Application and Evaluation of Optimal Configuration Estimation Scheme for Heterogeneous Clusters

Graduate Adviser : Shuichi Ichikawa 023719 Sho Takahashi

1 Background

Many parallel applications are targeted for clusters comprised of *homogeneous* processing elements (PEs). Since their performances are degraded by load imbalance on a *heterogeneous* cluster, it is necessary to distribute workloads considering the performance of each PE. It is a simple solution to invoke multiple processes on fast PEs (multiprocessing). Kishimoto and Ichikawa [1] constructed the execution-time estimation models from measurement results of HPL (High Performance Linpack), and showed that the (sub-)optimal configurations were actually estimated for multiprocessing. This study first examines Kishimoto's models on four applications, and then introduces a new model that is more accurate than Kishimoto's.

2 Execution-Time Estimation Model

2.1 Kishimoto's Models

Let N be the size of the problem. G_i is a group of PEs comprised of equivalent PEs in heterogeneous cluster. P_i is the number of PEs actually used for the job in G_i. M_i is the number of processes on each PE in G_i. P is the total number of processes in the cluster; i.e., $P = \Sigma_i P_i M_i$. T_i is the execution time of G_i, which is parameterized by N, P, and M_i. Total execution time T is estimated by $\max_i T_i$. The estimation function of T is designated by "execution-time estimation model" in the following discussion. Optimal configurations are estimated using the models of all possible configurations (P_i, M_i).

In case of HPL, T is given by Eq. (1), and thus T_i for $\exists(P_i, M_i)$ is represented by Eq. (2). Constant factors $k_0, ..., k_3$ are determined from the measurement results by the least squares method. This model is designated by *N-T* model [1].

It takes long time to construct N-T models, because they are constructed for all possible configurations (P_i, M_i). We can reduce the number of models by integrating N-T models into one new model that includes P as a parameter. Assuming that T_i is independent of the target of communication, this new model is given by Eq. (3), which is designated by *P-T* model. It takes shorter time to construct P-T models than N-T models, because P-T models are constructed from the measurements on G_is. Constant factors are extracted from the corresponding N-T models (PEs ≥ 2).

$$T(N,P) = \frac{1}{P} \cdot O(N^3) + P \cdot O(N^2) + O(N^2) \quad (1)$$

$$T(N)|_{P,Mi} = k_0 N^3 + k_1 N^2 + k_2 N + k_3 \quad (2)$$

$$T_i(N,P)|_{Mi} = \frac{k_0}{P} \cdot T_i(N)|_{P,Mi} + k_1 P \cdot T_i(N)|_{P,Mi} + k_2 \quad (3)$$

2.2 NP-T Model

Equation (1) is transformed to Eq. (4), using parameters N and P. This model is designated by *NP-T* model. An NP-T model includes more constat factors, and thus is expected to be more accurate than a P-T model. Since NP-T models can be constructed from the measurements on G_i, their construction time is the same as P-T models.

$$T_i(N,P)|_{Mi} = \frac{1}{P} \cdot (k_0 N^3 + k_1 N^2 + k_2 N + k_3) + P \cdot (k_4 N^2 + k_5 N + k_6) + k_7 N^2 + k_8 N + k_9 \quad (4)$$

3 Evaluation Methods

In this study, the following four benchmarks are examined on the heterogeneous cluster shown in Table 1. Table 2 summarizes the problem sizes (N) for measurement and evaluation. For each benchmark, N-T, P-T, and NP-T models are constructed and used to estimate the optimal configuration.

HimenoBMT measures the performance to solve Poisson's equation by Jacobi iteration for $N \times N \times N$ domain.

Hpcmw-solver-test is a benchmark for finite element method. $N \times N \times 1$ domain is examined here.

FFTE computes FFT of $N = 2^p 3^q 5^r$. In this study, N is fixed to 2^p. Since the process allocation is different when P contains a factor of 3 or 5, P-T and NP-T models for these cases are constructed separately.

HPL is a linear algebraic system benchmark. HPL is examined here to compare with Kishimoto's results.

Table 1: Evaluation environment

	G_1	G_2
PE	Xeon 2.8 GHz	Celeron M 1.5 GHz
OS	Redhat Linux 9	FedoraCore 3
Compiler, Library	gcc 3.2.2, ifc 8.1, mpich-1.2.6 (Buffer 8KB)	
P_i	$1 \leq P_1 \leq 8$	$0 \leq P_2 \leq 8$
M_i	$1 \leq M_1 \leq 2$	$0 \leq M_2 \leq 1$

Table 2: Measurement sizes (N)

	Measurement	Evaluation
HimenoBMT	32∼192 9 sets	32∼256 10 sets
hpcmw-solver-test	70 504 7 sets	70∼660 20 sets
FFTE	2^{12}∼2^{20} 9 sets	2^{16}∼2^{23} 8 sets
HPL	400∼6400 9 sets	1600∼9600 7 sets

4 Evaluation results

Figure 1 summarizes measured execution times of the estimated optimal configurations and the actual optimal configurations for various sizes.

For HPL and hpcmw-solver-test, (sub-)optimal configurations were estimated with NP-T models. Though N-T and P-T models also found (sub-)optimal configurations for interpolated N, their errors increased for extrapolated N, because parameter extraction fails for some cases.

For HimenoBMT, the estimation of P-T models and N-T models degraded at $N = 160$ and $N = 256$, respectively. NP-T models successfully estimated optimal or sub-optimal configurations for HimenoBMT.

For FFTE, the errors of N-T and P-T models become larger as N increases. NP-T models succeeded to estimate optimal or sub-optimal configurations.

In summary; Kishimoto's models degraded on some applications, while NP-T models succeeded to find better configuration for more applications.

In this study, a heterogeneous cluster with two kinds of processors was examined. The evaluations with more heterogeneous environment are left for future studies.

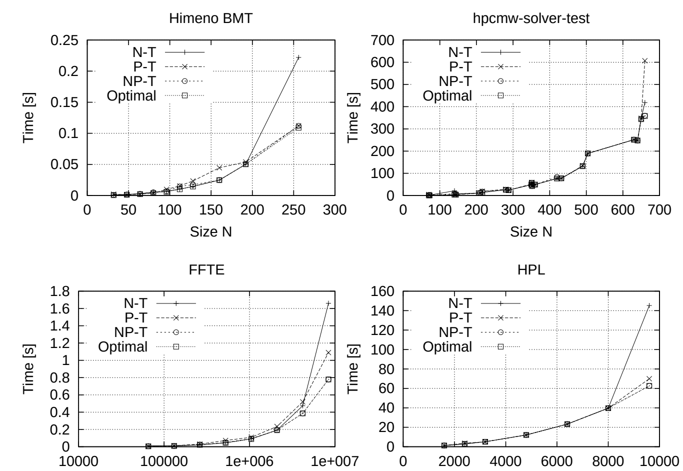

Figure 1: Evaluation results of four benchmarks

References

[1] Kishimoto, Y. and Ichikawa, S.: Optimizing the Configuration of a Heterogeneous Cluster with Multiprocessing and Execution-Time Estimation, *Parallel Computing*, Vol. 32, No. 7, pp. 691–710 (2005).