高可靠、强实时分布式系统的切换技术研究*

胡华平　金士尧　李宏亮

(国防科技大学计算机系　长沙　410073)

摘　要　本文首先给出某高可靠、强实时分布式系统及系统的两种切换方案的简要描述, 然后对切换技术的实现进行讨论; 最后, 在建立强实时双工系统的可靠性模型基础上, 对系统的两种切换方案进行分析与比较。研究结果表明: 对于本系统, 机组级切换优于结点机级切换, 且可以减少程序的复杂度, 提高系统的可靠性与实时性。

关键词　分布式系统, 实时系统, 可靠性, 切换技术, 时钟同步。

分类号　TP. 393

The Switch Technology Study of High Reliability, Hard Real-time Distributed System

Hu Huaping　Jin Shiyao　Li Hongliang

(Department of Computer, NUDT, Changsha, 410073)

Abstract　In this paper, a high reliability, a hard real-time distributed system and its two switch schemes are introduced at first. Then, the realization of switch technology is discussed. Finally, based on the reliability model of hard real-time dual working system, the two switch schemes are analyzed and compared. The conclusions show: The switch in cluster is superior to the switch in node, and it can not only improve the system reliability and real-time, but also reduce the complexity of the program.

Key words　Distributed systems, Real-time system, Reliability, Switch technology, Time synchronization.

实时系统是计算机应用领域中的一个重要分支, 它广泛应用于过程控制、工厂自动化、机器人系统、武器系统、等领域; 而且许多实时系统都是在高可靠性要求下进行工作, 任何不可靠因素和计算机的一个微小故障都可能导致难以预测的灾难性后果, 因此人们已把高可靠性作为衡量实时系统性能不可缺少的重要指标。同时, 要在分布式系统中引入实时特性, 还存在许多困难, 如网络的实时性、分布式系统的调度问题, 以及如何对系统进行可靠性设计与分析[1,2]。

对于强实时的分布式系统来说, 其中的互联通信时延和开销至关重要。通过对分布式系统互联时延的分析与研究, 可以得到下列结论[3]: 分布式系统互联时延是由互联网络软硬件时延、处理机收发开销、应用进程同步延时等部分组成; 在强实时分布式系统中, 由于时限为 ms 级, 因此多数应该采用紧密耦合的互联网络, 以使互联网络的延时最小, 而要进一步减少延时, 将主要取决于结点机的收发延时。

1　系统描述

本文所给出的分布式实时系统是一高可靠(系统的稳态可用度大于 0.9999) 与强实时(服务处理时间为毫秒级) 的系统, 整个系统可分成对称的两个部分, 每部分由输入输出结点机(IO)、计算结点机(C)、管理结点机(M) 等构成, 从而形成两部分的双工。多机之间的互联是通过专用的内部互联网络和网络集中器(Switch HUB) 实现的。系统内部通过专用的内部互联网络紧密耦合在一起, 以完成强实时微秒级的通信。系统外围通过双套快速以太网和网络集中器(Switch HUB) 互联的, 以完成弱实时通信。

* 国家部委预研基金的资助项目:
1999 年 2 月 5 日收稿
第一作者: 胡华平, 男, 1967 年生, 博士后, 副研究员

由于强实时部分是本系统设计的关键, 故本文主要针对强实时部分进行讨论。对于强实时部分, 有两种方案可供选择: 方案 1(结点机级切换) , 其体系结构如图 1 所示, 其信息流程如图 3 所示; 方案 2(机组级切换) , 其体系结构如图 2 所示, 其信息流程如图 4 所示。

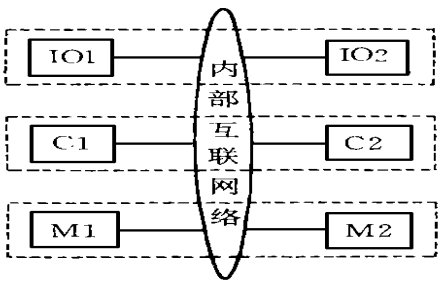

图 1 结点机级切换方案

Fig. 1 Switch scheme 1

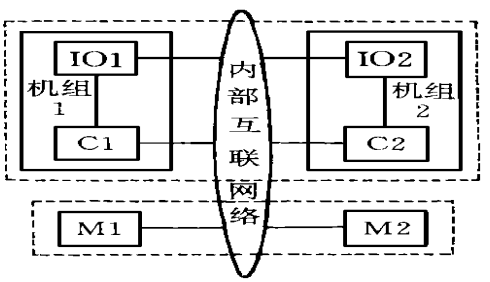

图 2 机组级切换方案

Fig. 2 Switeh scheme 2

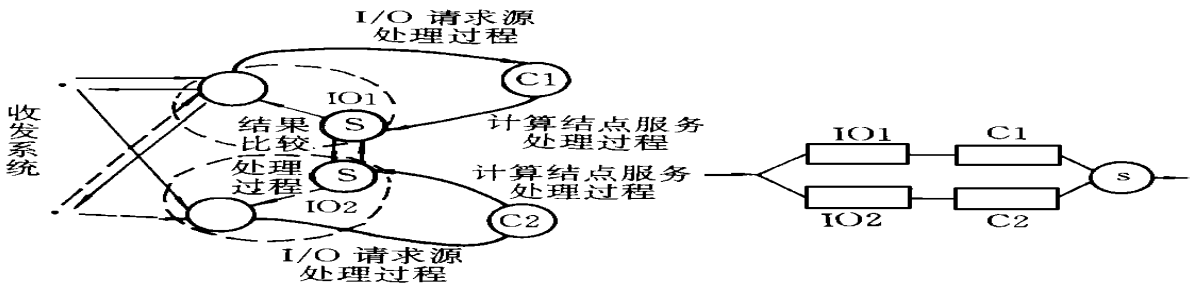

图 3 结点机级切换的信息流程

Fig. 3 Information process 1

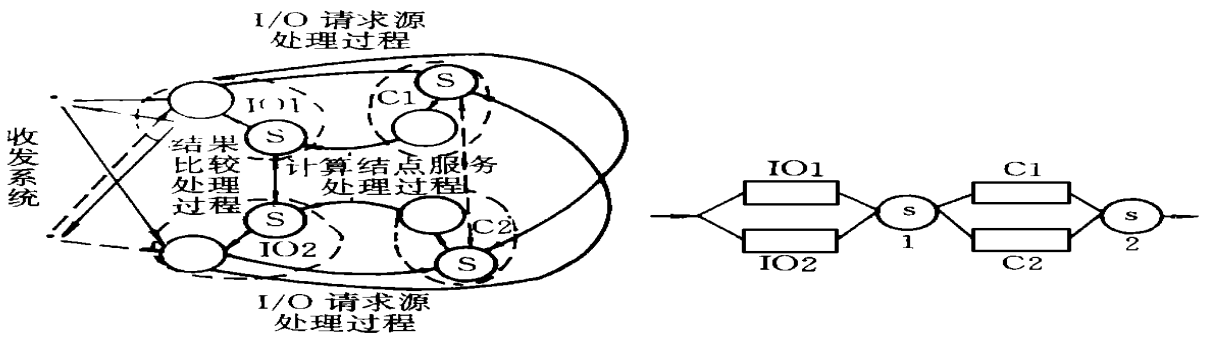

图 4 机组级切换的信息流程

Fig. 4 Information process 2

2 实现切换技术的关键点

研究在高可靠、强实时环境中实现双机(机组) 间的切换, 涉及到故障检测、故障恢复、时钟同步等关键技术。

2.1 结点机(机组) 的工作方式

(1) 方案 1

双工的结点机均可独立完成事务处理, 结点机间通过内部互联网络进行通讯。在正常情况下, 两结点机的处理结果需进行比较, 相同时方可将结果输出, 并送管理结点机备案, 或通过内部互联网络传送到下一双工的结点机上进行有关的处理; 若结果不相同, 则需进行有限复执, 以确定故障类型(偶然故障、永久故障或其它) , 并最终确定该双工结点机的工作状态(双工、单工、故障) 。

(2) 方案 2

双工的机组均可独立完成事物处理,机组间通过内部互联网络进行通讯。在正常情况下,两机组的处理结果需进行比较,相同时方可将结果输出,并送管理结点机备案;若结果不相同,则需进行有限复执,以确定故障类型(偶然故障、永久故障或其它),并最终确定该双工机组的工作状态(双工、单工、故障)。

2.2 故障检测

由于本系统是一高可靠、强实时的分布式系统,故障的出现不仅会延长事物处理的时间,而且还可能传播,影响其它相关的事物处理。为此,及时、准确地检测并确定故障是进行双机切换的关键。下面给出在本系统中所采用的两种故障检测方法。

(1) 在通信区设立结点机状态表

在通信区设立结点机状态表,每个状态表包括硬件状态字、软件状态字以及一个计数器。结点机故障与否是通过计数器来反映的。通过定期读取计数器的值,就可检测出结点机大部分故障。

(2) 建立质量评估报告

对每一个事务处理均产生一个质量评估报告,记录该处理过程中的信息(如所采用的计算方法、精度、自评等级等)。当该事务处理完毕时,两机组进行结果比较,若不相同,则进行有限次的复执,以避免偶然故障;此时,若结果仍不相同,则由驻留在IO结点机上的双工管理软件根据质量评估报告来确定最后的输出结果。

2.3 故障恢复

(1) 故障结点机(机组)的切出

当系统中某一结点机出现故障时,可将该结点机(或该结点机所在的机组)切出,进行故障维修。切出故障结点机(机组)时,系统的管理软件应修改结点机的状态表,此时,工作状态由双工变为单工。

(2) 故障结点机修复后的切入

由于另一结点机仍在正常工作,因此需设立缓冲区,并暂时停顿正常结点机的工作,以便复制现场。其主要技术点有:①在IO结点机上设立双缓冲(工作缓冲与后备缓冲)。当IO结点机接到切入信号时,即将事物输入指向后备缓冲区,而让工作缓冲区流空。流空后再进行现场复制,并最终交换两缓冲区,完成由单工到双工的转换;② 将需要复制的现场数据分类成块放置;③ 与应用程序尽量相互独立。

2.4 时钟同步

在强实时分布式系统中,为了满足系统实时处理事务的需求,各个处理结点之间的时钟误差要求为ms级。由于结点机时钟的初始设置通常有一定时差,而且时钟本身还会漂移,因此,系统中的各个时钟的误差将越来越大,从而造成系统不能正确、及时处理有关事务,故系统中各个结点机之间的时间必须保持高度的一致,并与标准时钟同步。在许多系统中,提供了网络时间协议软件,用来对系统中的结点机进行时钟同步。但是,这些网络时间协议软件适合于广域网、时间同步精度要求不是很苛刻的分布式环境,对于强实时分布式系统是不太适合的。

本文采用时钟服务器双机冗余的方法以满足强实时分布式系统对时钟同步的高精度和高可用的要求。具体实现方法为:时钟服务器与标准时钟相连接,接收标准时间信息,并调整本结点机的时间信息,使之与标准时钟同步。各个时钟客户端(结点机)通过内部互联网络与时钟服务器连接,它们之间采用TCP/IP协议进行通讯,以达到全系统时钟与标准时钟同步的目的。为了防止时钟服务器崩溃,在本系统中采用了双机冗余的方法以提高时钟服务器的可靠性,即采用了主辅两个时钟服务器,分别从不同的标准时钟接收时间信息。在通常情况下,时钟客户端与主时钟服务器的时间同步,当主服务器崩溃时,时钟客户端自动与辅时钟客户端的时间同步,辅时钟服务器转换为主时钟服务器。

3 切换方案的比较

定义1 可用度改进因子AIF(Availability Improvement Factor—AIF)[4]

定义可用度改进因子为

$$AIF = \frac{A_{方案1} - A_{方案2}}{1 - A_{方案1}} \times 100\% \qquad (1)$$

3.1 强实时双工系统的可靠性模型

由文献[5]知: 强实时双机系统不能容忍过长的故障诊断时间, 它存在着一个如何判别故障的问题, 因此系统的稳态可用度的表达式为

$$A = P_0 + P_2 = \frac{1 + 2\frac{\lambda}{\mu}}{1 + 2\frac{\lambda}{\mu} + 2(\frac{\lambda}{\mu})^2 + 2(1 - C)\frac{\lambda}{\beta}} \quad (2)$$

式中: C 为故障判别的成功率; β 为故障诊断率(平均故障诊断时间的倒数)。

对于结点机级切换方案, 其可靠性模型如图 5 所示, 对于机组级切换方案, 其可靠性模型如图 6 所示。图中, 凡涉及结果比较的结点机, 均由结果比较软件(用 SW 表示)和硬件(含系统软件与部分应用软件, 用 HW 表示)两部分组成, 未标明的结点机均是指结点机硬件(含系统软件与部分应用软件)。其目的在于分析结果比较软件对系统可靠性的影响, 并对切换方案进行比较。

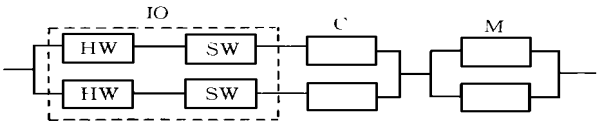

图 5 方案 1 的可靠性模型

Fig. 5 Reliability modle 1

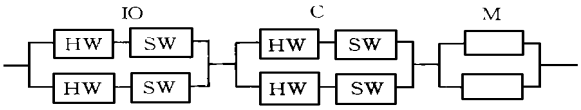

图 6 方案 2 的可靠性模型

Fig. 6 Reliability modle 2

本文采用综合考虑故障判别成功率和故障诊断率对系统可用度影响的可用度模型来分析两种方案的可靠性。假定所有结点机(IO、C、M)硬件的 MTBF 均为 10000h(小时), 所有结点机的 MTTR(含软件)均为 3h(小时)。将有关的可靠性和可维性参数、故障判别成功率 C= 0.8、故障诊断率 β= 6 等代入式(2), 可得系统的可用度及 AIF 随结果比较软件的 MTBF 变化的值, 其结果列于表 1。

表 1 系统的可用度及 AIF

Tab·1 System reliability and AIF

$MTBF_{SW}$ / 可用度	∞	50000	20000	10000	5000	1000
方案 1	0.9999792	0.9999777	0.9999754	0.9999716	0.9999637	0.999888
方案 2	0.9999795	0.9999767	0.9999724	0.9999651	0.999950	0.999804
AIF	-1.44%	4.48%	12.2%	22.9%	37.7%	75%

4 结论

(1) 由表 1 可知: 在一般情况下, 方案 1 的可靠性优于方案 2, 且其可用度改进因子(AIF)随着结果比较软件的 MTBF 减小而增大; 仅当结果比较软件的 MTBF 趋近于无穷大时, 即认为结果比较软件是完全可靠的情况下, 方案 2 的可靠性才优于方案 1, 而这只是一种理想情况。

(2) 从实时性角度来看, 方案 1 的实时性优于方案 2。这是由于方案 2 增加了一次结果比较, 因此增加系统的中断次数和时间开销, 从而影响系统的实时性。

(3) 对于本系统, 机组级切换优于结点机级切换, 且可以减少程序的复杂度, 提高系统的可靠性与实时性。

参考文献

1 胡华平. 分布式实时系统的可靠性模型. 计算机学报, 1997, 20(Supplement): 71～76

2 朱海滨等. 分布处理技术. 长沙: 国防科技大学出版社, 1997

3 金士尧, 王志英, 胡华平. 强实时高可靠的群机系统设计与论证. 计算机工程与科学, 1997, 19(A1): 1～5

4 Hu Hua-ping et al. Reliability Analysis of Reintegration Computer Systems. In: Ma huizu eda. Proceedings of ICRMS'96. Guanzhou: Publishing House of Electronics Industry, 1996, 116～122

5 胡华平等. 实时双机系统的可靠性分析. 模糊系统与数学, 1997, 11(3): 34～39