CSC532 Term Paper

The Modeling and Dependability Analysis of High Availability OSCAR Cluster System

Lixin Shen

Louisiana Tech University
eXtreme Computing Research Group
The Modeling and Dependability Analysis of High Availability OSCAR Cluster System

Lixin Shen

Abstract
OSCAR is widely used for building and maintaining a high-performance parallel computing system. In many cases, high availability requirement becomes as critical as high performance. In this paper, the current OSCAR cluster system is introduced. Some high availability consideration is discussed and the high availability OSCAR cluster system is presented. Continuous Time Markov Chain models are built and applied to these two systems. On the base of the models, the dependability of the two systems is analyzed and compared with. The results show that the high availability consideration is reasonable and it will highly improve the availability of the current OSCAR system.

Keywords: high availability, dependability, Continuous Time Markov Chain, OSCAR, cluster

1. Introduction
Open Source Cluster Application Resources (OSCAR) is a fully integrated, easy to install software packages designed to make it easy to build and use a cluster for high performance [1]. It has been widely used for building and maintaining a Beowulf cluster, which has become the fastest growing choice for building high-performance parallel computing systems [2]. In many cases, high availability (HA) requirement becomes as critical as high performance. If the future OSCAR is applied to telecom, ISP or Web industry, some high availability features must be involved in [3].

From the point of view of high availability, the current OSCAR release is fragile. Most of the system components present a single point of failure. Their failure rates will dominate the availability of the whole system. In order to fulfill the reliability requirement of a system, a too high reliability requirement will be applied to some key components. The overall price/performance rate of the system will raise and the main advantages of the Beowulf cluster system will be reduced. One solution falls within building redundancy in the system. Duplicating key components will highly improve the availability of the components and so does that of the whole system.

Dependability analysis of a system will show if the requirement will be meet from the design stage. It provides a good mechanism for examining the behavior of these systems. Continuous Time Markov Chains (CTMC) are a useful modeling formalism for dependability analysis of computer systems. They can easily handle many of the interdependencies and dynamic relationships among the system components [4].
The rest of this paper is organized as follows. In the next section, the architectures of current OSCAR cluster system and the HA OSCAR cluster system are introduced individually. In section 3, CTMC models corresponding to the two systems are built, and their dependability is analyzed in section 4. In section 5, a conclusion is provided.

2. System Characteristics

Before diving into the dependability analysis, system characteristic will be investigated. It will beneficial to give a description of the general layout of the system.

2.1 Current OSCAR Cluster Architecture

Figure 1 shows the architecture of an OSCAR cluster system supported by the current release. Each individual machine of a cluster is referred to as a node. Within the OSCAR cluster, there are two types of nodes, server and client. A server node is responsible for servicing the requests of client nodes. A client node is dedicated to computation. The OSCAR cluster will consist of one server node and a number of client nodes, where all the clients contain homogeneous hardware. The server node, along with serving as the gateway to the external network, contains the home directories of all users and runs the PBS server and scheduler. The clients each have a local copy of the operating system and other software, with the exception of NFS mounting the users' home directories from the server. All the nodes are connected by an internal Ethernet network, preferably through a dedicated switch [2].

![Fig. 1 An Architecture supported by the current OSCAR release](image)

2.2 HA OSCAR Cluster Consideration

The Open Cluster Group continues update the package and make sure it represents a current snapshot of best-known-practices for building and using cluster [1]. It shows that OSCAR cluster systems can be applied potentially in some fields, such as Telecommunication, Network or Web industry, in which high availability features are required.
Hardware duplication and network redundancy are common techniques for improving the reliability and availability of computer systems. For a HA OSACAR cluster system, the first thing that we can suggest is to duplicate the server of the current OSCAR cluster. The two servers drive the installation and setup of all other nodes. There are some choices for implementing the architecture, including the Active-Active, Active-Warm Standby and Active-Cold Standby. Currently the Active-Active configuration is the model of choice. The dual master nodes would be running redundant DHCP, NTP, TFTP, NFS and SNMP servers. If one master node dies for some reason, the other node will replace its functionalities.

Another suggestion that falls within building redundancy in OSCAR cluster system is to have high network availability that can be achieved by having two Ethernet ports on every node of the system. This allows every node to be present on two networks. Backed with Ethernet redundancy, we can provide high network availability.

Figure 2 shows the HA OSCAR cluster system architecture. Each of the duplicated servers connects to external network by two different lines. It will keep the system connecting to the external environment in case of one of the network system is down. Inside the system, each server connects to two switches separately. And each client node connects to each of the switches.
3. System Model

The modeling methodology chosen for this evaluation was a state-space model. Assumptions were made to reduce the size of the state space. Our analysis will focuses on servers and switches that effect cluster availability. We assume that the time to failure for each component is exponential distributed, with the parameters being $\lambda_{SV}$ for the servers and $\lambda_{SW}$ for the switches respectively. Furthermore, we assume that failed components can be repaired. Suppose the time to repair a server and the time to repair a switch are exponentially distributed with parameters $\mu$ and $\beta$. We also assume that whenever the system is down, no further failure can take place. Hence, for the current OSCAR cluster, when the server is down, the switch can not fail. Similarly when the switch is down, the server does not fail. For the HA OSCAR cluster, when both servers are down, the switches can not fail. Similarly when both switches are down, the servers do not fail.

3.1 Current OSCAR Cluster System Model

The CTMC model corresponding to the current OSCAR cluster system is shown in Figure 3. In state 1, both server node and switch are functioning properly. In state 2, the server node has a failure and in state 3 the switch has a failure. The system will be available for service in state 1, and will be unavailable for state 2 and state 3. The system goes from state 1 to state 2 when server failure occurs at rate $\lambda_{SV}$, and from state 1 and to state 3 when switch failure occurs at rate $\lambda_{SW}$. After server recovery at rate $\mu$, the system is back in state 1 from state 2. And, after switch recovery at rate $\beta$, the system is back in state 1 from state 3.

![Fig. 3 CTMC for current OSCAR cluster system](image)

3.2 HA OSCAR Cluster System Model

The CTMC model corresponding to the HA OSCAR cluster system is shown in Figure 4. The states and their corresponding system status are shown in Table 1. The system will be available for service in state 1, 2, 4 and 5, and will be unavailable for state 2, 6, 7 and
8. The system goes from one state to another at the rates mentioned in the corresponding arrow lines in Figure 4.

![Fig. 4 CTMC for HA OSCAR cluster system](image)

Table 1 System status

<table>
<thead>
<tr>
<th>State Number</th>
<th>Server Up</th>
<th>Switch Up</th>
<th>System Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>Up</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>Up</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>Down</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>Up</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>Up</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>Down</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>Down</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>Down</td>
</tr>
</tbody>
</table>

4. Availability Analysis

Let $\pi_i$ be the steady-state probability of state $i$ of the CTMC. They will satisfy the following equations:

$$\pi Q = 0$$

and

$$\sum_{i \in U} \pi_i = 1$$

where $Q$ is the infinitesimal generator matrix [5]. Let $U$ be the set of up states, the availability of the system $A$ is
\[ A = \sum \pi_k \]

### 4.1 Current OSCAR Cluster System Analysis

We compute the steady-state probabilities by balance equations:

\[ \pi_1 (\lambda_{SV} + \lambda_{SW}) = \pi_2 \mu + \pi_3 \beta \]
\[ \pi_2 \mu = \pi_1 \lambda_{SV} \]
\[ \pi_3 \beta = \pi_1 \lambda_{SW} \]

Now, since

\[ \sum_{i=1}^{1} \pi_i = 1, \]

we have

\[ \pi_1 = \frac{1}{E} \]
\[ \pi_2 = \frac{\lambda_{SV}}{E} \]
\[ \pi_3 = \frac{\lambda_{SW}}{E} \]

where

\[ E = 1 + \frac{\lambda_{SV}}{\mu} + \frac{\lambda_{SW}}{\beta}. \]

Thus, the steady-state availability \( A \) is given by

\[ A = \pi_1 = \frac{1}{E} \quad (1) \]

### 4.2 HA OSCAR Cluster System Analysis

We compute the steady-state probabilities by balance equations:

\[ \pi_1 (2\lambda_{SV} + 2\lambda_{SW}) = \pi_2 \mu + \pi_4 \beta \]
\[ \pi_2 (\lambda_{SV} + 2\lambda_{SW} + \mu) = \pi_1 \cdot 2\lambda_{SV} + \pi_3 \mu + \pi_5 \beta \]
\[ \pi_3 \mu = \pi_2 \lambda_{SV} \]
\[ \pi_4 (\beta + 2\lambda_{SV} + \lambda_{SW}) = \pi_1 \cdot 2\lambda_{SW} + \pi_5 \mu + \pi_6 \beta \]
\[ \pi_5 (\mu + \beta + \lambda_{SV} + \lambda_{SW}) = \pi_2 \cdot 2\lambda_{SW} + \pi_4 \cdot 2\lambda_{SV} + \pi_6 \mu + \pi_7 \beta \]
\[ \pi_6 \mu = \pi_5 \lambda_{SV} \]
\[ \pi_7 \beta = \pi_4 \lambda_{SW} \]
\[ \pi_8 \beta = \pi_5 \lambda_{SW} \]

Now, since

\[ \sum_{i=1}^{8} \pi_i = 1 \]

we have
\[ \pi_1 = \frac{1}{E} \]
\[ \pi_2 = \frac{2\lambda_{SV} / \mu}{E} \]
\[ \pi_3 = \frac{2\lambda_{SV}^2 / \mu^2}{E} \]
\[ \pi_4 = \frac{2\lambda_{SW} / \beta}{E} \]
\[ \pi_5 = \frac{4\lambda_{SV} \lambda_{SW} / \mu \beta}{E} \]
\[ \pi_6 = \frac{2\lambda_{SV}^2 \lambda_{SW} / \mu^2 \beta}{E} \]
\[ \pi_7 = \frac{2\lambda_{SW}^2 / \beta^2}{E} \]
\[ \pi_8 = \frac{4\lambda_{SV} \lambda_{SW}^2 / \mu \beta^2}{E} \]

where

\[ E = 1 + \frac{2\lambda_{SV}}{\mu} + \frac{2\lambda_{SW}}{\beta} + \frac{2\lambda_{SV}^2}{\mu^2} + \frac{4\lambda_{SV} \lambda_{SW}}{\mu \beta} + \frac{4\lambda_{SV}^2 \lambda_{SW}}{\mu^2 \beta} + \frac{4\lambda_{SV} \lambda_{SW}^2}{\mu \beta^2} \]

Thus, the steady-state availability \( A \) is given by

\[ A = \pi_1 + \pi_2 + \pi_4 + \pi_5 \]

\[ A = 1 + \frac{2\lambda_{SV}}{\mu} + \frac{2\lambda_{SW}}{\beta} + \frac{4\lambda_{SV} \lambda_{SW}}{\mu \beta} \]

\[ \frac{E}{E} \]

(2)

4.3 Comparison and Example

For an example system, we assume that \( \lambda_{SV} = 0.0001 \text{ hr}^{-1}, \lambda_{SW} = 0.00005 \text{ hr}^{-1}, \mu = 0.5 \text{ hr}^{-1}, \) and \( \beta = 1.0 \text{ hr}^{-1}. \) By formula 1 and formula 2, we can calculate the availability of the system. The availability for the current OSCAR cluster is 0.99975, and the availability for the HA OSCAR cluster is 0.9999999. The down time of the two systems in a year is 131 minutes and 0.05 minute (3 seconds) respectively.

5. Conclusions

By the analysis and comparison of the dependability by the current OSCAR cluster system and the HA OSCAR cluster system, we conclude that the availability of the HA OSCAR cluster system is higher than that of the current OSCAR cluster system. The high availability consideration is reasonable.
References
3 Ibrahim Haddad, Frederic Rossi, Chokchai (Box) Leangsuksun, Stephen Scott, Telecom/High Availability OSCAR Suggestions for the 2nd Generation OSCAR