分布式网络设备的业务不中断软件升级系统设计①

张敏狄　郭裕顺 (杭州电子科技大学 电子信息学院 浙江 杭州 310018)

摘　要： 以往用软件升级来增加新的特性和能力或软件维护是导致网络系统中断的主要原因之一。以冗余路由器为例，设计了一种能在分布式网络设备上实现业务不中断的软件在线升级系统，利用不中断转发(NSF)、状态转换(SSO)、热补丁、进程间通信(IPC)和主控硬件冗余等现有成熟技术在用户流量不中断转发的情况下真正实现了软件在线升级和版本替换，在升级过程中涉及到兼容性检查，回滚定时器设计等内容，有效地解决或减少了软件升级带来的业务中断现象，进一步提高了网络设备的高可用性。

关键词： 业务不中断；高可用性；兼容性技术；可回滚；热补丁

Design of ISSU System for Devices in Distributed Network

ZHANG Min-Di, GUO Yu-Shun

(School of Electronics Information, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Today software upgrades to have new features or capabilities, or to keep maintenance which is still the primary cause to inaccessibility of network system. The design of In-Service Software Upgrade (ISSU) system of redundant route processors takes advantage of Nonstop Forwarding with Stateful Switchover(NSF/ SSO), patching, IPC and hardware redundance to permit true in-service software upgrades or version changes while continuously forwarding user traffic. In the upgrading process of the compatibility check and rollback timer design are considered. So the ISSU system have faster upgrades, minimal impact to service and higher availability.

Keywords: ISSU; HA; compatibility; rollback; patching

随着 IP 技术的飞速发展，各种增值业务在互联网上得到了广泛的应用。新兴的 NGN/3G、IPTV 流媒体、大客户专线和 VPN 互联等重要的电信级业务，对 IP 电信网的可用性提出了很高的要求[1]。即便是相对较小的中断，也将对最终用户的服务体验产生负面影响。在这些自动化应用(交易、同步和备份)、全球化趋势以及其它“全天候流量”推动力的作用下，任何类型的网络节点中断都是无法容忍的。然而伴随着网络业务量的增多，新兴业务的出现，网络设备的软件版本更新显得越来越频繁，使得业务因软件版本维护或者升级而被迫中断，极大地影响了网络设备的高可用性要求。

1　引 言

网络高可用性技术[2,3]，是指一个产品或系统对客户持续服务的能力，基本可以归入容错技术，即在网络出现故障时，确保网络能快速恢复，它可以通过平均修复时间 MTTR(Mean Time To Repair)和平均故障间隔 MTBF(Mean Time Between Failures)两个指标进行衡量，MTTR 是指一个组件或设备从故障到恢复正常所需要的平均时间，MTBF 是指一个组件或设备的无故障运行平均时间，表示为：

$$Availibility = \frac{MTBF * 100}{MTBF + MTRR}$$

在承载网中，网络设备的可用性要求达到 99.999%，大致相当于设备在一年的连续运行中因各种可能原因造成停机维护的时间少于 5 分钟。因而当今的“始终运行”的网络不能再通过“非高峰期中断”的方法，来间隙的规划维护或升级工作了。据此各大

① 收稿时间:2009-09-26;收到修改稿时间:2009-11-05

网络设备生产商以提高 MTBF 或者降低 MTTR 来提高网络系统的高可用性为目的开始推出业务不中断升级 ISSU(In-Service Software Upgrade)系统[4]，打造持续运行网络。本文结合不间断转发技术、热补丁技术[5,6]以及更加严密的兼容性检查技术对分布式网络设备的 ISSU 系统进行了设计与改进，利用这种设计方法考虑到很多潜在的设备中断原因，并设法在问题实际发生时提供自动防故障安全机制和回滚处理，从而快速、自动地识别和隔离问题，恢复正常运行，更加确保了网络的稳定性和高可用性。

2 ISSU系统的构架与改进

2.1 业务不中断升级系统简介

分布式网络设备以冗余路由器为例，在硬件上拥有两块或两块以上主控板(以两块为主)，它们是整个设备的核心，承担整个系统的路由处理、资源管理、状态监测、网管代理等全局功能；另外拥有多块业务板，负责具体业务处理，如 IP 报文转发，MPLS 报文交换，QOS 保证等工作。硬件主控冗余的设计为分布式网络设备的 ISSU 升级提供了先决条件。

本文提出的 ISSU 软件升级系统是通过命令行执行顺序进行的，它结合主控冗余热备份、IPC[7]、热补丁以及 NSF/SSO 等外部模块拥有加载(LOAD)、倒换(RUNSWITCH)、确认(ACCEPT)和完成(COMMIT)四个升级过程实现了整个设备的升级，同时在升级过程中增加了三个异常回退(ROLLBACK)过程，包括手动回退和回滚定时器超时回退。当发现有任何异常，比如升级过程中出现故障、误操作以及升级超时等原因导致升级过程无法继续，我们都可以以手动回滚和等待超时的方式来终止本次升级，增强了升级系统的灵活性和容错能力。具体流程如图 1 所示。

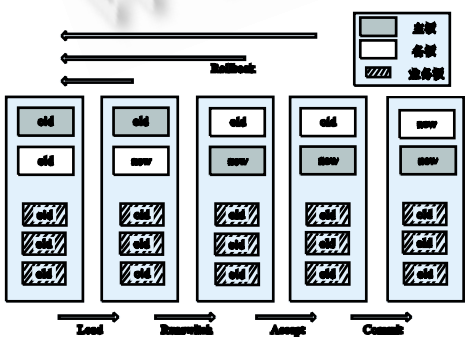

图 1 不中断业务版本升级过程

2.2 软件版本兼容性检查技术的改进

在升级的 LOAD 阶段，备板会加载升级软件版本并重新启动，导致主备板各自运行不同的软件版本，在这个过程中系统会通过板间交互同步不同的消息类型和数据结构，同时对两个版本进行兼容性检查，根据不同的检查结果选择合适的升级方式。兼容性可以分为静态和动态两种，静态即版本支持特性的兼容性，这种兼容性一旦版本产生就是确定不变的，而动态则指当前运行功能的兼容性，保证升级不影响设备当前正常运行。很多生产商仅通过对新老版本的静态兼容性检查来判断升级模式，本文提出的 ISSU 升级系统不影响设备当前正常运行的同时会根据代码的核心构架，注册的协议、数据报文以及配置消息的差异将兼容性检查结果划分为完全兼容、部分兼容、未知兼容和完全不兼容四类。

在完全兼容和部分兼容情况下进入正常的升级模式，利用主控冗余技术以及不间断转发技术进行主备倒换进入 RUNSWITCH 阶段；在完全不兼容的情况下，系统会进入 RPR(Route Processor Redundancy)升级模式，在 RPR 模式下，启动配置和引导注册表项需要在主备板之间同步，由于主备板之间加载的是完全不兼容的版本，在主备倒换的时候，备板成为新的主板必须完成 boot 引导过程，这个过程需要断流几十秒时间，但仍远比以前的断电升级节省时间。由于未知兼容存在着一定的升级风险，本升级系统将会退出该次升级，增强了升级的灵活性。

2.3 基于 ISSU 系统的不间断转发技术

上文提到的双主控板设计的好处是：其中一块是主用主控板，处于工作状态，另一块称作备用主控板，处于备份状态，主用主控板运行过程中，将所有静态配置信息和一部分动态信息备份到备用主控板，使得备用主控板具有和主用主控板相同的配置信息。当主用主控板因为硬件或者软件失效出现故障时，备用主控板接管失效主控板的工作，包括对控制平面和转发平面以及各业务板的控制，一定程度上提高了网络的高可用性。

不过，因为新的主用主控板在主备切换前不参与控制平面的处理，在切换后需要重新和邻居进行会话协商，所以虽然保存了完整的转发表项，但只能避免部分流量不中断。比如，二层业务，以及从本设备往外发送的流量可以不中断；另外，如果和邻居之间配置的是静态路由或静态 LSP 的话，邻居也会继续往发生倒换的设备发送流量，流量不中断。但如果和邻居

之间是动态路由协议或动态标签分发协议，和邻居之间的流量是会中断的，这是因为控制平面会话重置的情况下，邻居的控制平面会重新计算，选择它认为合适的路径。以 OSPF 协议为例，新 Master 在发出的 Hello 报文中没有原来邻居的 RID，会导致邻居把 OSPF 会话状态重置，并把和发生切换的设备相关的 LSA 删除，导致路由重新计算，如果有其他可选路径的话，流量会绕开发生主备切换的设备，如果没有可选路径，则需要等待 OSPF 重新收敛，在重新收敛之前，邻居是不会把流量发给发生主备切换的设备的。以上分析中可以看出，路由器进行主备切换时，在路由协议层面会与邻居之间发生震荡，这种邻居关系的震荡将最终导致路由震荡的出现，使得主备切换路由器在一段时间内出现路由黑洞或者导致邻居将数据业务进行旁路，进而会导致业务出现暂时中断。

本文提出的 ISSU 升级系统利用了不间断转发 NSF 和状态转换 SSO 技术有效地保证了 RUNSWITCH 阶段进行主备倒换，数据转发能够不间断地正常进行，从而保护网络各种流量几乎不受影响。为了实现不间断转发 NSF 技术，首先，要求路由器具有分布式体系结构，数据转发与控制分离，在发生主备切换时，备板必须能成功保存 IP/MPLS 转发表项(转发平面)。其次，根据需要，可能需要保存各种协议的状态(控制平面)。具体原理如图 2 所示。

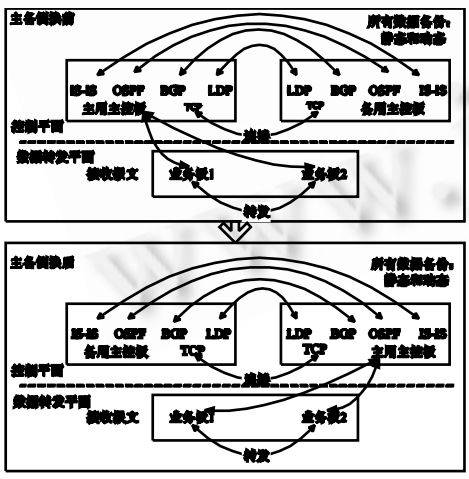

图 2　分布式网络设备的主备倒换不间断转发原理

对于所有信令协议，如 OSPF、IS-IS、BGP、LDP，备份所有数据，包括静态数据和动态数据。其中，BGP、LDP 在 TCP 上运行，这些协议备份每个 TCP 连接上的动态协议数据，也要求系统备份 TCP 链接状态。当主控板发生故障或者人工倒换时，备用主控板立即接管控制平面，所有控制平面应用协议包括 TCP 协议按照已备份的动态数据运行，OSPF/IS-IS 协议立即发送 Hello 报文，BGP/LDP 协议发送 keep-Alive 报文，这些协议与邻居的连接继续维持，不会中断，并借助邻居设备的帮助，可使路由协议包括 MPLS 信令协议与其它邻居断连和重新建立连接时不产生路由震荡，且主备倒换过程中业务板不会重启，业务板上的转发表被保留，可以达到继续业务转发的目的。

2.4 以补丁方式升级业务板

NSF 和 SSO 技术保证了主备倒换过程中业务板不会重启，继续业务不间断转发的目的，可并未实现业务板的软件升级，很多生产商会将业务板的新软件版本在主备倒换的过程中迅速加载到业务板内存里，然后依次快速重启来升级业务板，一定程度上保证了业务的延续性，缩短了业务中断时间，但仍会导致业务中断。

本文提出的 ISSU 升级系统有效的利用热补丁技术实现了不断电修正原有版本中的错误或者增加新的功能进而得到软件升级。补丁是计算机软件系统和软件工程学中的一个术语，一般是为了对系统中的某些错误进行修正而发布的独立的软件单元。它能够在不影响系统正常运行的情况下完成对系统错误的修正，也就是对系统进行动态升级。它在系统中保留一段内存空间，将新的函数实体以补丁文件的方式加载其中，根据要被替换函数的入口地址找到被替换函数的第一条执行指令，将其改为一条跳转指令，当其他函数要调用被替换函数时，CPU 根据跳转指令就会执行新的函数实体。

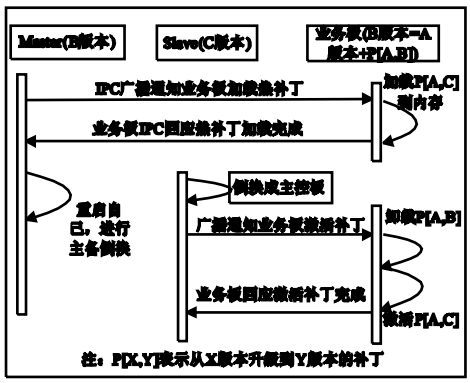

图 3　热补丁方式实现业务板软件的不中断升级

图 3 描述的是 ISSU 系统对业务板不中断业务升

级的一个简单过程，假设 A、B、C 版本是完全兼容的，在升级以前我们已经将业务板的升级软件、升级补丁和主控板升级软件一起打包在 C 版本的 APP 里。图中已经用 ISSU 系统通过补丁方式将 A 升级到了 B 版本，并完成 B 版本再次升级到 C 版本的 LOAD 过程进入 RUNSWITCH 阶段。

在这个阶段主控板会 IPC 广播通知业务板加载热补丁，待加载完成后主控板重启自己来触发主备倒换事件，待备用主控板成为新的主控板时 IPC 通知业务板激活热补丁生效，系统会先将从 A 到 B 的升级补丁卸载然后激活从 A 到 C 的升级补丁，然后回应主控板激活完成。当业务板补丁加载失败或者补丁不存在时，主控板会将业务板升级软件迅速加载到业务板内存里，然后依次快速重启来升级业务板，依然减少了业务中断时间。

3 ISSU系统的升级实现过程以及测试结果

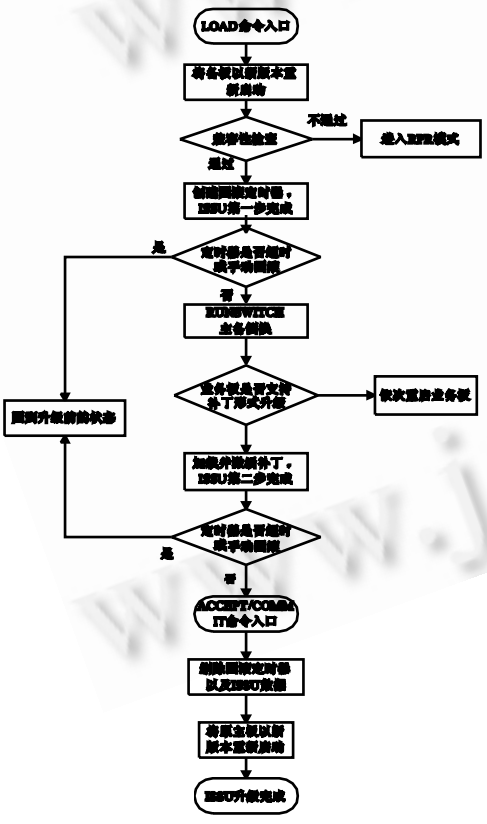

图 4 ISSU 系统的升级流程图

本文的ISSU系统是在linux系统基础上借助 H3C 的 C 语言软件平台实现的，下面以冗余路由器的升级过程为例，实现业务不中断软件升级。首先在路由器的主备板利用ftp或tftp协议将待升级软件加载到网络设备存储介质，它包括主控板软件版本和业务板升级补丁。图4为ISSU系统的升级流程图，用户通过issu load命令将备用主控板以新版本重新启动，通过版本兼容性检查后进入主备倒换阶段，同时加载并激活业务板的升级补丁，然后通过 accept 确认升级，最后通过 commit 将原主用主控板以新版本重新启动完成本次升级。

在冗余路由器上利用 Smartbits 工具在本 ISSU 升级系统通过发送和接收诸如 IP、ATM 信元、POS 等等形式的数据流，对吞吐量、延迟、包丢失率以及信元丢失率等性能指标进行分析。测试结果为：采用热补丁升级的方式，在软件版本完全兼容的情况下保持业务不中断，部分兼容情况下，兼容模块和协议保持流量不丢失，不兼容的协议和模块出现短暂的业务中断；RPR 升级模式下断流保证在 1 分钟之内，待升级成功后业务立即恢复正常。结果表明通过本 ISSU 升级系统一定程度上提高了 MTBF、降低了 MTTR，实现了分布式网络设备在业务不中断的情况下在线升级和版本替换。

4 结论

本文提出的 ISSU 升级系统在分布式网络设备上实现业务不中断软件版本升级，延长了网络的运行时间，大大提高了网络设备的高可用性。并且，随着分布式堆叠技术的逐渐成熟，网络设备开始变得越来越复杂，功能变得越来越强大，为 ISSU 升级系统的发展与应用提供了更加广阔的空间。

参考文献

1 曹濯钦,慕晓冬,郭文普,等.计算机网络技术及应用.北京:人民邮电出版社, 2005.
2 Garbin DA, Knepley JE. Design and analysis of high availability networks. Portugal, 2009. 1－6.
3 Cholda P. High availability network fundamentals: a practical guide to predicting network availability. Communications Magazine, IEEE 2002. 34－34.
4 Kankkunen A. Non-Service Affecting Software Upgrades for Multi-Service Routers. 2005. 16－19.
5 吕文安,薛先久.一种嵌入式系统软件补丁的实现和控制方法:中国, CN02125766.3. 2002-08-16.
6 陈瑜.面向嵌入式系统的在轨软件维护技术研究[硕士学位论文].杭州：浙江大学, 2006.
7 博韦,西斯特.深入理解 linux 内核.北京:中国电力出版社, 2007.