

# High-Performance Linpack Optimization

Gurusiddhesha Chandrasekhara  
Department of Informatics  
Technische Universität München  
guru.rvcs@gmail.com

Madhura Kumaraswamy  
Department of Informatics  
Technische Universität München  
madhura.kswamy@gmail.com

Madhukar Sollepura Prabhu Shankar  
Department of Informatics  
Technische Universität München  
madhutime@gmail.com

Lukas Grillmayer  
Department of Informatics  
Technische Universität München  
lukas.grillmayer@in.tum.de

February 22, 2015

## Abstract

The purpose of this project is to automate the process of tuning parameters for the Linpack benchmark to obtain the optimal configuration. A Persicope plugin was created to perform this task. The plugin was then tested on consumer grade machines and built on the SuperMUC supercomputer. Results show the best scenario for manipulation of single parameters.

## 1 Introduction

Supercomputers embody the climax of high-performance computing (HPC) systems. They provide computation power for governments, industries, and researchers. Applications include the general areas of cryptanalysis, simulation, and modeling. With advancing technology the need for more computation power increases constantly in order to solve more complex problems. Demand-driven, the computation capabilities of supercomputers increase rapidly and many institutions compete for the most powerful system. The Top500 project biannually updates a list of the 500 best supercomputers based on the LINPACK benchmark, which measures the system's floating-point rate of execution. At the

time of writing this paper, the best supercomputer is "Tianhe-2" at the National Super Computer Center in Guangzhou, China scoring 33.86 petaflops/s [1].

LINPACK is a library which solves a dense system of linear equation in double precision on distributed-memory systems. The theoretical peak performance of a system is generally not achieved. For different problem sizes the actual performance is measured and maximized. Along with the problem size, the benchmark result is also influenced by other parameters which are discussed later.

### 1.1 Project Description

The High Performance Linpack (HPL) benchmark has a configurable file called HPL.dat, which contains several parameters that may possibly impact the performance of a supercomputer. The goal of this project is to firstly, build HPL on the SuperMUC petascale system, and then, to identify which parameters impact the performance, with special regard to block sizes, type of process mapping (row-major or column-major) and type of swapping algorithm, and to determine how the problem size impacts scalability. Finally, to create a Persicope Tuning Framework (PTF) plugin that automates a new HPL.dat cre-

ation using "sed" (stream editor) and tunes the parameters in the SuperMUC system.

## 1.2 Outline

In section 2 tools are introduced which were used in the course of this project. Secondly, we discuss the tuning process and available benchmark parameters in detail before determining an initial configuration for a consumer grade laptop and SuperMUC. The implementation and functioning of the periscope plugin is described in section 4. Next, the benchmark results are presented, highlighting the impact of the automated tuning process. In conclusion we summarize our findings and discuss topics for future work.

# 2 Tools

## 2.1 Periscope

Periscope is an online distributed performance analysis tool[10] developed at the Technische Universität München that is used for dynamic tuning, which is a technique to automatically analyze performance issues of parallel MPI applications to evaluate single node performance[10]. Periscope comprises an analysis agent network, consisting of three agents, namely, the master agent, communication agents and analysis agents, which search for performance problems in one or more experiments, called iterations and use a monitoring interface to configure the measurements. After the end of the search, the detected problems are communicated to the master agent through the communication agents and the results of the performance problem detection are displayed on a user-interface.

The Periscope Tuning Framework (PTF) was developed under the AutoTune project [11] to combine and automate analysis and optimization of aspects such as energy consumption, inter-process communication and load balancing. PTF identifies tuning variants based on expert knowledge, evaluates the variants online and then, produces a report, which describes where and how to improve the code. PTF provides plugins for tuning high-level patterns for GPGPUs, HMPP codelets, MPI runtime, compiler

flag selection and the energy consumption [11].

## 2.2 BLAS

Basic Linear Algebra Subprograms (BLAS) are a collection of routines which provide basic vector and matrix operations. It is efficient and portable, which makes it suitable for usage in high performance calculations. It is required by HPL [4].

## 2.3 ATLAS

Automatically Tuned Linear Algebra Software (ATLAS) provides a complete BLAS API for C and Fortran77. For all operations it optimizes performance for a specific machine, allowing it to execute faster than plain BLAS. Since HPL requires BLAS it can be substituted by ATLAS [5].

## 2.4 OpenMPI

OpenMPI is an open source implementation of MPI (Message Passing Interface), a message passing library specification commonly used in the HPC community. Maintained by numerous academic, research, and industry partners this platform independent interface became the industry-standard for parallel programs in the last decade and has reached version 3.0 in September 2012. As it is designed for portability, efficiency, and flexibility, it runs on virtually any distributed memory, shared memory, and hybrid systems. Among others, the MPI interface has been implemented for C, C++, Fortran, and Python.

## 2.5 High-Performance Linpack

LINPACK addresses scalability in distributed memory systems by providing additional testing rules and environment through the Highly-Parallel LINPACK (HPL) NxN benchmark, which is used for measuring supercomputer performance in the Top500 list. HPL is a portable implementation of the HPL NxN benchmark, and generates and solves a random dense linear system of equations using Gaussian Elimination with partial pivoting and double precision floating point arithmetic. The HPL package is written in C and

requires the implementation of either the BLAS library, which can be automatically generated by the ATLAS tool, or the Vector Signal Image Processing Library (VSIPL). The HPL package provides testing rules to obtain the accuracy of the obtained solution, and timing rules to obtain the computation time to determine the best performance scenario.

The computational steps performed by HPL to obtain the benchmark rating are :

```

1  /* Generate and partition matrix
   * data among MPI computing nodes.
   */
2  /* ... */
3
4  /* All the nodes start at the same
   * time. */
5  MPI_Barrier (...);
6
7  /* Start wall-clock timer. */
8  HPL_ptimer (...);
9
10 /* Solve system of equations. */
11 HPL_pdgesv (...);
12
13 /* Stop wall-clock timer. */
14 HPL_ptimer (...);
15
16 /* Obtain the maximum wall-clock
   * time. */
17 MPI_Reduce (...);
18
19 /* Gather statistics about
   * performance rate (based on the
   * maximum wall-clock time) and
   * accuracy of the solution. */
20 /* ... */

```

## 2.6 Installation Script

We created a installation script (see appendix A) which automates the installation process of Periscope and HPL with all dependents. It was created for deployment on a clean installation of Ubuntu 14.04 LTS 64bit Desktop using VirtualBox. We suggest a hard disk size of 16GB or larger. The installation process

also depends on the system properties, the installation script can serve as useful guideline.

## 3 Tuning

### 3.1 Scalability

Scalability is an important aspect that must be considered while running problems on thousands of nodes. Strong scaling is the process of fixing a workload or the problem size and increasing the number of processors so that a program scales linearly when the speedup increases with the number of processing elements used. With respect to the total energy consumption, embarrassingly parallel codes achieve better energy efficiency with increasing system size[1]. Weak scaling is the process of increasing both the workload or the problem size and the number of processors so that the program scales linearly when the run-time stays constant while increasing the workload or problem size proportionally to the number of processors. HPL is a good weakly scalable application and gets more efficient with increasing memory [10].

### 3.2 Parameters

As discussed before HPL needs a set of parameters to derive a scenario for benchmarking. This is done via the HPL.dat file, which lists all available options. If multiple values are provided for individual parameters, HPL will execute all possible combinations in a separate scenario. In the following, parameters relevant for benchmarking are presented [6, 8]:

- Problem size (N): The problem should be as large as possible, whereas N denotes the size of the coefficient matrix. If the problem size exceeds the available memory, performance will decrease due to swapping. It is suggested to initially use about 80% of available memory.

$$N = \text{sqrt}((\#nodes * \#memoryPerNode * 0.8) / \text{sizeof(double)})$$

- Block size (NB): It is used for data distribution and computational granularity. Small values are desired as data distribution and load balancing improves. If chosen too small, the message overhead will increase and the available cache hierarchy might not be used properly. A value between 32 and 256 is suggested for modern systems.
- Process grid size (PxQ): The number of process rows (P) and columns (Q) depends on the number of available nodes and the physical network connection. It defines the number of nodes onto which HPL data should be distributed and executed, the process grid. Q must be slightly larger or equal to P. For simple Ethernet networks P should be very small since network scaling is limited.  $P*Q$  must be equal to the number of nodes in the system. If surpassed, additional nodes will not be utilized. The suggested range is:

$$P \leq Q \leq 3 * P$$

- Residual threshold: The residuals of the main algorithm will be compared to this value. The threshold is a real number of magnitude 1, whereas for negative values this threshold will not be checked. A suggested value is 16.0.
- Recursive factorization method (RFACT): Panel factorization is a matrix-matrix operation, which divides a large matrix into several submatrices. This is also known as LU decomposition. The left-looking and right-looking variants of this algorithm comply to the Doolittle method, dividing the matrix into lower and upper triangular unit matrices. The Crout variant is similar, but the returned lower matrix is not a unit matrix.
- Panels in recursion (NDIV): During each iteration of panel factorization the current matrix is divided into NDIV submatrices.
- Recursion stopping criteria (NBMIN): The panel factorization recursion stops when the matrix is composed of NBMIN columns or less.
- Panel factorization method (PFACT): After the factorization recursion stops, the algorithm con-

tinues with LU decomposition on a matrix-vector basis using a left-looking, Crout, or right-looking approach defined by RFACT.

- Lookahead depth (DEPTH): The lookahead depth defines the number of next panels which should be factorized immediately after being updated. For large problem sizes a value of 1 is suggested.
- Panel broadcasting method (BCAST): After panel factorization is complete, the resulting values are sent to other process columns using messages. Using different approaches leads to changes in execution time and throughput. There are multiple broadcast algorithms available:
  - Increased-ring: Every process has to send a message the next process consequently.
  - Increased-ring (modified): Process 0 sends a message to process 1 and 2 each.
  - Increasing-2-ring: Processes are split into two groups, whereas process 0 sends a message to the second group and acts as the first element of the first group.
  - Increasing-2-ring (modified): The modification to increasing-2-ring is to additionally apply the increasing-ring algorithm to the first group.
  - Long (bandwidth-reducing): The message is divided into Q equal pieces to be distributed to Q different processes in a synchronized fashion. The submessages are then exchanged between sets of two processes until every process received all parts of the original message.
  - Long (bandwidth-reducing modified): The modification to long (bandwidth reducing) is to additionally apply the increasing-ring algorithm.

The best method depends on the problem size and hardware performance. If the platform nodes are expected to outperform the network, "Long (bandwidth reducing)" (Lng) or "Long

(bandwidth reducing modified)” (LnM) can be recommended.

- Process mapping(PMAP): This parameter specifies how the MPI processes are mapped to the nodes. It is recommended to use row-major mapping for systems with multi-processor computer nodes.
- Swap algorithm (SWAP): This parameter specifies the swapping algorithm used during all tests. Available algorithms are based on binary-exchange (bin) and spread-roll (long). For mix, the binary-exchange will be used for a number of columns less than the swapping threshold before switching to spread-roll. The swapping threshold is only used for mix. For bin and long, the equilibrium phase can be enabled / disabled.
- Storage of panel triangle (L1, U): These parameters specify whether the lower (L1) and upper (U) triangle matrix should be stored in no-transposed or transposed form.
- Memory alignment: This parameters specifies the memory alignment HPL uses to allocate memory. The value is measured in double.

For the purpose of this project following parameters were considered for optimization:

- Problem size (N)
- Block size (NB)
- Process grid size (PxQ)
- Process mapping (PMAP)
- Swap algorithm (SWAP)
- Panel broadcasting method (BCAST)

Appendix B shows the optimal configuration for a Lenovo G510 laptop and appendix ?? shows the optimal configuration for SuperMUC.

## 4 Implementation

### 4.1 Periscope Plugin

Describe Flow Chart, Steps to be adapted etc.

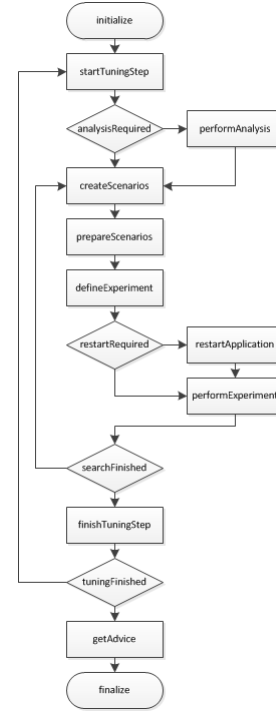


Figure 1: PTF flow

### 4.2 Parameter Manipulation - sed

## 5 Evaluation

Present data, findings, optimal parameters

## 6 Conclusion

Our Framework supports optimization of execution time and flops by changing one flag.

## 7 Future Work

The developed plugin currently only optimizes one parameter during each run. As a result dependencies between two or more parameters which might influence the overall performance are not considered. Optimizing for multiple parameters will prolong the

execution time significantly as each combination has to be checked.

In the course of this project "flops per second" was used as benchmark metric. Although widely used to classify supercomputers the resulting value can only serve as a baseline for comparison. In general, users of HPC systems desire a minimal execution time and/or low cost. For example "Tianhe-2" has a lower power to performance ratio than "Titan", second place in TOP500, which result in higher usage costs [1]. Thus, the fastest TOP500 system not necessarily yields the best results for every stakeholder. To optimize by execution time, the presented periscope plugin can be adapted to determine the average execution time over several runs and use it as benchmark metric. Likewise, the energy consumption of a system can be monitored using likwid-powermeter [2, 3].

Further, HPL solves dense matrix-matrix multiplication problems, but real-life HPC applications also contain differential equations, which need high bandwidth, low latency and irregular data access patterns, which HPL cannot execute efficiently. A new benchmark, called the High Performance Conjugate Gradient (HPCG) can deal better with lower computation-to-data-access ratios [9]. Moreover, LINPACK doesn't address issues like system latency and memory hierarchy.

## References

- [1] TOP500 Supercomputer Site, *November 2014*.  
<http://www.top500.org/lists/2014/11/>
- [2] Leibnitz Supercomputing Centre, *likwid - light weight performance tools*.  
<https://www.lrz.de/services/software/programmierung/likwid/>
- [3] Liwkid, *LikwidPowermeter*.  
<https://code.google.com/p/likwid/wiki/LikwidPowermeter>
- [4] BLAS, *Basic Linear Algebra Subroutines*.  
<http://www.netlib.org/blas/>
- [5] ATLAS, *Automatically Tuned Linear Algebra Software*. <http://math-atlas.sourceforge.net/>
- [6] HPL, *HPL Tuning*.  
<http://www.netlib.org/benchmark/hpl/tuning.html>
- [7] Luszczek and A. Petitet, *The LINPACK Benchmark: Past, Present and Future*, 2001, University of Tennessee, USA
- [8] Netlib, *HPL Algorithm*.  
[http://www.netlib.org/benchmark/hpl\\_oldest/algorithm.html](http://www.netlib.org/benchmark/hpl_oldest/algorithm.html)
- [9] Jack Dongarra and Michael Michael Heroux, *Toward a new metric for ranking high performance computing systems*, Sandia Report, SAND2013-4744, Volume 312, 2013
- [10] Tibidabo, *Making the case for an ARM-based HPC system*, Future Generation Computer Systems, Elsevier, 2013
- [11] Miceli, Civario, Sikora, Csar, Gerndt, Haitof, Navarrete, Benkner, Sandrieser, Morin, and Bodin, *AutoTune: A Plugin-Driven Approach to the Automatic Tuning of Parallel Applications*, Lecture Notes in Computer Science, p. 328-342, 2013, Springer Berlin Heidelberg, Germany

## A Installation Script

```
1 #!/usr/bin/env bash
2 cd ~
3
4
5 ### PERISCOPE ###
6
7 # install dependencies
8 apt-get install -y bison flex build-essential gfortran libace-dev
9     libxerces-c-dev libboost-all-dev doxygen automake autoconf libtool m4
10     libswitch-perl git mpich2
11 # clone repo
12 git clone http://periscope.in.tum.de/git-releases/PTF-PPE-Lecture.git
13 # prepare build
14 cd PTF-PPE-Lecture
15 ./bootstrap
16 cd ..
17 mkdir build
18 cd build
19 # configure
20 ../PTF-PPE-Lecture/configure --prefix=$HOME/install/periscope --enable-
21     developer-mode
22 # find out number of cores
23 NP='cat /proc/cpuinfo | grep processor | wc -l'
24 # build framework
25 make -j $NP
26 make install
27 export PATH=$HOME/install/periscope/bin:$PATH
28 echo "export PATH=$HOME/install/periscope/bin:$PATH" >> ~/.bashrc
29
30 ### BLAS ###
31
32 # For more information see following resources:
33 # http://linuxtoolkit.blogspot.de/2013/03/running-linpack-hpl-test-on-
34     linux.html
35
36 mkdir -p ~/src/
37 cd ~/src/
38 wget http://www.netlib.org/blas/blas.tgz
39 tar -zxvf blas.tgz
40 cd BLAS
41 gfortran -O3 -std=legacy -m64 -fno-second-underscore -fPIC -c *.f
42 ar r libfblas.a *.o
```

```

40 | ranlib libfblas.a
41 | rm -rf *.o
42 | export BLAS=~/.src/BLAS/libfblas.a
43 | ln -s libfblas.a libblas.a
44 | mv ~/.src/BLAS /usr/local/
45 |
46 |
47 | ### ATLAS ###
48 |
49 | # For more information see following resources:
50 | # http://math-atlas.sourceforge.net/atlas_install/node6.html
51 |
52 | cd ~
53 | # disable cpu throttling
54 | # Note: CPU throttling might not be necessary for virtual machines
55 | apt-get install -y gnome-applets
56 | apt-get install -y gnome-applets-data
57 | #UNABLE TO LOCATE PACKAGE# apt-get install -y libpanel-applets2-0
58 | /usr/bin/cpufreq-selector -g performance &
59 |
60 | # download latest ATLAS library
61 | wget http://www.netlib.org/lapack/lapack-3.5.0.tgz
62 | wget http://sourceforge.net/projects/math-atlas/files/Stable/3.10.2/atlas3
63 |   .10.2.tar.bz2
64 | tar -xvf atlas3.10.2.tar.bz2
65 | mv ATLAS ATLAS3.10.x
66 | cd ATLAS3.10.x
67 | mkdir Linux_C2D64SSE3
68 | cd Linux_C2D64SSE3
69 | # get unique dir name
70 | # enter SRCdir
71 | # create BLDdir
72 | # enter BLDdir
73 | ../configure -b 64 --shared --with-netlib-lapack-tarfile=/home/$USER/
74 |   lapack-3.5.0.tgz
75 | # Note: These next steps might take very long...
76 | make build
77 | # tune & build lib
78 | make check
79 | # sanity check correct
80 | # answer
81 | make ptcheck
82 | # sanity check parallel
83 | make time
84 | # check if lib is fast
85 | make install
86 | # copy libs to install dir
87 |
88 |
89 | ### OpenMPI ###
90 |
91 | # For more information see following resources:
92 | # http://www.sysads.co.uk/2014/05/install-open-mpi-1-8-ubuntu-14-04-13-10/

```



```

83 cd ~
84 # download latest package
85 wget https://www.open-mpi.org/software/ompi/v1.8/downloads/openmpi-1.8.4.
    tar.gz
86 apt-get install -y libibnetdisc-dev
87 tar -xvf openmpi-1.8.4.tar.gz
88 cd openmpi-1.8.4.tar.gz
89 ./configure --prefix="/home/$USER/.openmpi"
90 make
91 make install
92 export PATH=$HOME/.openmpi/bin:$PATH
93 echo "PATH=$HOME/.openmpi/bin:$PATH" >> ~/.bashrc
94 export LD_LIBRARY_PATH=$HOME/.openmpi/lib/:$LD_LIBRARY_PATH
95 echo "LD_LIBRARY_PATH=$HOME/.openmpi/lib/:$LD_LIBRARY_PATH" >> ~/.bashrc
96
97
98 ### HPL ###
99
100 cd ~
101 # download latest package
102 wget http://www.netlib.org/benchmark/hpl/hpl.tgz
103 tar -xvf hpl.tgz
104 cd hpl
105 # configure
106 ARCH="Linux_PII_CBLAS"
107 cp setup/Make.$ARCH .
108
109 # change parameters and practice sed
110 sed -i '/^ARCH*/ c\ ARCH = $(ARCH)' Make.$ARCH
111 sed -i '/^TOPdir*/ c\ TOPdir = $(HOME)/hpl' Make.$ARCH
112 sed -i '/^MPdir*/ c\ MPdir = /usr/lib/openmpi' Make.$ARCH
113 sed -i '/^MPinc*/ c\ MPinc = -I$(MPdir)/include' Make.$ARCH
114 sed -i '/^MPlib*/ c\ MPlib = $(MPdir)/lib/libmpi.so' Make.$ARCH
115 sed -i '/^LAdir*/ c\ LAdir = /usr/local/atlas/lib' Make.$ARCH
116 sed -i '/^CC*/ c\ CC = /usr/bin/mpicc' Make.$ARCH
117 sed -i '/^LINKER*/ c\ LINKER = /usr/bin/mpicc' Make.$ARCH
118
119 make arch=$ARCH
120 cd bin/$ARCH
121
122 # run HPL
123 mpirun -np 4 ./xhpl

```

## B HPL.dat for Laptop

```
1 HPLinpack benchmark input file
2 Innovative Computing Laboratory, University of Tennessee
3 HPL.out      output file name (if any)
4 6            device out (6=stdout,7=stderr,file)
5 3            # of problems sizes (N)
6 108304 346573 368233 # N — fantastically important; see ahead
7 1            # Number of block sizes
8 80 112 96    # Block sizes
9 0            PMAP process mapping (0=Row-,1=Column-major)
10 1           # of process grids (P x Q)
11 25          # Ps
12 26          # Qs
13 8.0         threshold
14 1           # of panel fact
15 0 2 1       PFACTs (0=left, 1=Crout, 2=Right)
16 1           # of recursive stopping criterium
17 4 2         NBMINs (>= 1)
18 1           # of panels in recursion
19 2           NDIVs
20 1           # of recursive panel fact.
21 1 2 0       RFACTs (0=left, 1=Crout, 2=Right)
22 1           # of broadcast
23 0 3 1 2 4   BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
24 1           # of lookahead depth
25 0           DEPTHS (>=0)
26 2           SWAP (0=bin-exch,1=long,2=mix)
27 256         swapping threshold
28 0           L1 in (0=transposed,1=no-transposed) form
29 0           U in (0=transposed,1=no-transposed) form
30 0           Equilibration (0=no,1=yes)
31 8           memory alignment in double (> 0)
```