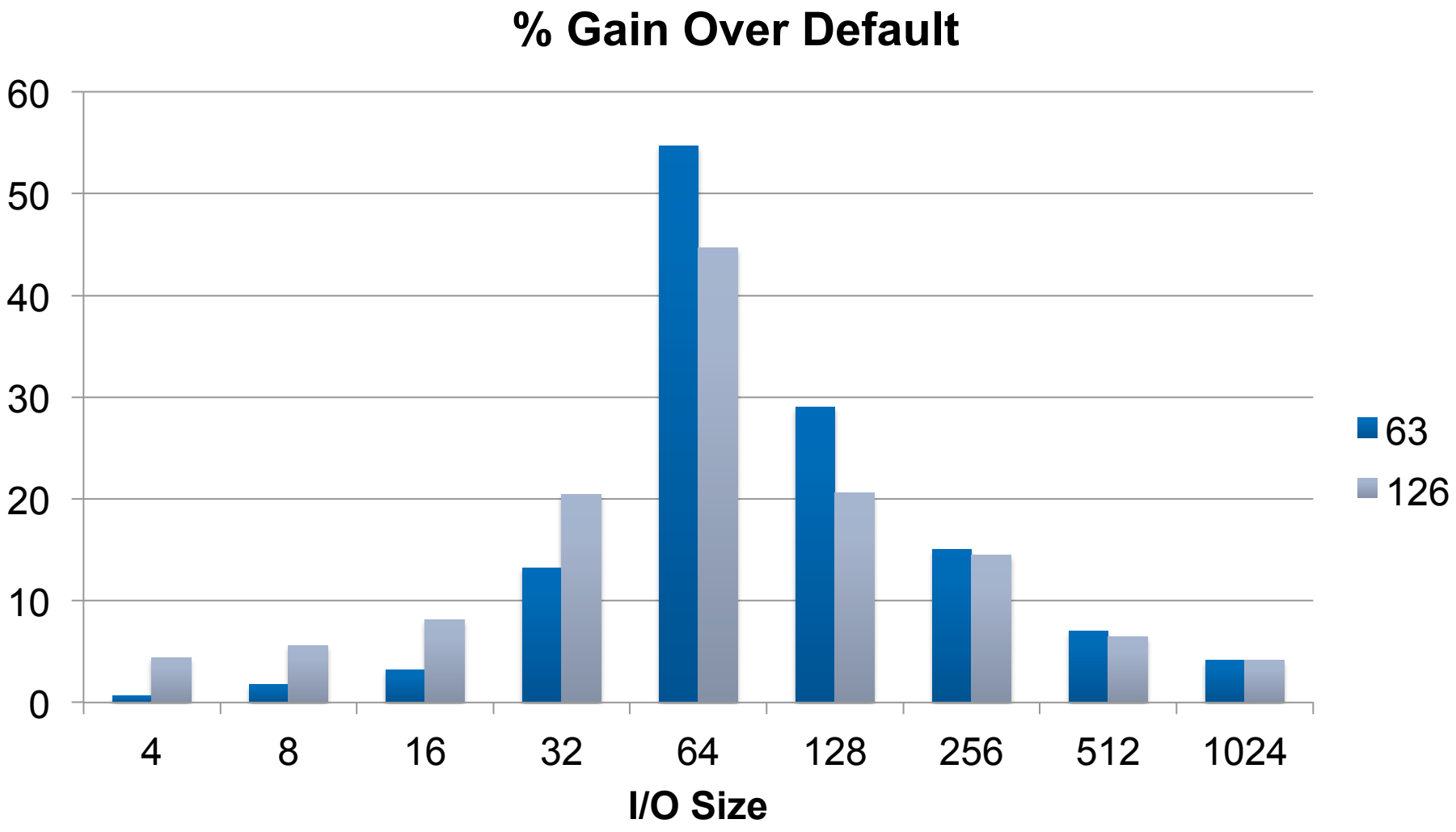
www.youtube.com/watch?v=bm_uPtNsd1Y
Apr 15, 2015 - Uploaded by RichReport
In this video from LUG 2015 in Denver, John Fragalla from Seagate presents: **Lustre** Network (**LNET**) Router ...

Lustre Tuning - Obsolete Lustre Wiki
wiki.old.**lustre**.org/manual/**Lustre**Manual20_HTML/**LustreTuning**.html
This chapter contains information about **tuning Lustre** for better performance and includes the following sections: Optimizing the ... **Tuning LNET** Parameters.

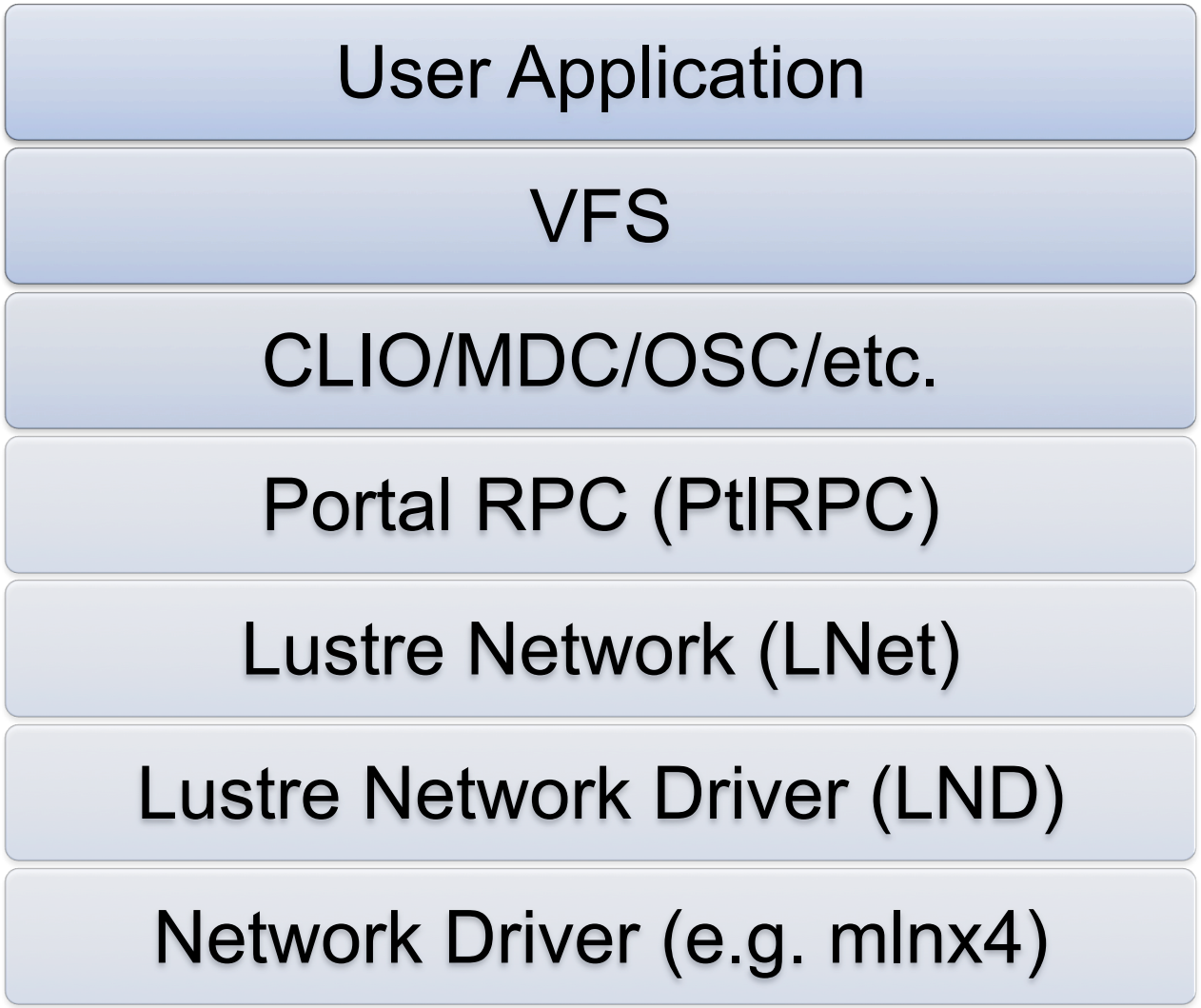
[PDF] LNET Configuration - ORNL Lustre Activities
lustre.ornl.gov/ecosystem/documents/**Lustre**Eco2015-Tutorial2.pdf
Mar 3, 2015 - 3 types of nodes to consider. – **Lustre** Client. – **Lustre** Server. – **LNET** Router. • All **tuned** differently. – Some commonalities. • Let's take a look.

[PDF] Oak Ridge National Laboratory Lustre Tuning and ...
lustre.ornl.gov/**lustre**101-courses/content/C1/L5/**LustreTuning**.pdf
Tuning recommendations from OLCF experience. • Multi-rail **LNET configurations** . •

Why Tuning Matters



(Some of) The Software Stack



LNet's Credit System - Sends

- **Every lnet_send() takes a peer credit and network interface credit**
  - Except for the loopback NI: 0@lo
- **Peer Credit**
- **Network Interface Credit**

Peer Credits

- Governs the number of concurrent sends to a single peer.
- Set with an LND's peer_credits module parameter
  - e.g "ko2iblnd peer_credits"
- Point-to-point
  - End-to-end flow control accomplished at higher layer. e.g. max_rpcs_in_flight

```
# cat /proc/sys/lnet/peers
nid                      refs state  last   max   rtr   min    tx   min queue
10.149.2.72@o2ib            3    up    -1   126   126   126   126   110 0
```

- "tx" is the number of peer credits currently available for this peer
- "min" is the smallest number of peer credits seen
- Negative credit count indicates the number of messages awaiting a credit

Network Interface Credits

- Governs the number of concurrent sends to a single network
- Set with an LND's credits module parameter
  - e.g. "ko2iblnd credits"
  - Shared across all CPU partitions (CPTs)

```
# cat /proc/sys/lnet/nis
nid                     status alive refs peer  rtr   max    tx   min
10.149.4.5@o2ib             up    -1    9  126    0  2048  2048  1796
```

- "max" is total available (i.e. value of ko2iblnd credits)
- "tx" is the number currently available
  - Negative number indicates number of messages awaiting a credit
- "min" is the low water mark

LNet's Credit System – (Routed) Receives

- Additional credit accounting when routers receive a message destined for another peer
- These credits account for resources taken on the router node
- Peer Router Credit
- Router Buffer Credit

Peer Router Credit

- **Governs the number of concurrent receives from a single peer**
- **Prevent single peer from using all router buffer resources**
- **Set with module parameter "ko2iblnd peer_buffer_credits"**
  - (Or "ksocklnd peer_buffer_credits"; Default is 0 for both
  - At network initialization, if zero, or LND does not provide value, then uses the "lnet peer_buffer_credits" module parameter: Default is 0
  - If LND and LNet value is zero, then LND's peer_credits value is used
- **Router takes a credit for the peer it's receiving *from* (lnet_post_routed_recv_locked())**
- **A credit is given back when the receive completes (lnet_return_rx_credits_locked())**

Router Buffer Credit

- Router has limited number of three different sized buffers: tiny, small, large
- Router buffer credits ensure we only receive if an appropriate buffer is available
- Tiny buffers for 0 byte payloads
- Small buffers for Single page payloads
- Large buffers for payload > single page
- Number of buffers of each type defined with LNet module parameters:
  - options lnet tiny_router_buffers (At least 512 per CPT)
  - options lnet small_router_buffers (At least 4096 per CPT)
  - options lnet large_router_buffers (At least 256 per CPT)

LNet Credits and 4MB RPCs

- 4MB I/Os associate additional LNet memory descriptors with a bulk operation. LNet MTU is still 1MB
- The total number of messages (and thus credits) required to complete bulk reads and writes is lower for transfers > 1MB
- *lctl set_param osc.*.max_pages_per_rpc=1024*

	256 pages/RPC	1024 pages/RPC
1 MB Write	1 RPC, 2 Credits	1 RPC, 2 Credits
2 MB Write	2 RPCs, 4 Credits	1 RPC, 3 Credits
3 MB Write	3 RPCs, 6 Credits	1 RPC, 4 Credits
4 MB Write	4 RPCs, 8 Credits	1 RPC, 5 Credits

# of bulk write RPCs sent and peer credits taken for bulk transfer by a single client for different sized writes (read case is the same)

o2iblnd parameters

- peer_credits, peer_credits_hiw, credits
- concurrent_sends and map_on_demand
  - Control number of Work Requests per Queue Pair
  - # WRs = (map_on_demand + 1) * concurrent_sends

```
# ibv_devinfo -v | grep max_qp_wr
        max_qp_wr:                      16351
```

```
LNetError: 16485:0:(o2iblnd.c:869:kiblnd_create_conn()) Can't create QP:
-22, send_wr: 16448, recv_wr: 256
```

- map_on_demand
  - Disabled by default (value of 256 in above equation)
  - Reduce number of work requests per queue pair at the cost of using FMR for transfers >= map_on_demand
  - Need LU-3322 to mix map_on_demand on/off on different peers

o2iblnd parameters - cont.

- **Fast Memory Region tuning:**
  - Memory Region registration is “heavy”
  - FMR pools are lightweight by comparison
  - Grow at runtime, so just need sane starting point
  - fmr_pool_size
  - fmr_flush_trigger
  - fmr_cache
  - Note: FMR eventually going away
    - http://article.gmane.org/gmane.linux.drivers.rdma/29040

LNet Testing How-To

- **Establish a baseline:**
  - Reference materials
  - ib_read_bw, ib_write_bw, etc.
  - lnet_selftest
    - See Lustre Ops Manual Chapter 24.
- **Set goals**
  - Peer-to-peer performance
  - Bulk performance, I/O size
  - Memory usage
  - Message rate
- **Iterate, iterate, iterate**
  - Make an educated guess
  - Measure with lnet_selftest
  - repeat

Helpful Links and References

- **Understanding Lustre Filesystem Internals**
  - http://users.nccs.gov/~fwang2/papers/lustre_report.pdf
- **LNet Configuration**
  - http://lustre.ornl.gov/ecosystem/documents/LustreEco2015-Tutorial2.pdf
- **Lustre Tuning and Advanced LNet Configuration**
  - http://lustre.ornl.gov/lustre101-courses/content/C1/L5/LustreTuning.pdf
- **InfiniBandTM Architecture Specification Volume 1**
  - http://www.infinibandta.org
- **Lustre Resiliency: Understanding Lustre Message Loss and Tuning for Resiliency**
  - http://wiki.lustre.org/Lustre_Resiliency:_Understanding_Lustre_Message_Loss_and_Tuning_for_Resiliency
  - http://goo.gl/upcN3l
- **Linux Kernel Networking: Implementation and Theory**
- **Minimizing Lustre Ping Effects at Scale on Cray Systems**
  - https://cug.org/proceedings/attendee_program_cug2012/includes/files/pap166.pdf

Questions?

Chris Horn
hornc@cray.com