

Honda Grid System for Large-scale Heterogeneous PC-cluster Group

Fumito MAEKAWA*

Yoshihiro YAMAMOTO*

Hidemoto NAKADA**

ABSTRACT

In order to enhance the CAE computation environment, a Honda grid system for a large-scale heterogeneous PC-cluster group was developed. This system is necessary for increasing the operating rate of and effectively utilizing resources (CPUs), the number of which increases rapidly, along with the shift of the CAE computation environment from that of a supercomputer containing a high performance CPU to a PC cluster consisting of a large number of PCs installed on a rack.

By carrying out tuning that matches the environment based on a high quality scheduler while developing unique functions, the operating rate of the CPU was able to be increased. This, together with the effect of reinforcing the resources, enabled the mean computation time, including the waiting time, to be reduced to approximately one-tenth, regardless of the increase in the job number and job scale.

Also, using the same method, a CAD desktop PC could be effectively utilized, and the CAE computation environment on the PC was also enhanced.

1. Introduction

Up until several years ago, super computers that use a few powerful CPUs to perform high-speed calculations were predominately used to carry out CAE calculations. However, such systems are expensive and demand a large investment in order to expand the system scale.

In contrast, given that PC performance increases constantly over time, investment costs can be kept down, while securing performance approaching that of a super computer by introducing a PC cluster formed by clustering several PCs together.

However, to maintain the same level of performance, a PC cluster with 64 CPUs versus eight super computer CPUs is needed, resulting in an increase in system scale, from several dozen CPUs originally to several hundred CPUs.

In addition, as a result of heightened expectations regarding increased accuracy in CAE analyses and shorter development cycles, the amount of model data increases.

Further additions to the PC cluster in order to compute this data in an equivalent amount of time increases the number of CPUs needed, approaching 10 000 CPUs (core).

Clearly, system operations will break down in a situation like this, given the job management and operation techniques performed manually up to this point. This leaves no choice but to quickly automate operations and increase efficiency.

Honda has applied grid computing⁽¹⁾ technology as a means of addressing these issues. The system functions so that if there is a calculation request, the calculation is performed using unspecified resources connected somewhere on the in-house LAN and the result is then returned (Fig. 1).

This paper describes the technology and research topics related to the Honda Grid System (Fig. 2), which has succeeded in reducing the man hours associated with operation and management, while achieving the highly efficient operation of several thousand CPUs.

* Automobile R&D Center

** National Institute of Advanced Industrial Science and Technology

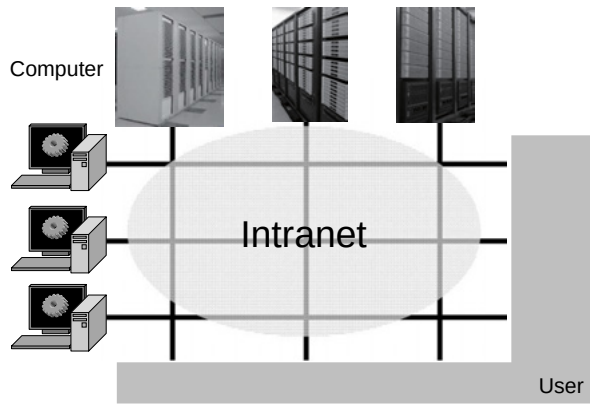


Fig. 1 Honda grid system

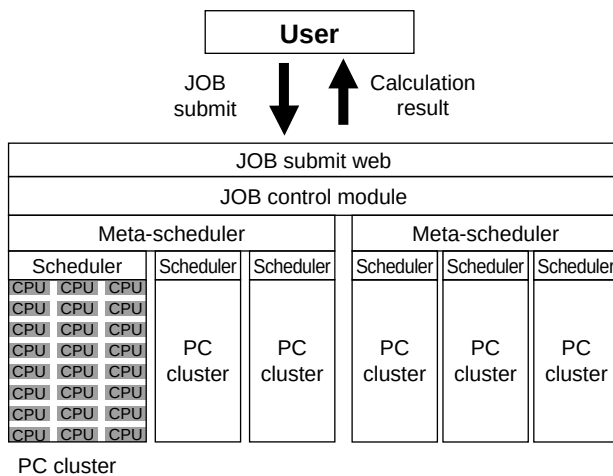


Fig. 2 Outline of Honda grid system

2. System Goals

The goal of developing the Honda Grid System is to take a PC cluster group consisting of PCs from different manufacturers that have different levels of performance, installed at different times, and run that PC cluster group efficiently, to reduce job calculation times without increasing the hours of operation.

The following goals were set for standard calculation of a large-scale analysis performed in-house.

- (1) A CPU availability rate of 85% or more
- (2) A calculation time, including wait times, of 7 hours or less
- (3) Reduction in management and operation man hours
- (4) Effective utilization of PCs for CAD

3. System Configuration and Implemented Functions

3.1. Job Scheduling Function

In building a grid system in a large-scale PC cluster environment, the most important function is job

scheduling. This function distributes jobs submitted to the system to each CPU to be used for calculation based on the status of CPUs being managed and policy of job processing. The software used to fulfill this role is called a “scheduler.”

Since the selection of a scheduler is the key to success even in building the Honda Grid System, the items to be evaluated were chosen and, of these, the following requisite features were established for the Honda grid. The system should:

- (1) Support both Linux and Windows
- (2) Be strong at (or capable of being made strong at) performing parallel calculations
- (3) Have a good performance record and offer stable operations
- (4) Keep costs down even with an increased number of CPUs

As a result of thorough investigation, Condor⁽²⁾ was selected.

3.2. Check Point Restart Function

This function temporarily saves information in the middle of calculations (check point), so that calculations can be restarted later based on this intermediate information (Fig. 3).

Since the number of CPUs and number of jobs are both low in a conventional computing environment, processing can be performed according to a simple first-in, first-out method where only the number of tasks being executed were limited for each user.

However, as the number of jobs increases along with the increase in CPUs, waste begins to arise in both job processing and CPU utilization conditions when using simple processing. To help eliminate this waste and achieve efficient job processing and CPU utilization, a Check Point Restart function was introduced. If a CPU is available, the restriction on the number of jobs is canceled and jobs that are accumulating are processed. If a high-priority job is submitted to the system after this, a CPU can be made available by taking the check point of an executing job and temporarily suspending it. The high-priority job is then processed. When other calculations are completed and a CPU becomes available, the suspended job is restarted from its check point.

The challenging task involves determining the job to temporarily suspend. Suspending a job once processing has started may give rise to feelings of distrust among users. Honda therefore considered the issue of execution priority to determine the order jobs are processed from a reverse perspective, in which an evacuation priority is determined for temporarily suspending jobs. By increasing the evacuation priority of jobs with later processing start times (lower processing priority), jobs that start processing later can be temporarily suspended as necessary.

A processing algorithm has been realized based on this thinking and, as a result of PR aimed at users, Honda was able to introduce the algorithm as a function that did not

give rise to distrust. In addition, because the implementation of this system differs for each application, a processing method suitable for each application was established.

By implementing this function, overall job processing performance was increased by 16%.

3.3. Meta-scheduler Function

Because Honda has used multiple calculation servers (PC clusters) at the individual department and application level up to this point, there was a tendency for server loads to fluctuate in accordance with the task load. Although differences in the load among servers would

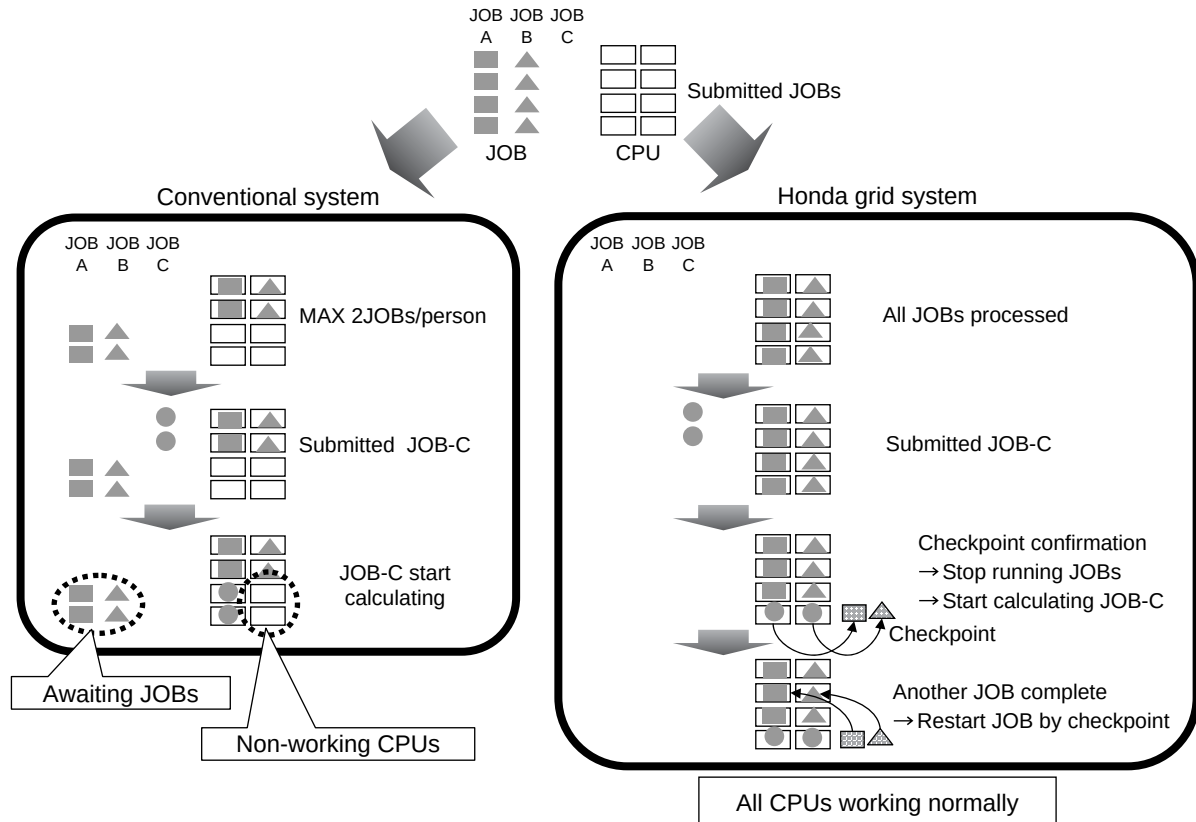


Fig. 3 Checkpoint and restart

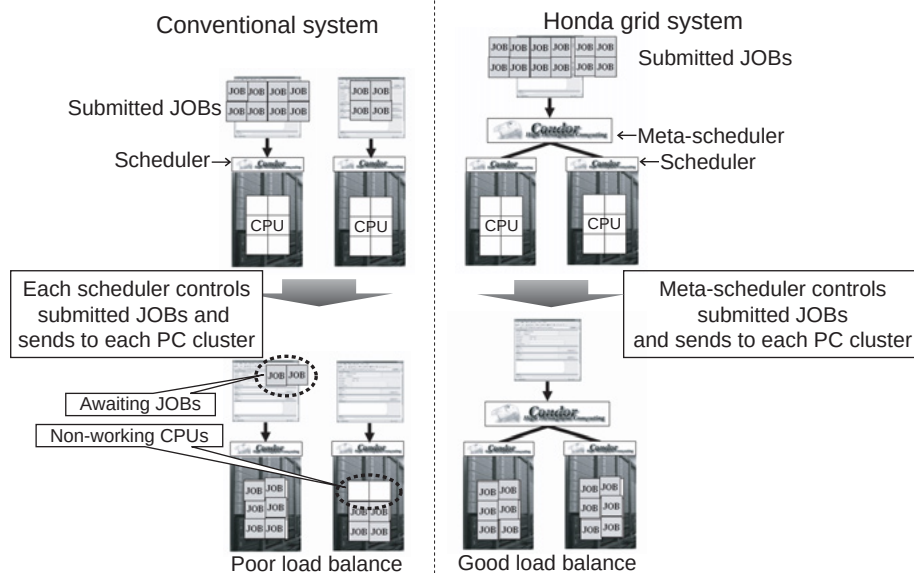


Fig. 4 Meta-scheduling

therefore arise, resulting in inefficient operations, it was almost infeasible to manually perform the task of assigning jobs in order to settle these differences.

A meta-scheduler function that manages multiple servers at once and distributes jobs appropriately has been developed in order to reconcile these differences automatically (to achieve load balance) (Fig. 4).

The flow of this function is as follows:

- (1) When a job is submitted, the meta-scheduler (parent) exchanges information with the scheduler (child) for each server under it, and the job is then passed to an appropriate scheduler based on that server's load information.
- (2) The scheduler to which the job is passed distributes the job to available CPUs underneath it and calculations begin.
- (3) When calculations have ended, the child scheduler informs the meta-scheduler that the job has ended. Finally, the job is removed from the pool.

This function allows operators to balance the load among servers and job execution performance of the overall CAE calculation environment is increased as a result.

3.4. Simple Job Submission Function

Since the method of submitting jobs to the system and the Message Passing Interface (MPI) software used for parallel calculations differ for each application being used, creating a job submission script is complicated, and mistakes in scripts arise due to users creating the scripts themselves, resulting in job issues.

By putting job submission that calls for these man hours on the Web, mistakes in scripts are minimized by automatically creating job submission scripts, thus reducing both the user man hours for submitting jobs and job calculation issues.

Furthermore, Honda determined a standard policy

regarding the order in which jobs submitted to the system are processed and developed a program for assigning priority in line with this policy. This allowed Honda to provide equal service to all users by counteracting unequal conditions among users regarding job processing.

3.5. Automatic System Monitoring Function

In the Honda grid system, where massive parallel calculations are processed in a large-scale environment, the early detection of issues is the most important factor related to stable operation. However, once a system reaches a scale of this size, monitoring by human operators as in the past is no longer feasible. New means of monitoring therefore need to be constructed, even for early recovery, when an issue is detected.

The following measures have therefore been taken with this system.

- (1) A 24-hour monitoring system is configured by launching a tool for constant automatic monitoring of the grid system. Under this system, an e-mail is sent to an administrator if an issue is detected.
- (2) Issues can be investigated and handled smoothly when they occur because grid system monitoring tasks have been split up and divided between the overall system department and grid-related departments.
- (3) Hotlines have been established with hardware vendors to achieve early resolution of hardware-related issues. With specialists in place on the vendor's side, issues can be investigated and handled smoothly.

These measures serve to reduce the man hours and time needed to recover from issues.

3.6. Benefits of the Honda Grid System

Figure 5 shows the benefits of this system configuration.

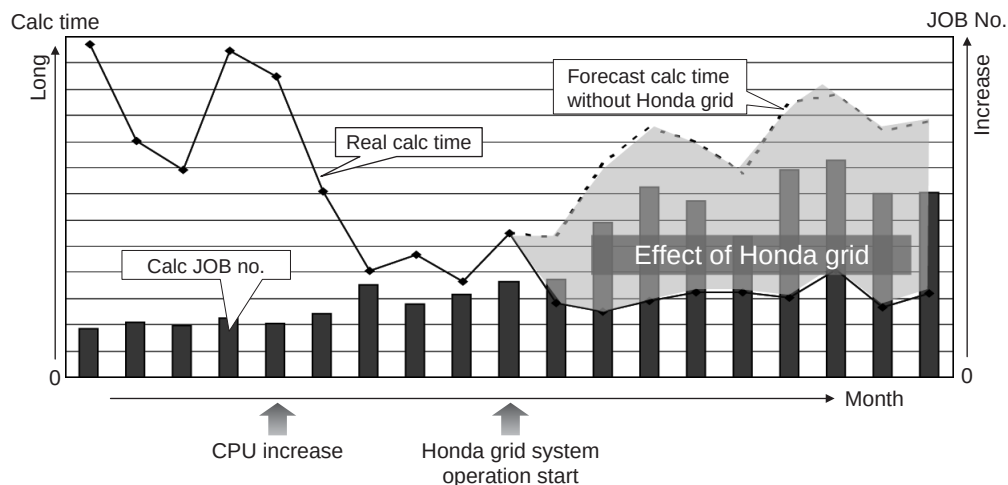


Fig. 5 Result of Honda grid system

4. Grid Application Development on the Desktop PC Group

Although many various CAE applications are being used under Windows today, prompt, responsive performance is reduced to the point that other tasks cannot be executed during calculations. However, even though one frequently sees cases of PCs going unused due to a supervisor being out of his or her seat for a meeting or for some other reason, it is not easy to assign calculations to these PCs. In addition, although one might like to perform calculations continuously throughout the night when no one is using the system, the environment is not set up to do so.

To meet these demands, a grid application using Condor, mentioned previously, has been developed for use on a desktop PC group (Fig. 6).

The issue at this point is what criteria to use in determining that a PC is in idle status. In other words, how should the system determine whether someone is merely out of their seat for a short time or that a PC is not being used for an extended period due to a meeting or other such reason?

The system has been designed so that machines

- (1) whose mouse and/or keyboard has not been touched for a set period of time, and
- (2) whose CPU has not been running for a set period of time, are assumed to be idle PCs to which jobs can be assigned.

The system is therefore designed so that if a machine

being used to process a job is used by a user, calculations are immediately stopped and restarted on another idle machine.

In addition, since users do not use their PCs in the middle of the night, the system has been designed so that the number of PCs needed to process submitted jobs is estimated, and only those PCs needed are run during the night to perform calculations.

These specifications were used to tune an effective and reliable environment for job processing without placing a load on PC users.

Furthermore, reduction in the man hours for submitting jobs to one tenth was achieved by allowing jobs to be submitted over the Web.

5. Conclusion

Having developed several functions for effectively running a large-scale PC cluster environment, Honda has built the Honda Grid System. The benefits of this system are listed below.

- (1) The CPU availability ratio has been increased from 50% to 85% and computing performance has been increased 1.7 times.
- (2) Along with this CPU increase, calculation time (including wait times) has been reduced to approximately one-tenth.
- (3) By forming desktop PCs into a grid, Honda has successfully realized a CAE calculation service under a Windows environment.

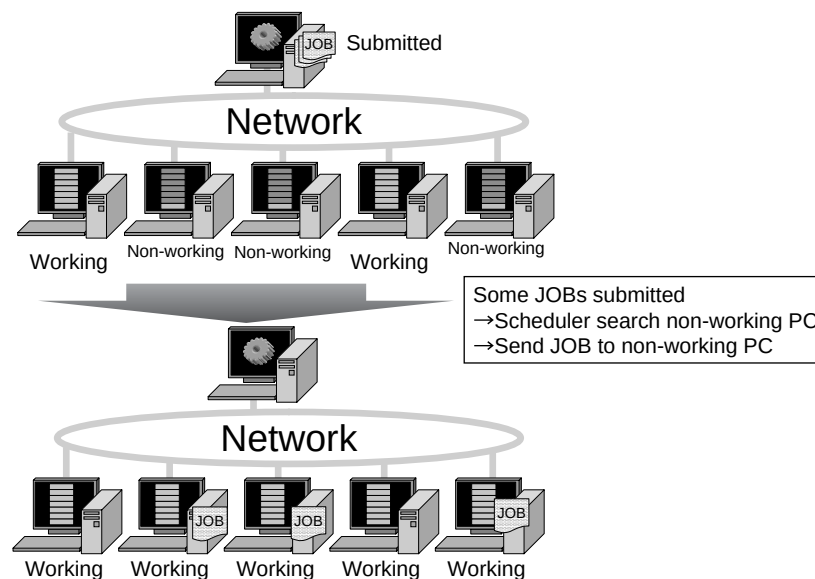
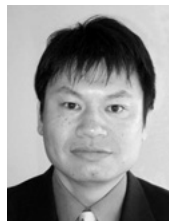


Fig. 6 Honda grid system on CAD PC

References

- (1) <http://www.jpgrid.org>, 2008/5/20
- (2) <http://www.cs.wisc.edu/condor/>, 2008/4/1

■ Author ■



Fumito MAEKAWA



Yoshihiro YAMAMOTO



Hidemoto NAKADA