Computer Science - Research and Development manuscript No.
(will be inserted by the editor)

Performance Analysis and Prediction for distributed homogeneous Clusters

Heinz Kredel · Hans Günther Kruse · Sabine Richling · Erich Strohmaier

March 15, 2012

Abstract We present a new performance model based on the roofline concept for the analysis and performance prediction of distributed computing clusters. The background for our performance modeling is the 28 km InfiniBand interconnection between two bwGRiD clusters each consisting of 140 compute nodes in day-to-day production use. The model is used to analyze the MPI performance of inner-cluster communication compared to inter-cluster communication. We compare the new modeling results to our earlier stochastic model [RHKK2010] where we could give an estimate on the bandwidth requirements for doubling the performance of an application (LinPack in the simplest example). We will derive some bounds for the size of regions in a cluster and the scaling of the maximal speed-up for the region-region-interconnected network.

Keywords performance model · performance prediction · inter-cluster communication · roofline model

1 Introduction

In [2,3] we presented a stochastic model to analyze and predict performance of distributed computing clusters. In this paper we will present a new model based on the roofline concept [19,20]. The performance modeling was inspired by the requirements to operate two grid clusters over a distance of 28 km as a single compute resource. We wanted to choose between two options: allowing applications to span over the long distance interconnect or to restrict applications to run on one side only. Although there are applications like Monte Carlo simulations with low communication requirements, some others, like LinPack, have higher communication requirements. With our model we could predict, that the bandwidth between the remote clusters will always be too low (under given budgetary constraints) to run communication intensive applications across two sites. Given this insight we set-up the batch system so that all jobs will run completely on one side only.

With the new model, not using any stochastic tools, we attempt to reach the same performance prediction figures. As the two clusters will have to be replaced in 2013 we attempt to use the new model for questions arising from the sizing of the new cluster. For example, how many nodes / CPUs / cores would be optimally within one 'island' (a region with fully-non-blocking InfiniBand interconnection), or is there a lower bound for such a number, or what will be the maximally expected speed-up depending on the configured bandwidth between 'islands' respectively the long distance interconnection.

1.1 Related work

The performance of InfiniBand for SAN and WAN cluster connections has been studied in [13], [15] and [16]. Our performance modeling is based on the Linpack benchmark with its specific computational workload and communication patterns. For the study of other application workloads in a wide area setting see for example [6] or [14]. The previous stochastic model has

H. Kredel, H.G. Kruse (retired)
IT-Center, University of Mannheim, Germany, E-mail: kredel@rz.uni-mannheim.de, kruse@rz.uni-mannheim.de

S. Richling
IT-Center, University of Heidelberg, Germany, E-mail: richling@urz.uni-heidelberg.de

E. Strohmaier
Future Technology Group, Lawrence Berkeley National Laboratory, Berkeley, E-mail: estrohmaier@lbl.gov

also been applied successfully to (distributed) Web applications [17,18]. Further related work is mentioned in the paper as required.

1.2 Outline

In Section 2 we introduce the bwGRiD cooperation and infrastructure. Section 3 presents our new performance model and discusses advantages over the old stochastic model and the application to bwGRiD. Finally Section 4 draws some conclusions.

2 Distributed bwGRiD clusters

In this section we summarize the background for our performance modeling. We focus only on the aspects which are needed in the understanding of the rest of the paper. Other aspects, like operating and administration considerations are discussed in [4,5]. This section contains summaries and revised parts of [3].

Part of the German grid activities (*D-Grid Initiative*[1]) is a community project of the universities of Baden-Württemberg (BW) *bwGRiD*[2]. It consists of compute clusters at the universities in Stuttgart, Ulm (together with Konstanz), Karlsruhe, Tübingen, Freiburg and Mannheim together with Heidelberg. Besides local storage at each site, there is a central storage unit in Karlsruhe. The funding requires all sites to provide access to all D-Grid virtual organizations (VOs) by at least one middle-ware from the supported D-Grid software stack. So bwGRiD architecture is a distributed system with local administration by the IT-Centers of each university.

The objectives of the bwGRiD project are verifying the functionality and the benefit of Grid concepts for the HPC community and learning how to manage organizational and security problems. The project should further develop new cluster and grid applications, solve license difficulties and enable the computing centers to specialize for certain application areas relevant to their home university. To improve resource utilization and application support, each cluster is configured to allow a transparent use for all users (from all universities and VOs).

2.1 Hardware

The hardware consists of 10 bladecenters in Heidelberg and 10 bladecenter in Mannheim. Each bladecenter contains 14 IBM HS21 XM blades and each blade contains

- 2 Intel Xeon CPUs, 2.8 GHz
- each CPU with 4 Cores
- 16 GB Memory
- 140 GB Hard Drive (since January 2009)
- Gigabit-Ethernet (1 Gbit)
- InfiniBand Network (20 Gbit)

This makes a total of 1120 CPU cores on each side.

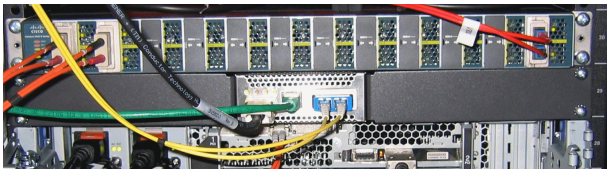

Fig. 1 Obsidian Longbow interconnection

The concept for a tight interconnection of the two clusters in Heidelberg and Mannheim was developed in 2008 and was finally operational in mid 2009. The main technical part is InfiniBand over Ethernet over fibre optics which is provided by the Longbow adapter from Obsidian (see figure 1, [9]). This adapter has an InfiniBand connector (black cable) and a fibre optic connector (yellow cable) and does the packaging of the InfiniBand protocol within optical Ethernet. The theoretical bandwidth is 10 Gbit/sec for the Obsidian and 20 Gbit/sec for a single InfiniBand connection. The universities already had a dark fiber connection between the two IT-centers which was used for (file) backup and fast campus connection up to then. An additional component from ADVA had to be used for the transformation of the white light from the Longbow to one color light transmitted over the dark fibre. Neglecting the costs for the existing dark fibre the interconnection costs about 150 TEUR.

2.2 Performance

The MPI performance over such an interconnection is shown in figure 2. Our measurements find a latency for a local (inner cluster connection) of ~ 2 μsec and 145 μsec over the interconnection to the other site. The bandwidth is 1400 MB/sec local and 930 MB/sec over the interconnection. For message sizes of 10^9 byte the remote bandwidth is 50 % lower than the local value. For smaller message sizes, the situation is worse. In summary our experiences with the interconnection network are as follows.

- The cable distance from Mannheim to Heidelberg is 28 km (18 km linear distance in air). So the light needs at least 143 μsec for this distance, according to a refractive index of 1.53 in SiO_2 (see [21]).

[1] www.d-grid.de
[2] www.bw-grid.de

– The latency is high: 145 μsec = light transit time +2μsec compared to a local latency of only 1.99 μsec point-to-point.
– The bandwidth is as expected: about 930 MB/sec over the interconnect compared to the local bandwidth of 1200-1400 MB/sec between two nodes. The theoretical bandwidth is 20 Gbit/sec for a single InfiniBand connection and 10 Gbit/sec for the long distance.

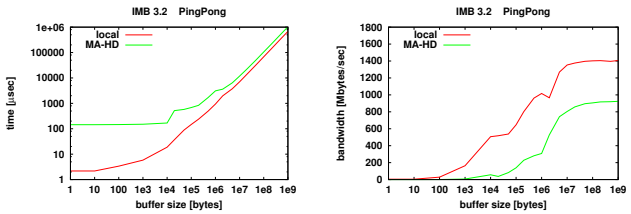

Fig. 2 MPI latency and bandwidth Mannheim-Heidelberg

2.3 Operating Considerations

As cooperating IT-Centers we have to operate the bwGRiD clusters for the HPC users at our universities and scientists from virtual organizations in Germany. To minimize system administration we descided to run both clusters as single system image. This implied a fast InfiniBand interconnection to access the two parallel Lustre filesystems on both sides for jobs running on some side. The filesystems have been connected with the InfiniBand interconnection because the ethernet interconnection was too slow (1Gbit shared for 14 nodes).

Under these constraints it was not clear if we should allow jobs with node numbers greater than 140 or to improve through-put to allow jobs to run on nodes on both sides. There are perfectly scaling codes, which could profit from nodes connected with different communication bandwidths (e.g. Monte Carlo simulations). However, we decided to take a conservative approach and first check how other applications would perform in such a situation. As an application with higher interconnection bandwidth requirements we looked into the LinPack benchmark. The findings of this and earlier papers indicated that it is not advisable to allow jobs spanning nodes on both sides under a 10Gbit/s connection and that is not feasable to buy enough bandwidth to allow at least Linpack to run optimally. So in order to avoid bad user experience with poor scaling applications we decided to keep all nodes of a job on one side.

One question for the next generation system is if the expected new bandwidth will allow the execution of jobs running on both sides. We will try to answer this question at least for LinPack and if time permits for other applications.

2.4 Next generation bwGRiD

The current bwGRiD clusters will reach five years of operation time in 2013. So the performance of the hardware will be no longer competitive and will have to be replaced. For the next generation of bwGRiD hardware we expect the following performance figures

– expected bandwidth of 56 Gbit/s for InfiniBand between two nodes
– expected bandwidth of 450 Gbit/s (as 8×56 Gbit/s) for inter cluster communication at one 'side'
– expected bandwidth of 160 Gbit/s (as 4×40 Gbit/s) for the longer distance between the sites
– the maximal possible bandwidth will be 480 Gbit/s (as 12×40 Gbit/s) limited by dark-fiber technology and the existing cables.

3 Performance Modeling

This section describes a simple and transparent model for the analysis of a distributed homogenous architecture, such as two sites in Heidelberg and Mannheim – interconnected by InfiniBand.

In two earlier publications [2,3] we presented some good results in performance predictions by a stochastic approach. Indeed most of applications in High Performance Computing (HPC) are straight-on and the stochastic method seems not to be the appropriate tool.

In order to avoid these complications we propose a new approach, based on the ideas of MultiCore-Analysis by Williams, Waterman and Patterson [19] and the work of Hill and Marty [20]. The reasons for applying these concepts are the analogy of MultiCore-Systems and homogenous clusters. Our model describes two cluster regions I_1 and I_2, each with n CPUs (nodes / cores) and each CPU with a theoretical arithmetic performance l_{th} [GFLOP/sec]. The interconnection of the CPUs in a single region has a bandwidth b^I [GByte/sec], the bandwidth between the two regions will be b^E [GByte/sec]. The load consists of a number of arithmetic operations $\#op$ and the data $\#b$, measured in bytes.

3.1 Analysis of a single Region

First we discuss the model of single region I_1 or I_2. The total time t_v for the load ($\#op$, $\#b$) will be split up in

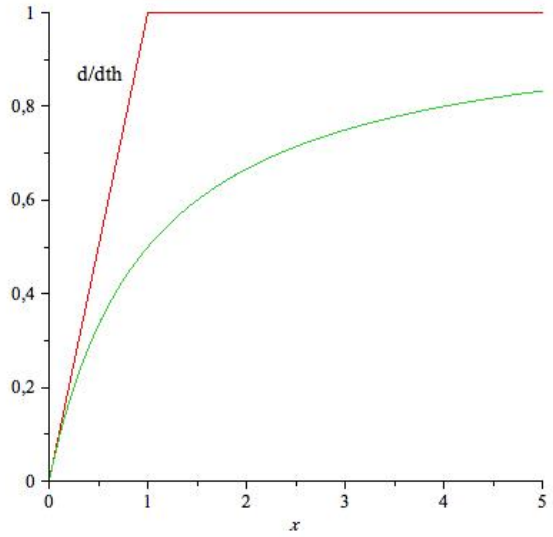

Fig. 3 Throughput for additive (green) and overlapping (red) concepts.

two phases *computation* t_R and *communication* t_C and we calculate

$$t_R \sim \frac{\#op}{d} \qquad \text{and} \qquad t_C \sim \frac{\#b}{b^I}. \tag{1}$$

d is the throughput of the cluster region. The total time t_V can be computed by simple addition $t_R + t_C$ or by $\max(t_R, t_C)$ in the case of overlapping phases:

$$t_V \sim \begin{cases} \frac{\#op}{d} + \frac{\#b}{b^I} & \text{additive} \\ \max\left(\frac{\#op}{d}, \frac{\#b}{b^I}\right) & \text{overlapping} \end{cases} \tag{2}$$

We do not know the real throughput d, but we know the theoretical value d^{th} and replacing d by d^{th} we get

$$t_V \geq \begin{cases} \frac{\#op}{d^{\text{th}}}\left(1 + \frac{\#b}{b^I}\frac{d^{\text{th}}}{\#op}\right) & \text{additive} \\ \frac{\#op}{d^{\text{th}}}\max\left(1, \frac{\#b}{b^I}\frac{d^{\text{th}}}{\#op}\right) & \text{overlapping} \end{cases} \tag{3}$$

by some simple algebraic manipulations. Introducing the arithmetic intensity $a = \#op/\#b$ [19], and the variable $a^* = d^{\text{th}}/b^I$, it follows for the real throughput d with the dimensionless parameter $x = a/a^*$

$$d \leq \begin{cases} d^{\text{th}}\frac{x}{x+1} & \text{additive} \\ d^{\text{th}}\min(1, x) & \text{overlapping} \end{cases} \tag{4}$$

The shape of the functions $x/(x+1)$ and $\min(1, x)$ is shown in Fig. 3 which favourites clearly the overlap of communication and computation. In order to be more realistic we prefer the additive case in our further analysis.

The explicit dependance from the number n of CPUs with throughput l_{th} and the bandwidth b^I is hidden in the dimensionless parameter x. Therefore we elaborate with $d^{\text{th}} = n \cdot l_{\text{th}}$

$$x = \frac{a}{a^*} = \frac{\#op}{\#b} \cdot \frac{b^I}{d^{\text{th}}} = \frac{1}{n} \cdot \frac{\#op}{\#b} \cdot \frac{b_0^I}{l_{\text{th}}} \cdot \frac{b^I}{b_0^I} = \frac{x' \cdot z}{n} \tag{5}$$

Further we define $x' = (\#op/\#b) \cdot (b_0^I/l_{\text{th}})$ and $z = \frac{b^I}{b_0^I}$ where b_0^I represents a reference bandwidth. The result for the throughput is

$$d \leq l_{\text{th}} \frac{x'z}{1 + \frac{x'z}{n}} \to (x'z) l_{\text{th}} \Big|_{n\to\infty} \tag{6}$$

and for the speed-up

$$Sp = \frac{d(n)}{d(1)} = \frac{1 + x'z}{1 + \frac{x'z}{n}} \to 1 + x'z \Big|_{n\to\infty} \tag{7}$$

The last limit suggests a constant shape of Sp and therefore a decreasing efficiency $Sp/n \to 0$ with $n \to \infty$, reflecting Amdahl's law (strong scaling). Figure 4 shows the behavior of Sp for different values x' and z (increasing bandwidth).

Since x' depends on problem size, we observe a simple scaling by variation of z (bandwidth). From further interest is $n_{1/2}$, the number of CPUs which guarantees half of the theoretical performance, a simple calculation results in $n_{1/2} = x'z$, which shows the same scaling. $n_{1/2}$ may be a parameter to **estimate a lower bound for the number of CPUs** in a single region – dedicated to the same class of applications and characterized by x' or $a = \#op/\#b$.

3.2 Analysis of two interconnected Regions

In this second step we discuss an application running on two homogenous cluster regions I_1 and I_2, supposing a symmetric distribution. That means in each region we have the same part $(\#op/2, \#b/2)$ of the total load. Our strategy in order to determine the throughput and the speed-up is quite similar to the previous section 3.1 – also we restrict ourselves to the additive case.

Let $\#x$ the number of bytes exchanged between the two regions, b^E the bandwidth of interconnection and $t_V^{(1)}$ the total time for the load $(\#op/2, \#b/2)$ in one region. It follows $t_V \sim t_V^{(1)} + \#x/b^E$, in which the term $\#x/b^E$ summarizes the communication time between the two regions. With the arithmetic intensity $a = (\#op/2)/(\#b/2)$ and $a^* = d^{\text{th}}/b^I$ we get for the total time t_V

$$\begin{aligned} t_V &\geq \frac{(\#op/2)}{d^{\text{th}}}\left(1 + \frac{a^*}{a}\right) + \frac{\#x}{b^E} \\ &= \frac{\#op}{2d^{\text{th}}}\left[1 + \frac{a^*}{a} + 2\frac{d^{\text{th}}}{b^E}\frac{\#x}{\#op}\right] \end{aligned} \tag{8}$$

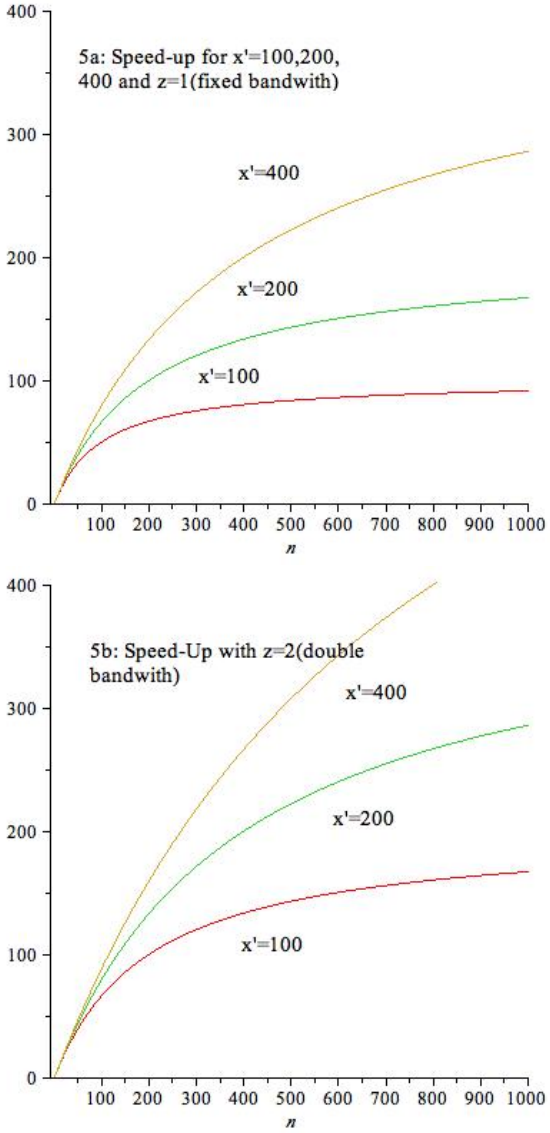

Fig. 4 Speed-up Sp for different values x' and z.

and finally for the throughput

$$d \leq 2d^{\mathrm{th}} \frac{1}{1 + \frac{a^*}{a} + 2\frac{d^{\mathrm{th}}}{b^E}\frac{\#x}{\#op}} \tag{9}$$

With the replacements $d^{\mathrm{th}} = n \cdot l_{\mathrm{th}}$, $x = a/a^* = x'/n$, $x' = (\#op/\#b)/(l_{\mathrm{th}}/b^I)$ and the rewriting

$$\begin{aligned} y &= \frac{\#op}{\#x} \cdot \frac{b^E}{2d^{\mathrm{th}}} = \frac{1}{2} \cdot \frac{1}{n} \cdot \frac{\#op}{\#b} \cdot \frac{b^I}{l_{\mathrm{th}}} \cdot \frac{\#b}{\#x} \cdot \frac{b^E}{b^I} \\ &= \frac{1}{2}\frac{x'}{n}\left(\frac{\#b}{\#x}\right)\left(\frac{b^E}{b^I}\right) \end{aligned} \tag{10}$$

$$y' = \frac{1}{2}x'\left(\frac{\#b}{\#x}\right)\left(\frac{b^E}{b^I}\right) \tag{11}$$

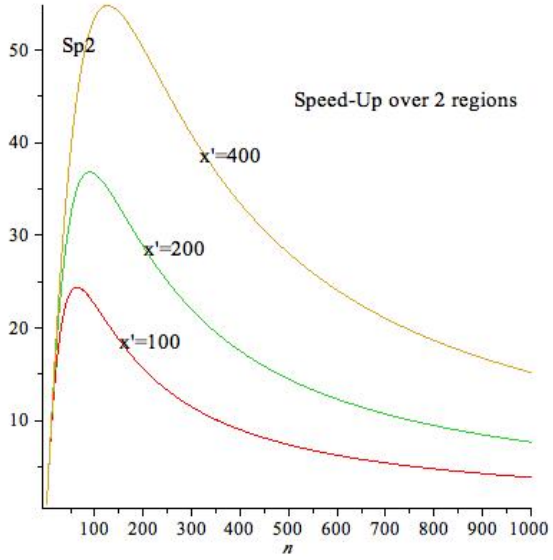

Fig. 5 Speed-up $Sp2$ for $z' = 1$ and different values of x'.

follows

$$d \leq 2n\, l_{\mathrm{th}} \frac{1}{1 + \frac{1}{x} + \frac{1}{y}} = 2l_{\mathrm{th}} \frac{x'y'}{x' + y' + \frac{x'y'}{n}} \tag{12}$$

Taking account that the interconnection is a shared medium, we have to substitute $b^E \to b^E/p(n)$, in which $p(n)$ is the number of communicating CPUs. Therefore we have to modify the previous expression

$$d \leq 2l_{\mathrm{th}} \frac{x'\frac{y'}{p}}{x' + \frac{y'}{p} + \frac{x'y'}{n \cdot p}} = 2l_{\mathrm{th}} \frac{x'y'}{px' + y' + \frac{x'y'}{n}} \tag{13}$$

and can calculate the speed-up

$$Sp2 = \frac{x' + y' + x'y'}{p(n)x' + y' + \frac{x'y'}{n}} \to 0\Big|_{n\to\infty} \tag{14}$$

presumed that $p(n)$ is a monotone increasing function of n, an assumption which is reasonable. In Fig. 5 we show the behavior of $Sp2$ for some values of x', y' and $p(n) = \alpha \cdot n$ with $\alpha = 1/20$.

The expression for $Sp2$ and the shape of the function in Fig. 5 suggests the existence of a maximum, determined by $(dSp2(n)/dn) = 0$

$$\frac{dp(n)}{dn}x' - \frac{x'y'}{n^2} = 0 \Rightarrow n_E^2 \cdot \frac{dp(n)}{dn} = y' \tag{15}$$

If we know $p(n)$, such as $p(n) = \alpha \cdot n$ with $0 < \alpha < 1$, Eqn. 15 can be solved easily. The result is $n_E = \sqrt{y'/\alpha}$ and the maximum of speed-up behaves like

$$Sp2(n = n_E) = \sqrt{\frac{y'}{\alpha}} \cdot \frac{1 + x' + x'/y'}{1 + x' + \sqrt{\frac{y'}{\alpha}}} \tag{16}$$

The value of n_E may be an realistic **estimation of the optimale size of a region** (for a fixed class of

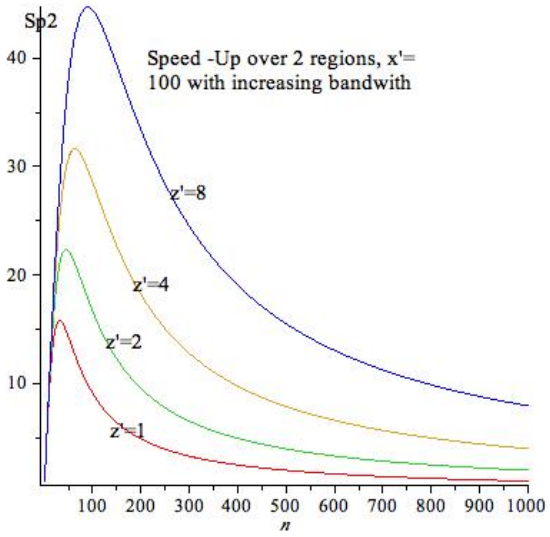

Fig. 6 Speed-up $Sp2$ for $x' = 100$ with increasing bandwidth z'.

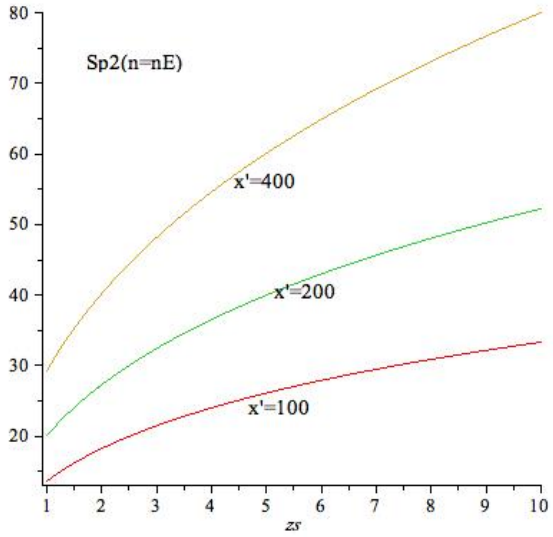

Fig. 7 Max. Speed-up $Sp2$ by variation of bandwidth $zs = z'$.

applications, characterized by the parameter x'). According section 3.1 we can calculate $n_{1/2}$ by $d(n = n_{1/2}) = (1/2) \cdot 2 \cdot n \cdot l_{\text{th}}$, which results in the equation $n_{1/2} \cdot p(n_{1/2}) + (y'/x') \cdot n_{1/2} - y' = 0$, which is a simple quadratic type.

Much more interesting is the discussion of Eqn. 14 in the focus of the application and the interconnection bandwidth. We refer to Eqn. 11, defining $r = \#b/\#x$ and $z'' = b^E/b^I$, then for the speed-up follows

$$Sp2 = \frac{2 + (1 + x')(rz'')}{2p(n) + (1 + \frac{x'}{n})(rz'')} \tag{17}$$

Inspecting this, we will observe that an increasing bandwidth $z'' > 1$, represents a scaling of the data $\#b$ and vice versa. That is the reason why we introduce $z' = r \cdot z''$ as a new scaling variable, which implies

$$Sp2 = \frac{2 + (1 + x')z'}{2p(n) + (1 + \frac{x'}{n})z'} \leq \frac{x'z'}{2p(n) + \frac{x'z'}{n}} \tag{18}$$

Figure 6 shows the behavior of the approximation, like the corresponding speed-up in Sect. 3.1. But remember that in the definition of $z'' = b^E/b^I$ the internal bandwidth is now the reference bandwidth b^I_0.

In the context of a qualitative discussion we understand better the scaling in the external bandwidth $b^E(z')$ if we compute $Sp2(n = n_E)$ at the maximum. The calculation is straight-forward and has a little bit crowdy result. But it is possible to derive a very good approximation

$$Sp2(n = n_E) \sim \frac{x'}{1 + 2\sqrt{\frac{2\alpha x'}{z'}}} \tag{19}$$

for the exact formula. Increasing z (bandwidth b^E) implies a monotone growth of the maximal speed-up (Fig. 7), but the effect is not linear.

If we fix z'_0 and ask for an z', which doubles $Sp2(z'_0)$ we get

$$\sqrt{\frac{z'}{z'_0}} = \frac{4\sqrt{\frac{2\alpha x'}{z'_0}}}{2\sqrt{\frac{2\alpha x'}{z'_0}} - 1} \tag{20}$$

Using the approximation in Eqn. 18 directly yields $z' = 4z'_0$, which is very inaccurate.

3.3 Modeling bwGRiD

Now we will apply our results from Section 3.1 and Section 3.2 on real configurations, summarized in Section 2 and described in [1,5] in detail. The two sites/regions consists of 2×140 nodes (with 8 cores per node; core=cpu) interconnected by InfiniBand over Ethernet.

The relevant data of the hardware are

- $l_{\text{th}} = 8.5$ [GFLOP/sec] for a core
- $b^I = 1.5$ [GByte/sec], $b^I_0 = 1.0$ [GByte/sec]
- $b^E = 1.0$ [GByte/sec]

For the load we choose LinPack with the problem size $n_p = 10000, 20000, 30000, 40000$:

$$\#op \sim \frac{2}{3}n_p^3 \qquad \text{and} \qquad \#b \sim 2n_p^2 \cdot w \tag{21}$$

w represents the number of bytes per word ($w = 8$). The arithmetic intensity is $a = \#op/\#b = n_p/(3 \cdot w) = n_p/24$ and the parameter $a^* = l_{\text{th}}/b^I = 8.5/1.5 = 17/3 < 6$. Hence it follows $x' = a/a^* = n_p/136$ and

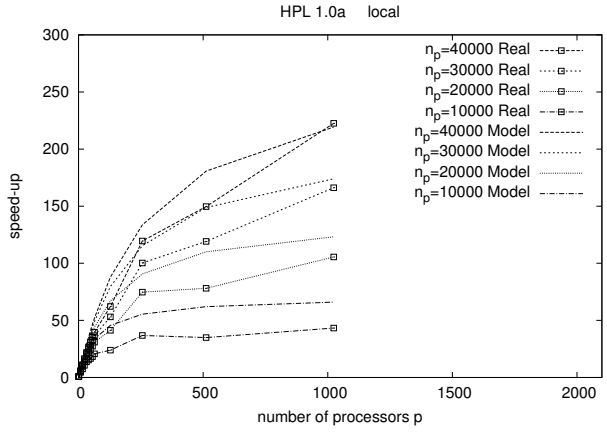

Fig. 8 Speed-up comparison of measurements and model for one region.

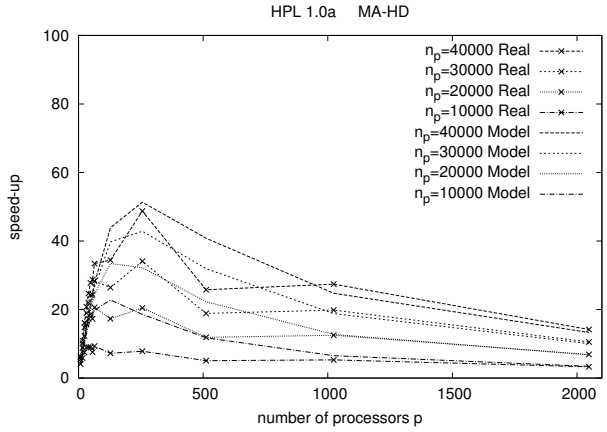

Fig. 9 Speed-up comparison of measurements and model for two regions.

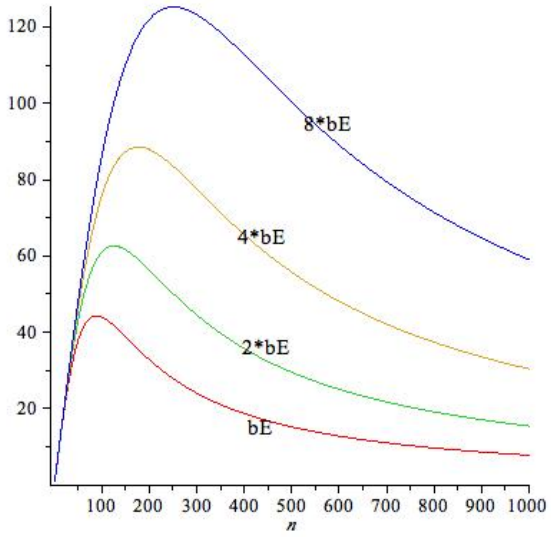

Fig. 10 Speed-up bwGRiD1 for two regions and varying bandwidth b^E.

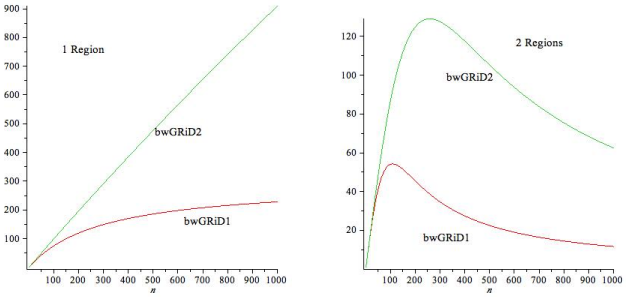

Fig. 11 Speed-up in bwGRiD1&2 for one and two regions with $n_p = 40000$.

we can realize our model for a single region $Sp = (1 + x'z)/(1 + x'z/n) = (1 + n_p z/136)/(1 + n_p z/(136n))$ (Fig. 8), which agrees very good with the empirical data, choosing $z = 1$ (higher values imply a simple scaling of n_p).

In the case of two interconnected regions we use Eqn. 17 and Eqn. 18 of Sect. 3.2 $Sp2 \leq x'rz'/(2p(n) + x'rz'/n)$. The value of the parameter x' remains, $z' = b^E/b^I = 2/3$ describes the influence of interconnection. Further we need $r = \#b/\#x$, with the pessimistic assumption $\#x = 0.5n_p^2$ it follows $r = 4$. The number $p(n)$ of communicating nodes will be a monotone function in n, we choose $p(n) = \alpha \cdot n$ with $0 < \alpha < 1$, which is a free parameter. With $\alpha = (1/30)...(1/20)$ we get a very good correspondence with the measured data (Fig. 9).

In order to estimate the influence of a higher bandwidth b^E, we fix $n_p = 40000$ and vary $z = 2/3...16/3$ (Fig. 10).

The doubling of maximal speed-up by changing the bandwidth is calculated by Eqn. 20. With $\alpha = (1/20)$, $n_p = 40000$, $z'_0 = 8/3$ and $x' = np/136$ follows $z'/z'_0 = 11/2$. That means we have to increase the interconnected bandwidth b^E by a factor of 5.5 to achieve our goal.

The same procedure will be needed to estimate the performance of the bwGRiD2, which will be implemented in the future – data are

- $l_{\text{th}} = 10 - 16$ [GFLOP/sec] per core
- $b^I = 6.0$ [GByte/sec], $b^I_0 = 1.0$ [GByte/sec]
- $b^E = 15$ [GByte/sec]

It follows $a^* = 15/6$, $x' = (n_p/60)$, $z = 15/6$ with $l_{\text{th}} = 15$ [GFLOP/sec]. We compare the cases of 1 and 2 regions in the bwGRiD1 and bwGRiD2 at $n_p = 40000$ (Fig. 11). As long as the ratio b^I/b^E remains constant, it is not important, if we take theoretical or real values for the bandwidths.

The calculation in one region only, shows for the bwGRiD2 an ideal linear speed-up with $n < 1500$, anyway remains the asymptotic behavior $Sp(n \to \infty) = \text{const}$. Running LinPack over two regions we get a factor of 3

in favour of bwGRiD2, due to the better values of the performance parameters l_{th}, b^I and b^E, the asymptotic behavior $Sp2(n \to \infty) = 0$ does not change.

4 Conclusions

The presented analysis of homogenous cluster systems without any stochastic tools via a modification of the roofline model shows some simple and insightful results.

We can describe throughput and speed-up by $2-3$ dimensionless scaling variables, which summarize all important hard- and software-oriented characteristics. In a qualitative context we can reproduce the empirical data (LinPack measurements) of the bwGRiD1 and can produce further a reliable performance prediction for the future bwGRiD2. Furthermore we are able to derive some bounds for the size of regions in a cluster, depending of the scaling variables of application (x') and bandwidth (z,z'). From special interest is the scaling of the maximal speed-up. For doubling its value at the fixed bandwidth z'_0 we find that the new bandwidth z' must fulfill the condition $z/z_0 \sim (2+W/\sqrt{W})^2$, with $W = 2 \times \alpha \times x'/z'_0$ (see section 3.3) – this will be an important parameter for the configuration of the region-region-interconnected network.

But a detailed analysis is still missing, which focuses on the original roofline model for the multicore-nodes and on inhomogeneous clusters with asymmetric load distribution. This will be considered in future investigations, including applications like Matrix-Matrix-Multiplication (MMM) or Fast-Fourier-Transform (FFT).

Acknowledgments

We thank our colleagues Rolf Bogus, Hermann Lauer and Steffen Hau as well as the bwGRiD team for the help in the construction and operation of the interesting hardware and the optimization of the connection, which is the basis for this paper. One of the authors, H.G. Kruse, thanks LBNL/Berkeley for the hospitality during a research visit. The inspiring and exciting atmosphere favored the becoming of this work.

bwGRiD is a Member of the German D-Grid initiative and is funded by the Ministry of Education and Research and the Ministry for Science, Research and Arts Baden-Württemberg.

References

1. bwGRiD, Member of the German D-Grid initiative, funded by the Ministry of Education and Research and the Ministry for Science, Research and Arts Baden-Württemberg, Universities of Baden-Württemberg, 2007-2010. URL http://www.bw-grid.de/, 2007-2010, accessed May 2010.
2. H. Kredel, H.-G. Kruse, S. Richling, Zur Leistung von verteilten, homogenen Clustern, PIK, Vol. 2, 2010, pp. 166–171.
3. S. Richling, S. Hau, H. Kredel, H.-G. Kruse, Operating Two InfiniBand Grid Clusters over 28 km Distance, Proc. 3PGCIC-2010, IEEE, 2010.
4. S. Richling, S. Hau, H. Kredel, H.-G. Kruse, Operating Two InfiniBand Grid Clusters over 28 km Distance, International Journal of Grid and Utility Computing, Vol. 2, No. 4, pp. 303 - 312, 2011.
5. S. Richling, S. Hau, H. Kredel, H.-G. Kruse, A Long-distance InfiniBand Interconnection between two Clusters in Production use, Proc. Supercomputing, November 12-18, 2011, IEEE, 2011.
6. M. Merz, M. Krietemeyer (eds.), IPACS integrated performance analysis of computer systems - benchmarks for distributed computer systems, Logos Verlag, Berlin, 2006.
7. HLRS, MPI benchmark of InfiniBand over fiber optics, http:// www.hlrs.de/, accessed May 2010.
8. InfiniBand, High performance network, http://de. wikipedia. org/ wiki/ InfiniBand or http://www.infinibandta. org/, accessed Jan 2012.
9. Obsidian, High performance network, http://www. obsidianresearch. com/, accessed May 2010.
10. LinPack und HPL, Linear Algebra Package and High Performance Linpack, http:// www.netlib.org/ benchmark/ hpl/, accessed Jan 2012.
11. H.G. Kruse, Leistungsbewertung bei Computer-Systemen, Springer, 2009.
12. L. L. Peterson, B. S. Davie, Computernetze, dpunkt, 2004.
13. U. Schlegel, K. Grobe, David Southwell, 100 Gbit/s DWDM InfiniBand Transport over up to 40 km, http://tnc2009. terena.org/core/getfile7b9a.pdf ?file_id=10, accessed May 2010.
14. A. Plaat, et. al., Sensitivity of parallel applications to large differences in bandwidth and latency in two-layer interconnects, Future Generation Computer Systems, 2001, pp. 769–782.
15. Weikuan Yu, Nageswara S.V. Rao, Jeffrey S. Vetter, Experimental Analysis of InfiniBand Transport Services on WAN, International Conference on Networking, Architecture, and Storage, 2008, pp. 233-240.
16. S. Carter, M. Minich, N. Rao, Experimental evaluation of InfiniBand transport over local- and wide-area networks, SpringSim '07: Proc. 2007 spring simulation multiconference, 2007, pp. 419–426.
17. H. Kredel, H.-G. Kruse, I. Ott, Lastverhalten und Systemkonfiguration von Web-Applikationsservern, Praxis d. Informationsverarbeitung und Kommunikation (PIK), De Gruyter Saur, Vol. 3, pp. 215-223, 2011.
18. H. Kredel, H.-G. Kruse, I. Ott, Performance analysis and performance modeling of Web-applications, Proc. 3PGCIC-2011, IEEE, pp. 115-122, 2011.
19. S. Williams, A. Waterman, D. Patterson, Roofline: an insightful visual performance model for multicore architectures, Commun. ACM, 52, Vol. 4, pp. 65–76, 2009.
20. M. D. Hill, M. R. Marty, Amdahl's Law in the Multicore Era, Computer J., Vol. 41/7, pp. 33–38, 2008.
21. A. Popa, What is the speed of light in a fiber optic cable? http://www.madsci.org/posts/archives/2001-02/983369337.Ph.r.html accessed Jan 2012.