NAMD Package Benchmarking on the Base of Armenian Grid Infrastructure

Armen Poghosyan\textsuperscript{1*}, Levon Arsenyan\textsuperscript{1}, Hrachya Astsatryan\textsuperscript{2}, Mikayel Gyurjyan\textsuperscript{2}, Hovsep Keropyan\textsuperscript{2}, Aram Shahinyan\textsuperscript{3}

\textsuperscript{1}International Scientific Educational Center, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia
\textsuperscript{2}Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia
\textsuperscript{3}Institute of Applied Problems of Physics, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia

Email: *sicnas@sci.am

Received November 22, 2011; revised December 24, 2011; accepted January 9, 2012

ABSTRACT

The parallel scaling (parallel performance up to 48 cores) of NAMD package has been investigated by estimation of the sensitivity of interconnection on speedup and benchmark results—testing the parallel performance of Myrinet, Infini-band and Gigabit Ethernet networks. The system of ApoA1 of 92 K atoms, as well as 1000 K, 330 K, 210 K, 110 K, 54 K, 27 K and 16 K has been used as testing systems. The Armenian grid infrastructure (ArmGrid) has been used as a main platform for series of benchmarks. According to the results, due to the high performance of Myrinet and Infini-band networks, the ArmCluster system and the cluster located in the Yerevan State University show reasonable values, meanwhile the scaling of clusters with various types of Gigabit Ethernet interconnections breaks down when interconnection is activated. However, the clusters equipped by Gigabit Ethernet network are sensitive to change of system, particularly for 1000 K systems no breakdown in scaling is observed. The infiniband supports in comparison with Myrinet, make it possible to receive almost ideally results regardless of system size. In addition, a benchmarking formula is suggested, which provides the computational throughput depending on the number of processors. These results should be important, for instance, to choose most appropriate amount of processors for studied system.

Keywords: Parallel Molecular Dynamics; NAMD Benchmarking; Grid

1. Introduction

It is fact that computational Grids \cite{1-5} consists of various computational layers. The computational resources can be integrated within the organization-institution, country, region, and worldwide. In order to ensure that Armenia would not stay behind in this important area, an appropriate national Grid infrastructure has been deployed on basis of available distributed computational resources.

Particularly in 2004, the first high Performance computing cluster (Armcluster) in the South Caucasus region had been developed in Armenia. Now the Armenian Grid infrastructure \cite{6-8} consists of seven Grid sites located in the leading research (National Academy of Sciences of the Republic of Armenia, Yerevan Physics Institute) and educational (Yerevan State University, State Engineering University of Armenia) organizations of Armenia. Apart from computing and storage resources, core Grid services \cite{9} which enable seamless access to all resources are provided to the national user communities. Armenian leading research and educational organizations actively engage in different International Projects such as \cite{10-13}. The Armenian National Grid Initiative has been established in 2009 and participates as a partner in the policy board of European Grid Initiative \cite{14}.

Interest in modeling of complex systems using molecular dynamics (MD) simulation has increased dramatically \cite{15-17}, and the parallel implementation makes it possible the fully understanding interesting phenomena and events, which occurs on long timescale and impossible to get from real experiments. During last decade the usage of parallel computational resources and super-computers leads to the significant progress in bio-systems modeling \cite{18-21}. Increasing in system dimensions and simulation time became possible with the linear increase of computational resources of distributed computing infrastructures. A number of MD software packages, like NAMD \cite{22}, GROMACS \cite{23}, CHARMM \cite{24} and AMBER \cite{25} are widely used and the most commonly
used packages are free available NAMD and GROMACS with open source codes, which are aimed at the high performance simulation with parallel support. The GROMACS developers claim as a “fastest MD” code, meanwhile the NAMD is a most scalable and efficient on parallel runs. It should be noted that both packages use the Message Passing Interface standard for communication between the computational nodes. Recently, the comparison of NAMD and GROMACS has been done by us [26], where comparable feature analysis of both packages has been carried out. It was stated that the GROMACS has been displayed as faster as NAMD, which is probably due to united atom character, meanwhile NAMD is more suitable for simulation of relatively small systems and for detailed analysis of the system in all atom character. It was also established that NAMD shows linear increase with increase of number of processors, however GROMACS receives saturation and even goes to the worst results.

The parallel scaling of GROMACS (version 3.3) molecular dynamics code has been studied by Kutzner and coworkers [27]. They have claimed the high single-node performance of GROMACS, however, on Ethernet switched (HP ProCurve 2848 switch) clusters, they find the breakdown in scaling, when more than two nodes were involved. They have tested 3Com 3870, 3Com 5500, HP 3400CL/24 and D-Link DGS 1016D switches for up to 10 nodes and have observed no change (same results as in case of HP 2848). For comparison, the authors performed the benchmarks with Myrinet-2000 interconnection.

The scaling of NAMD to ~8000 processors of Blue Gene/L system has been presented in [28]. They achieved 1.2 TF of peak performance for cutoff simulation and ~0.99TF with PME method. The corresponding speedup values were 5048 and 4090. The Blue Gene architecture has up to 65,536 dual core processors (i.e. 2^16 nodes) connected by a special auxiliary torus interconnection. The NAMD scaling has been performed on 3000 processors at Pittsburgh supercomputing center [29].

In order to better understand the parallel behavior of NAMD package, a series of benchmarks have performed within the ArmGrid infrastructure by using different types of interconnections and processor features. The purpose of current research is to evaluate the parallel performance of NAMD package and estimate the role of interconnection and processor performance. The results has practical meaning to the end users to effectively port and use computational resources of the Grid sites similar to the investigated clusters.

2. Benchmarks and Results

The NAMD package is a C++ based parallel program, which is implemented using CHARM++ communication library [30]. NAMD is parallelized via hybrid force/spatial decomposition using cubes (patches) with larger dimensions than the truncation radius is. The speedup estimation of the NAMD (version 2.7) package has been benchmarked on the base of Armcluster and Grid sites located in the State Engineering University of Armenia (SEUA), Yerevan State University (YSU) and Yerevan Physics Institute (YERPHI). The usage of the above mentioned computational resources dedicated by the following factors:

- Different interconnection technologies including Myrinet (ArmCluster), Infiniband (YSU), Gigabit Ethernet (SEUA, YERPHI).
- Different node architectures including Intel Xeon 3.06 GHz (ArmCluster) and Quad Core Intel Xeon (SEUA, YSU, YERPHI).

Though the nodes of SEUA, YSU and YERPHI Grid sites based on Intel Quad Core Xeon architectures, they use different types of network interconnections, mainboards, processors and other components: SEUA—MSI X2-108-A4M/E5420 2.5 GHz, YSU—HP ProLiant BL460c/E5405 2.0 GHz, YERPHI—Dell PE1950 III/E5420 2.5 GHz.

The system of ApoA1 (with 92224 atoms) available on the official web page of the NAMD package—is used as a benchmarking system (lipid bilayer with lipoprotein A1 in water environments). Particularly 1 fs timestep, the PME electrostatics, 12 Å van der Waals forces truncated at and 109×109×78Å³ cell size are used. There are lots of benchmarking results on mentioned system and therefore, it is reasonable to examine and compare to already existing data and test the computational resources.

In molecular dynamics simulations, the parameter which describes the speed of calculation, is expanding days per ns (days/ns). The mentioned parameter has been therefore examined (Figure 1 plots the computational throughput in days per ns versus the number of processors).

As can be seen from Figure 1, good results achieved on SEUA and YERPHI Grid sites if we take into account single processor. We have received about 21 days per ns experiment on SEUA and YERPHI Grid sites, and it should be noted that we have obtained almost same value using 2 processors on ArmCluster. It is naturally to suppose, that SEUA and YERPHI Grid sites with 2.5 GHz processors treat data more quickly than ArmCluster with 3.06 GHz and the reason is the caching and 32/64 bit differences. The 32 bit 3.06 GHz ArmCluster processor deal with data slower about twice than 64 bit 2.5 GHz SEUA and YERPHI Grid sites, however further increase of processors claim the importance of interconnection rather than processor performance. Before 16 processor, one can see almost continuously decrease, meanwhile
A. POGHOSYAN ET AL

Figure 1. NAMD performance and the estimation of days/ns against the number of processors.

Figure 2. NAMD performance and the estimation of days/ns on SEUA accordingly for 1000 K, 330 K, 210 K, 92 K and 54 K atoms systems.

Further increase of processors lead to the worst value for SEUA and YERPHI Grid sites. Because of using low latency and high-bandwidth Myrinet and Infiniband networks, the Armcluster and YSU Grid site scale well and show better results than SEUA and YERPHI Grid sites, however, due to Infiniband interconnection, the estimation of YSU sites shows rather good values than Armcluster with Myrinet support.

The difference is probably due to processor performance, as already mentioned, the caching. It is established that the large cache is perfectly suited for NAMD. It is important to note that the best result for the system of 92,224 atoms achieves on YSU Grid sites—48 processors. It is established that interconnection plays important role and it is obvious from Figure 1, that in comparison with simple Gigabit Ethernet, the Myrinet and Infiniband accelerates up to 4 - 7 times. The estimated speed of calculation was about 0.5 days per ns (48 processors in YSU site), which is rather good result.

To check the performance of Gigabit Ethernet equipped SEUA cluster depending on system size and to reveal the optimal number of processors, the additional testing have been performed. In this regards, we have tested 54 K, 210 K, 330 K and 1000 K systems on SEUA cluster and together with 92 K results, the data are shown in Figure 2. It is obvious that the changing of system size does not influence on the results when we take into account the systems up to 330 K atoms and we still see the breakdown in scaling at the optimal number of processors (at 16 point). After the 16 processors, with the increase of number of processors, we see also the increase of simulation time. In addition, one can see that with the decrease of system size, the sharply increase of the simulation duration occurs (the results of 92 K atoms on 24 processors is almost same as 210 K atoms on 16 processors). In case of 210 K and 330 K atoms, the difference between 16 and 24 processor results is about 0.7 - 1 days, meanwhile for 54 K and 92 K systems the differences are ~1.9 and ~2.2 days correspondingly. Hence, one can assume, that in some manner, Gigabit Ethernet equipped clusters have limitations and in our case the $2 \times [\text{Node}] = 16$ is the optimal number of processors. However, it is most important to note, that this assumption is true, when the system size do not exceed the so called “critical amount” of atoms and therefore, for small systems, it is recommended to use $2 \times [\text{Node}]$ processors to avoid wasting computational resources. On the other hands, the further increase of system size (testing of large 1000 K systems) shows that so called “critical point” (16 processors) disappears (no breakdown is observed) and we see the decrease of estimated days per ns with the increase of the processors. As already mentioned, NAMD is parallelized via hybrid force/spatial decomposition, where for each pair of neighboring cubes (called patches) an additional force computation object is assigned, which, in its turn, can be independently mapped to any processor. For relatively small systems, the problem is that the increasing of processors lead to the spending more time on communication, however, the further increase of system size claims that the increasing of processors is more efficient rather than any type of interconnection between processors.

To testify above mentioned suggestion, an additional benchmarks on Armcluster have also been performed in order to clarify the Myrinet equipped ArmCluster’s features depending on system size. Together with now-standard 92 K system, the 210 K, 110 K, 54 K, 27 K, 16 K systems are also examined and the curves are shown in Figure 3.
We see that increasing the system size lead to the increase to the speed of calculation, as well as, if we use the united atom character instead of all atom accelerate parallel simulation compared with all atom model.

The main aim of this work is to estimate and extrapolate our benchmark findings. As one can see, there are some peculiarities depending on system size, and therefore we were trying to get in some manner a “universal” formula, which will describe the behavior of changing. According to our testing, depending on coefficients, the following formula is better describing the curves:

\[ E_d = \text{Estimated}_{\text{days} \text{per} \\text{ns}} = \alpha \frac{N}{N_p^2} + \beta \]  

where \( \alpha \) and \( \beta \) are coefficients, which describes the physical nature of cluster (processor type, frequency, etc.) and the network (bandwidth, latency) correspondingly, the \( N \) is number of atoms, and the \( N_p \) is a number of processors. This is surely a roughly estimation, however, for ArmCluster, after the testing, we obtained results near the testing points. We have estimated the coefficients and following results were obtained.

\[ \alpha_1 = \alpha_o \\
\alpha_2 = \alpha_1 + 2 \\
\alpha_3 = \alpha_2 + 2, \quad n = 1 \cdots N_p \]

\[ \ldots \\
\alpha_n = \alpha_{n-1} + 2 \]

i.e. the coefficient can be interpret as \( \alpha_n = \alpha_{n-1} + \text{Duality} \), where \( \alpha_o \) is defined as processor frequency, i.e., in case of ArmCluster—\( \alpha_o = 3.06 \). The network characterized parameter \( \beta \), which is roughly set to zero, is estimated to be as follow:

\[ \beta(\text{network}) = \begin{cases} f(\chi, \lambda, N_p, N) & \text{if } N < N_c \\
\{0 \text{ or } 1\} & \text{else} \end{cases} \]

where \( N_c \) is a so called critical (or optimal) number of processors, which depends on system size and network type. The network coefficient \( \beta \) and the complex function \( f(\chi, \lambda, N_p, N) \) depends on many factors, like network bandwidth, latency time, etc., however the increasing processors \( N_p > N_c \) shows that the network parameter displays as just a correction \( \beta \to 0 \). The testing and formula estimated data are shown in Table 1. We see a value drift for 4 - 16 processor range, meanwhile from 16 to 48 processor data are in well agreement with estimated findings and even roughly estimation lead to the good comparison results for Myrinet networked ArmCluster. In addition, we also estimated 1000 K large system performance on 40 processors, and we received 13.16 days per ns, and the corresponding calculation shows 13.14, which is surely excellent agreement.

The further step is the estimation of other cluster data, namely the estimation of Gigabit Ethernet equipped SEUA cluster data. According to suggested Formula (1), we have calculated and compared the data, where correspondingly the \( \alpha \) and \( \beta \) coefficients are set to follows: \( \alpha_o = \alpha_{o-1} + \text{Duality} \) \( (\alpha_o = 2.5) \), and the \( \beta \) correction is set to be 1. The data for 1000 K large system is shown in Figure 4. As one can see the testing points are somewhat in agreement with suggested formula data.

To check the formula, we have also performed benchmarks of 210 K system on Blue Gene/P supermachine (IBM Blue Gene/P: PowerPC 450 processors, a total of 8192 cores) at Bulgarian Supercomputing Centre. The benchmarking data together with data according to Formula (1) are shown in Table 2. In overall, we find the good agreement with estimated findings.
The next estimated parameter, which describes the parallelization, is the speedup $\sigma$ coefficient. The speedup $\sigma$ measures the efficiency of using multiple processors with respect to a single one. If we take into account the communication time between processors, according to Amdahl’s law, the speedup can be interpreted as follows.

$$\sigma = \frac{1}{J_p / N_p + J_s + C(N_p)} \tag{2}$$

where $N_p$—number of processors, $J_p$ and $J_s$ is the amount of job, which can be done in parallel and in serial ($J_p + J_s = 1$) and the $C(N_p)$ is a function of the number of processors and describes the network depending on the network bandwidth and the latency time. In case of ideal parallelization (Figure 5), all the operations would be perfectly parallelized—$J_p = 1$ ($J_s = 0$ and $C(N_p) = 0$), which means $N_p$ times faster operations. No gain observes when $J_s = 1$.

Here, the speedup (i.e. the speedup relative to one processor) was calculated by simple dividing executing time on single-processor $T_i$ to the executing time on $N$ processors—$T_N$. $\sigma = \frac{T_i}{T_N}$. The speedup calculation profiles are exemplified on Figure 5. One can see that high performance infiniband shows ideally speedup value and the breakdown in scaling is observed in case of SEUA and YERPHI Grid sites after 24 processors, as it

\begin{table}[h]
\centering
\caption{Comparison of testing and formula data on BlueGene/P supermachine.}
\label{tab:bluegene_p}
\begin{tabular}{llllllll}
\hline
\textbf{Processors} & \multicolumn{2}{c}{\textbf{210 K atoms}} & \multicolumn{2}{c}{\textbf{110 K atoms}} & \multicolumn{2}{c}{\textbf{92 K atoms}} & \multicolumn{2}{c}{\textbf{54 K atoms}} & \multicolumn{2}{c}{\textbf{27 K atoms}} \\
\hline
256 & 0.922 & 312 & 0.461 & 1024 & 0.230 & 2048 & 0.115 & 4096 & 0.057 \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{speedup.png}
\caption{The speedup calculation on different grid sites.}
\end{figure}
is obvious from Figure 1. The breakdown of SEUA and YERPHI Grid sites is probable due to overloading of Gigabit switches. The ArmCluster also shows linear increase, however, is a bit lower than YSU with Infiniband support.

3. Conclusions

The results should be important to choose the most appropriate amount of computational resources for various types of interconnections and studied systems sizes. As a result of series of benchmarks, a formula has been obtained to provide the computational throughput depending on the number of processors, which, in our opinion, should be testified with other benchmarks in the literature.

It is stated the in contrary to high performance Myrinet and Infiniband clusters, for Gigabit Ethernet there is limit of optimal number of processors for relatively small systems. The further increase of the system size shows that the increasing of processors is more suitable than the any type of interconnection between processors: after 16 processor, Gigabit Ethernet equipped clusters shows breakdown in scaling, however for less than 16 CPUs, it scales very well. Therefore, it is expected to study the systems with various sizes in order to receive some peculiarities for GROMACS software package and as well to check and compare GROMACS data with exiting NAMD results.

4. Acknowledgements

This work was supported in part by the European Commission under EU FP7 project HP-SEE (under contract number 261499).

REFERENCES


