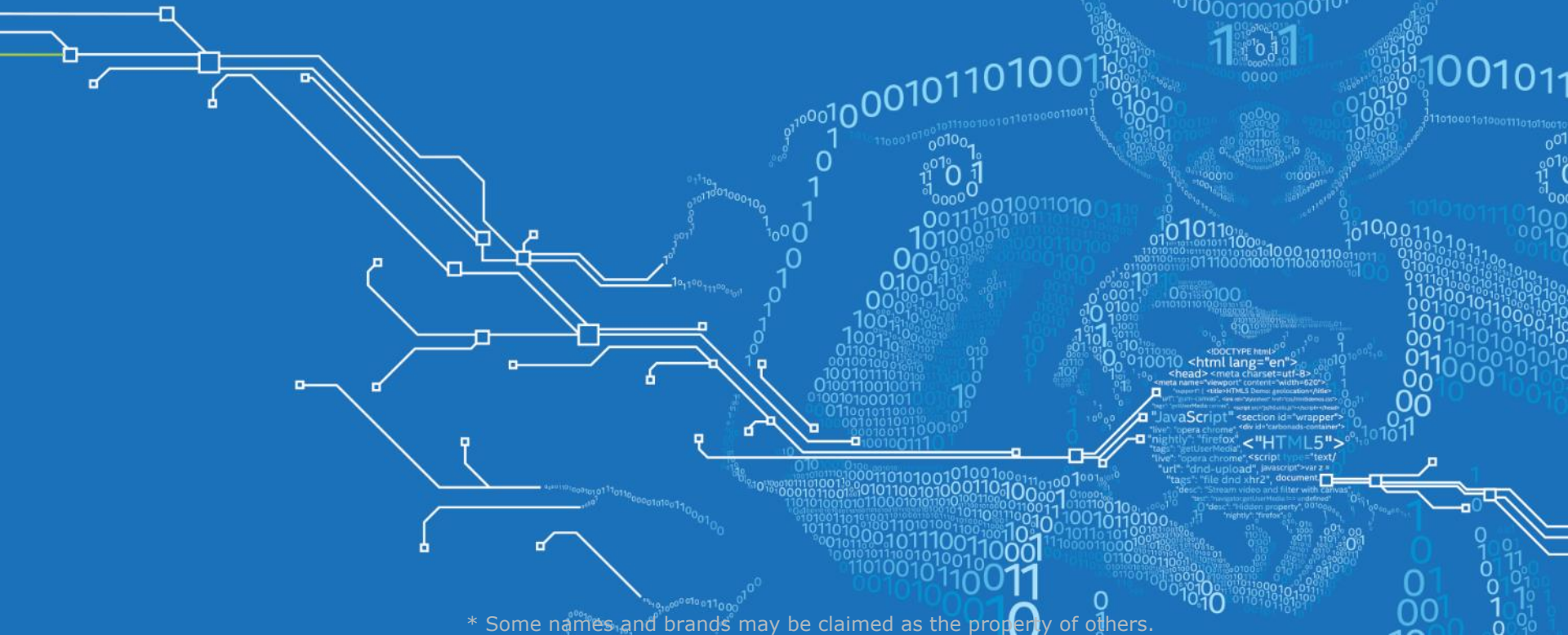
Optimization of Lustre* performance using a mix of fabric cards

LAD 2015

Dmitry Eremin



Agenda

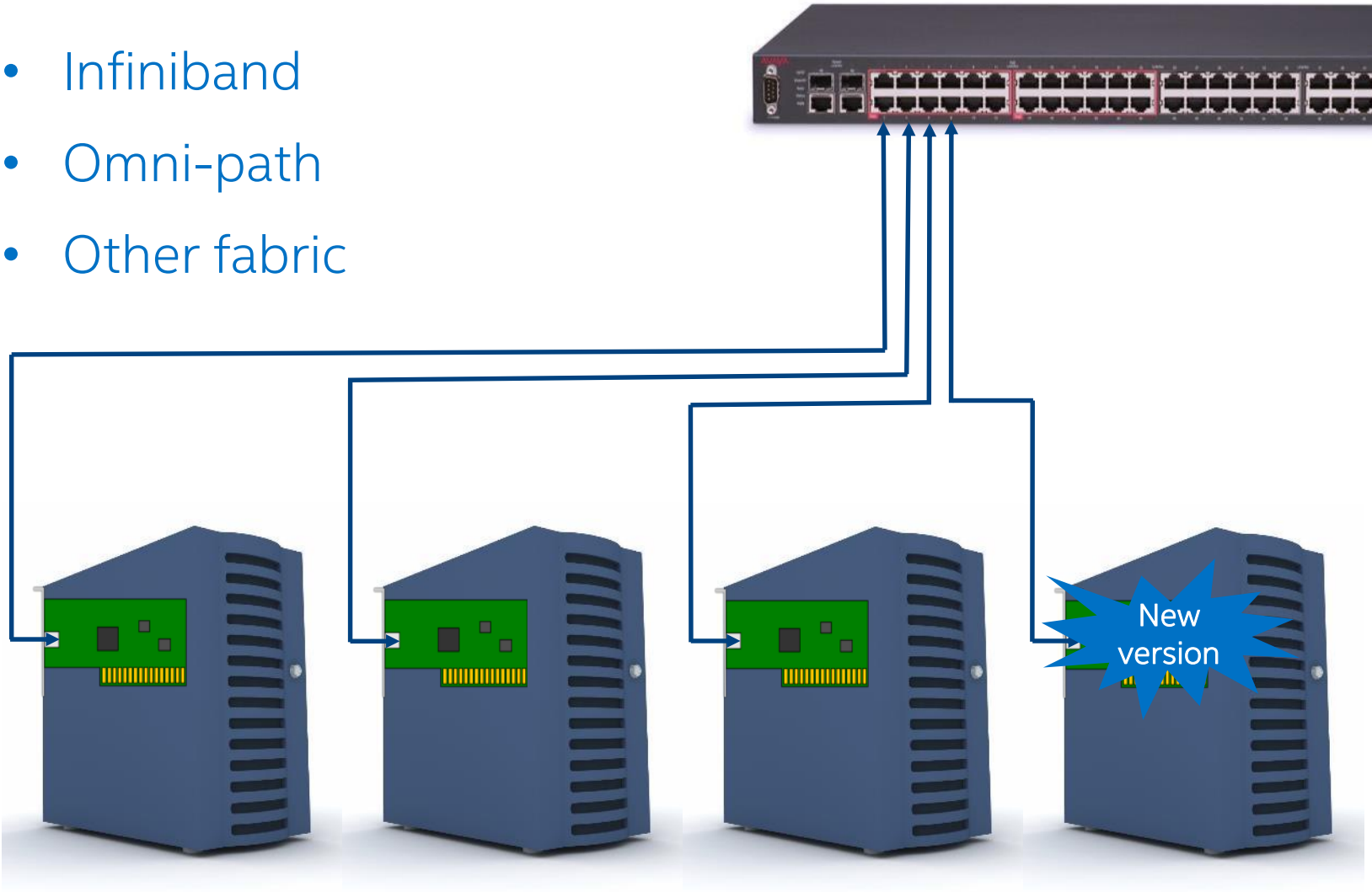
- High variety of RDMA solutions
- Network optimization
- LND tunables and new ko2iblnd.conf
- Co-existence of different cards
- Conclusion

High variety of RDMA solutions

- Infiniband
- Omni-Path
- RoCE
- iWARP

Typical Configuration

- Infiniband
- Omni-path
- Other fabric



Network optimization

Intel® True Scale Infiniband:

- With default LNET configuration the performance is much worse than with LNET tuned

Intel® Omni-Path:

- Maximum performance can be obtained with LNET tuning

Mellanox* Infiniband:

- Don't need to tune LNET for old cards but for new it can be tuned

LND tunables

- ko2iblnd.ko has 24 tunable parameters
- Parameters are read only once during the loading module into the kernel
  - at boot or when Lustre is first mounted
- All sets of parameters in single ko2iblnd.conf

Example of /etc/modprobe.d/lustre.conf

options lnet networks="o2ib0(ib0)"

options ko2iblnd peer_credits=128 peer_credits_hiw=64 ...

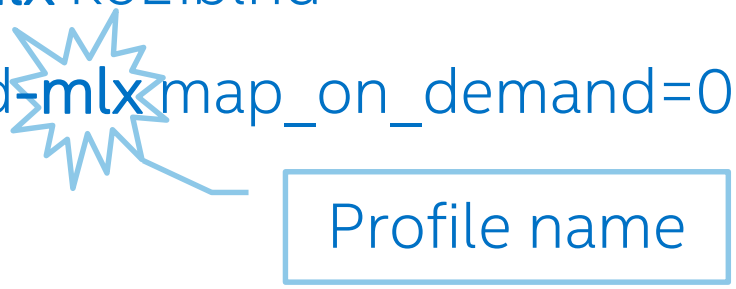
New /etc/modprobe.d/ko2iblnd.conf

alias ko2iblnd-**opa** ko2iblnd

options ko2iblnd-**opa** map_on_demand=32 ...

alias ko2iblnd-**mlx** ko2iblnd

options ko2iblnd-**mlx** map_on_demand=0 ...



alias ko2iblnd-**xxx** ...

options ko2iblnd-**xxx** ...

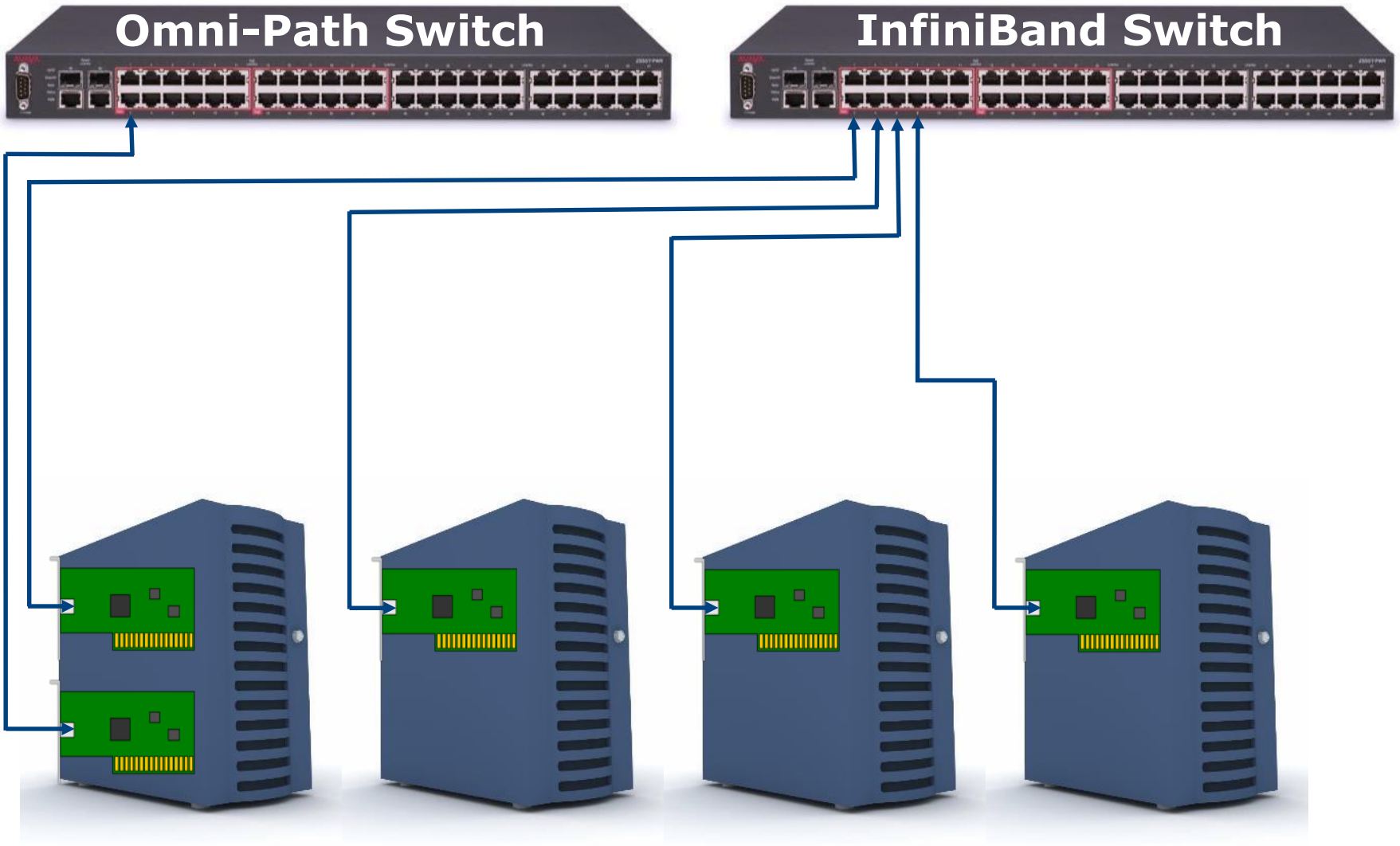
Card detection and tunables selection

- Handled via **/usr/sbin/ko2iblnd-probe** script at runtime
- Any user-space tools can be used for card detection
  - for example: “lspci -v | grep Mellanox”
- Tunables can be selected individually for each
  - type of card
  - version of card
  - specific user defined feature

Example of /usr/sbin/ko2iblnd-probe

```
INFINIBAND="/sys/class/infiniband"
PROFILE=""
if [ -d $INFINIBAND ]; then
    for dev in `ls -d $INFINIBAND/* | sed -e "s#^$INFINIBAND/##" -e 's#_[0-9]*$##'`; do
        case $dev in [...] esac
    done
    # Set profile name according priority
    if [ ... ]; then
        PROFILE="-xxx"
    fi
fi
exec /sbin/modprobe --ignore-install ko2iblnd$PROFILE $CMDLINE_OPTS
```

Mixed Configuration



Co-existence of different cards

- Single OFED distribution should support all cards
- OFED API hides the details about cards
- LND tunables are common for all network interfaces now
  - The fixes are coming (LU-6850, LU-3322 and LU-7101)
- The performance of routers depend on the slowest card

Conclusion

- Currently it isn't possible to auto-tune LND optimally inside the kernel since the OFED API hides the details
- Unfortunately, there isn't a single set of parameters that provide optimal performance for different cards
- Card detection and tunable selection is handled via /usr/sbin/ko2iblnd-probe at runtime when the ko2iblnd module is loaded, either at boot or when Lustre is first mounted

Legal Information

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visithttp://www.intel.com/performance.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.
- The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
- Intel and the Intel logo, are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others


intel
Software