容错那些事

北京长久斯捷有限公司
Beijing Betaone Sysjob Ltd.

前言：

近年来我国高速路建设事业发展迅猛，也为 IT 行业提供了巨大的商业空间。因为高速路项目招投标管理规范，对先进技术的使用表现积极，使得众多 IT 商家趋之若鹜，大量先进的产品和技术得以在高速路行业获得成功应用和推广。

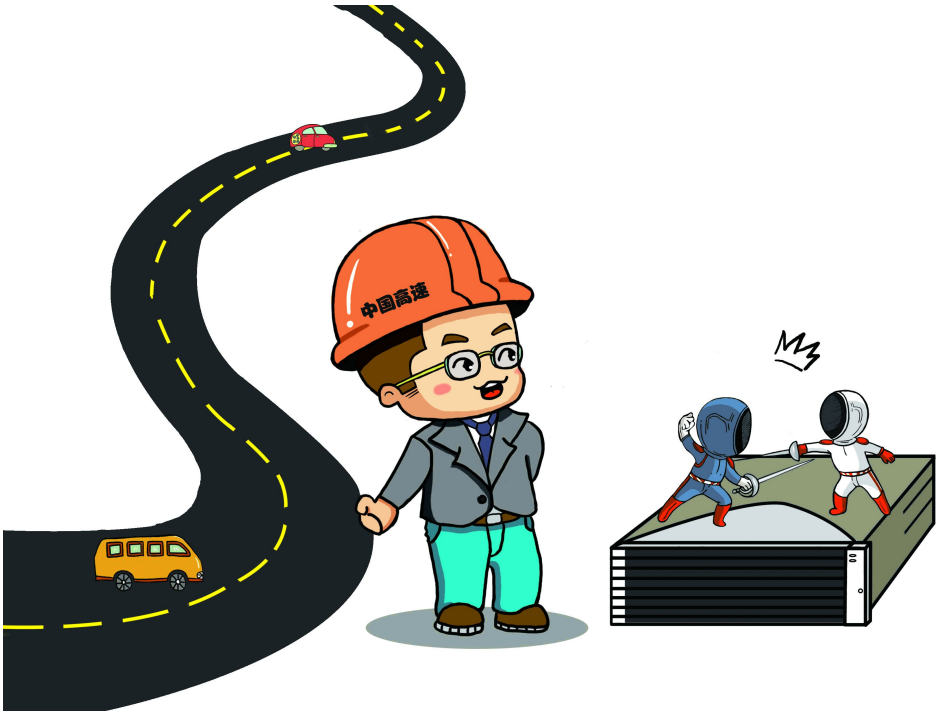
容错服务器面向监控中心、收费中心、收费站等关键业务，在动辄数千万人民币的项目标的中所占份额很小， 一般很难超过一百万人民币。但是因为高速路行业的吸引力和发展前景，硬容错服务器商家在这一方小天地中也展开了一场妙趣横生的竞争。

硬容错服务器在由日本 NEC 生产和直销的同时，美国容错公司（Stratus）也贴牌销售同款产品，继工业自动化领域之后，NEC 和 Stratus 两家各施手段以价格和服务为手段，又在中国高速路行业展开了全面的竞争。

本文以近年来 IT 技术的发展历程为背景，介绍了容错技术曲折起伏的发展道路，对硬容错服务器的核心技术和 HA 双机系统的不足做了探究，同时对现在的竞争格局进行了分析。

良性的竞争对我国高速路行业是一件好事，竞争可以为业主和集成商在价格和服务方面提供更加多样化的选择，良性竞争同时也对商家有物竞天择式的推进。只有通过合理的竞争，利益的天平才能向用户倾斜，让店大欺客的旧格局一

去不再复返。



导读：

一． 三十一年还旧国，落花时节读华章。

容错技术初介

二． 疾风知劲草，板荡识英雄。

刀光剑影的 IT 十年

三． 无可奈何花落去。

早期硬容错技术的衰落

四．你是我胸口永远的痛。

HA 双机容错

五．似曾相识燕归来。

硬容错服务器

六．一枝独秀不是春。

硬容错服务器的竞争

容错那些事

一．三十一年还旧国，落花时节读华章。

容错技术初介

容错技术至今走过了三十年。三十年前容错技术横空出世，其后命运多舛几近消亡，现在又浴火重生再呈燎原之势。三十年来翻翻滚滚起起伏伏，期间有多少利益缠斗人心向背锥心泣血的故事，现在说起来也是饶有趣味。

容错技术只是 IT 领域的一个小角色，和 CPU、网络、数据库等大牌不能相提并论。但是见一叶而知秋，窥一斑而见全豹，从容错技术的故事中可以看到这些年 IT 发展的一些规律。

可以这样简单地下一个结论，一般来说， IT 技术是面向应用的，容错技术是面向 IT 的。

按奥运会“更高更快更强”的口号做以下分工，IT 技术追求的是更高更快，而容错技术则关注更强。

高和快源于显象的应用需求，如更快的 CPU,更大的网络带宽，更好的人机界面，而容错技术则是保证这样的应用系统更健康更强壮。

面向社会应用的 IT 技术在最开始研发时是不会考虑容错的，就像 INTEL 的 CPU 或微软的 DOS 操作系统，甚至是 ORACLE 在最研发时都无暇顾及容错。关心容错的不是实验室的开发人员，也不是最终使用者，而是应用系统的建设者，除非后二者合一。这样说看似不可思议，其实从最终用户的角度出发，他并不太关心承包商是如何搭建开发这套应用系统的，而更关心好不好用。

“好不好用”可以从两个角度解读：一是操作复杂度，二是系统稳定性。系统的稳定性就是容错技术的研究对象。有一个大家熟知的 IT 术语“Availability”就是在讲“好不好用”，只是这个词目前更侧重系统稳定性。生活中的道理也是同样的，没有一个家庭会关心冰箱内部的电路设计，只关心使用是不是方便，东西是不是结实。

二． 疾风知劲草，板荡识英雄。

刀光剑影的 IT 十年

西方国家大规模的 IT 系统建设始于 1980 年，我国当时有十年的滞后期，中国的石油、海关、交通、 烟草、测绘、航空、铁路、金融等行业的 IT 建设大致是从 1990 年开始进入了十年的建设期。

今天我们生活中无处不在的 IT 服务，如票务、同城通兑、异地存取、财税联网等很多都源自那时。特别需要说明的是，上世纪九十年代的 IT 建设，最开始和互联网没有任何关系的，只是到 2000 年前后因特网快速普及，这些 IT 应用系统才从以前的主机-多用户模式或 C/S 分布式模式升级到现在常见的 B/S 浏览器模式。

1990-2000，中国 IT 的黄金时段。那时真是八仙过海各显神通，中关村就是从那是形成规模的，现在已经在市场上难觅踪迹的一些产品，在当时可是风靡一时。

海关上用的 DECnet 网络，银行里用的 Informix 数据库，证券里用的 SCO UNIX，图像处理上用的 SGI 工作站，甚至还有 IBM 的 OS/操作系统，大学里教的 PDP-总线，机房里基于 MOTROLA CPU 的苹果电脑……，十年之后再看，都是昙花一现，现已芳踪难觅。

话说天下大事久分必合，十年的血拼尘埃落定，十年的中原逐鹿终于有了结果。

- Intel + Windows 的 PC 格局无可撼动；
- Oracle 在四大数据库厂家中独占鳌头；
- Cisco 在八大网络制造商中脱颖而出；
- IBM、HP 宝刀不老；
- Dell、联想小心翼翼地跟上了脚步。

当看到这些一统天下的王者时，不由地想起一句中国古话：真是一将成名万古枯啊。在这些王者阴影下，又有多少曾经威风八面的 IT 公司已万劫不复。

这一场市场洗牌用了十年时间，结果却是如此惊心动魄。曾经熠熠生辉的 CDC、DEC、SUN、Compaq、Infomix 等，恍如一夜之间竟已销声匿迹，MOTORLA、Unisys、SGI，NB 等一蹶不振至今都无法东山再起，那些曾自以为天之骄子的 IT 白领同志们，在 2000 年前后也都是惶惶不可终日。

市场格局尘埃落定，IT 在经过十年建设期后，也已逐渐进入使用期。

IT 系统从研发建设转入实用后，其管理权就从开发商转移到了客户手中，一般来说是各单位的信息中心。这很像道路建设的模式，高速路建好后要移交给运营公司。1990 年时的 IT 格局有点类似我们的 1949 年，那一年是建立国家到管理国家的分水岭。

三．无可奈何花落去

早期硬容错技术的衰落

“更高更快”的事告一段落了，“更强”的事就该粉墨登场了。一个正在运转的系统，其安全健康现在成了头等大事，创业难守业更难的道理在这里显然也成立。

容错技术终于迎来了第一个春天，这个时间节点在国外是 1985 年，在中国是 1995 年。

随着大批应用系统投入运转，容错技术的前景看似是繁花似锦，但天有不测风云，接下来的五年中所发生的情况一定让你目瞪口呆，历史在这里和容错技术开了一个不大不小的玩笑。

为了下面的讨论， 在这里先简单论述一下容错技术。

容错技术，**Fault Tolerant。**

广泛地说，容错技术涵盖很广，内存的 ECC 校验、网络的奇偶校验、CLUSTER 集群、UPS 不间断电源、热插拔板卡，这些都可以算作容错技术。

狭义而精准地定义容错技术，就是在计算机平台上如何实现故障下的作业连续和数据完整。

容错技术从用户的角度说得最明了：作业不停，数据不丢！

CPU 故障、内存故障、总线故障、硬盘故障、电源故障、网卡故障，甚至还有软件死机，这些问题在开发过程中还来得及讨论解决，一旦应用系统上线运行，银行排队的储户，售票点的旅客，高速收费口的司机，还有等生产报表的领导，他们不会再给你解释故障停机的机会。

真正意义上的容错技术是硬件的彻底冗余，程序运行同步到指令级别，只有这样，才能做到作业不停，数据不丢，这也就是所谓的硬容错技术。

为了强调安全性，甚至可以配置多套硬件冗余故障时自动切换，这就提到了容错技术中“度（degree）”的概念，2 度冗余指硬件双份配置，业界目前的记录是 4 度冗余。

还有一些行业因其特殊性把容错技术做了变化和扩展。如轨道交通中采用“三取二”模式，机车上的一个传感控制点同时由三台独立的计算机采集和控制，每次采集和控制信息都需经过比对，只有三台中至少两台计算机的信息一致才能发起控制。飞行控制中每个传感点按四个独立通道发送数据，四通道数据进行比对后才能被认可。

HA 双机容错系统算不算容错技术？答案是肯定的。但是这个答案中有很多

无奈心酸，看了下文就知道为 HA 双机容错是一种局促的容错手段， HA，那真是没有办法的办法。

说到硬容错技术，有几个公司的名字必须要提及，美国的 Tandem 公司，Stratus 公司和 Sequent 公司。 这几个公司的命运就是容错技术起伏的缩影。

在上世纪八十年代随着西方信息化建设的普及，容错技术率先在美国登场，Tandem 和 Stratus 主要从事硬容错业务， Sequent 公司则从事类似的集群技术。硬容错系统进入中国已是就是九十年代初期。

当时硬容错技术可谓是阳春白雪，一套硬容错系统或集群系统动辄上百万美元，真是皇帝女儿不愁嫁，一年不开张，开张吃三年。国家气象局和西单商场等大型客户当时均耗资数千万人民币配备了类似的硬容错系统。

但是这样昂贵的代价以及繁杂的技术怎么可能在越来越普及的 IT 市场上普及呢？除此以外，这些硬容错系统还有一个致命的缺点，那就是系统的封闭性。硬件结构和操作系统甚至是 CPU 都是各自为政，你购买了这样的系统回去后还要花费大量的时间做软件移植。

阳春白雪的硬容错技术啊，像你这样的王谢堂前燕，如何不能飞入寻常百姓家，可怎么存活呢？结果真是印证了那句话：其兴也勃焉，其亡也忽焉。从 1995 年到 2000 年，短短五年，市场给了高高在上的硬容错技术一记响亮的耳光，只

要看看这几个公司的命运就知道了：

Tandem 公司：成立于 1974 年，1997 年被康柏公司收购，后康柏公司又被 HP 公司再购。今天在 HP 的 Nonstop 产品中或还能见到一点 Tandem 的身影，但是 Tandem 这个名字却永远地退出了 IT 界。Tandem 公司同时还完成了另外一个轮回：该公司的创始人 James Treybig 当年是从 HP 出走创建了这家公司，二十年后尘归尘土归土，Tandem 又回到了 HP 的手中，因果循环令人嗟呀。

Stratus 公司：成立于上世纪七十年代，和 Tandem 一样，走过了辉煌以后于 1998 年被 Ascend 收购，一年后国际投资公司 Investcorp 又注资回购 Stratus，两年后和日本 NEC 公司达成协议，由 NEC 生产新一代硬容错服务器，Stratus 贴牌后销售。

Sequent 公司：1999 年被 IBM 收购，其业务已并入 IBM 的 Unix 阵营中。

在这期间，SUN，MOTOROLA，HP 等公司也曾在硬容错产品上做过尝试，终是一击不中全身而退，没有像上述几家全军覆没。

在应用系统大面积普及的形势下，本该大放光彩的硬容错技术为什么落得个山穷水尽的局面？问题不在硬容错技术本身，而是硬容错技术的先驱们太不把市场和用户当回事了。

昂贵的价格，封闭的系统，繁杂的操作，在 IT 标准逐渐规范统一开放的今天就显得太不合时宜了。即便你貌若天仙 ,在世界大同的今天你如果还是三寸金莲，你还是会被社会所摒弃的。

谁是 IT 发展的推手？客户需求！谁是技术和产品的生死判官？用户!

这个道理在今天说来如此简单，在当年店大欺客的时代竟然是无人理会，但是实践是检验真理的唯一标准。真希望华为苹果等国内外 IT 后继大佬深以为戒，不要让历史悲剧重演。

四．你是我胸口永远的痛

HA 双机容错

中国的客户可怜，没有那么多资金去玩早期的硬容错，中国的客户幸运，也没有在这里烧钱。

但是容错需求是现实摆在那里的，因为生活还是要继续的，这时，我们称之为“穷人的容错系统”登场了，这就是 HA 双机容错系统。

单服务器仅靠 MTBF（平均无故障时间）来保证可靠性，这对于关键业务是无法接受的，应用系统不能赌运气，对这个问题的认识目前都已是共识，所以不再赘述。

HA 双机容错系统的出现一下就把系统的故障率降低到了每年几天甚至是数小时。

HA 系统的结构其实很好理解，两台服务器通过链路定期相互侦测，所谓链路一般是指网络或 RS 串口，两台服务器事先通过容错软件做好脚本配置，当一台服务器发生故障时，另一台服务器可以接管应用继续运行。

看上去很美。代价不过是两台服务器再加上容错软件（当然，所需要的系统软件或应用软件也需要双份）， 成本不过 5 万人民币左右。1995-2005 年是 HA 双

机系统的黄金时期，铁路、证券、银行、医院等陆续装备了大量的 HA 双机系统。

本文开篇即已提到，HA 双机系统是一种心酸无奈的选择，其原因有二：

一是当时我们没有低端平台硬容错系统的选择；

二是 HA 双机系统是一柄双刃剑，价格的确低廉平台的确通用，但是其痼疾也是一枚定时炸弹。

我们再回到容错系统的定义：运行不间断，数据不丢失。

双机系统在这两点上革命都不彻底。

双机系统相互侦测是不连续的，间隔一般在秒级，通常是 1-3 秒。一秒代表什么？对于现在的 CPU，一秒就可以执行一亿条指令。而且双机系统为避免不必要的切换，还要将数次的侦测结果合并判断后再做出切换决定。

一旦开始切换，数据库，网络 IP 等都需要时间启动或重配，应用系统也需要同步，这个时间大致在数分钟到半小时之间。

如果说切换时间根据不同的 HA 软件还可逐步优化，关于数据丢失问题则更是双机系统的痼疾，这个问题在 HA 双机系统上没有解决的可能。

我们知道，数据首先是存放在内存中，当需要的时候才会保存在硬盘中，这和我们使用 WORD 时需要经常点击保存是一个道理。但当 CPU、总线、内存、电源等关键部件故障时，没有来得及保存的内存数据必然丢失。这时即便你将应用

切换到另一台服务器上，因为数据发生了断篇，作业也不会连续了。

举一个例子。高速路收费时是一次收款对应一张发票，如果你收了款并已通过收费系统界面录入，此时如果突然发生硬件故障，收费系统内的数据库还没有做二次提交，此时数据还在内存中，随着故障的发生数据必然丢失，等系统切换后你将遇到尴尬的局面，票款记录不符！此时如果不出票司机不干，如果出票，票款轧帐又对不上。

高速路收费毕竟每次额度都不大，如果是银行存款或证券交易，就会造成很大的麻烦。但这还不是最危险的，如果是在轧钢生产线上，这就真是事故了，如果在军工试验场，这就会上升为责任。

笔者自 1995 年开始从事容错技术的应用，包括基于智能磁盘阵列的数据安全存储和基于 HA 软件的双机容错。此后数年， 在全国石油、铁路、航空、银行、证券等行业设计实施了近千套 HA 双机容错系统，涉及 WINDOWS、SCO UNIX、UNIXWare、IBM/AIX、SUN/Solaris、HP/UX、UNIX /SVR4 等多种系统平台以及 Oracle、 Informix、 Sybase、 DB2、 SQL/Server 等数据库。

笔者当时真是感觉到了 HA 是“穷人的容错机”这句话，几十万人民币的代价即可实现一定的在线容错能力，较好地保证了作业连续性。

但是，HA 的顽疾也让人触目惊心 故障时 HA 系统的内存数据丢失的现象时

有发生，造成银行轧帐不齐或网点票款对应不符，曾造成长时间的系统停滞，领导发怒储户骂街，银行证券单位内外当时是一片混乱。

一套套正在上线运行的 HA 系统在我眼里逐渐变成了一颗颗定时炸弹，因为内存数据丢失在 HA 层面上是无法解决的。作为从事容错技术的一员，当时多么希望在 PC 平台上能有一台价廉物美的硬容错服务器，让我再也不用担心内存数据丢失，再也不用设定复杂的 HA 软件，再也不用小心翼翼地编写切换脚本。

梦想成真，千禧年过后， Intel 平台的硬容错服务器真的面世了。

五．似曾相识燕归来

硬容错服务器

IT 的事往往就是这样，车到山前疑无路，柳暗花明又一春。

为了追求网速，先从同轴电缆演变为双绞线，双绞线内芯又从 2 根增为 8 根。可是天长地久应有尽，再宽的路也担不住太多的车，相互间的电磁干扰让这种方式走到了尽头。山穷水尽之时，光纤横空出世。单路串行按理说是最落后的通讯构架，在光速的配合下却是熠熠生辉。这不仅是技术的成就，更是思想的光芒。

容错技术也是如此。那些高端大气上档次的大型容错机因为系统封闭百姓不会使，因为昂贵百姓买不起。但是如果将这种技术移植到 PC 平台上，不就可以峰回路转了吗？

答案是简单的，过程是困难的。

因为硬容错技术涉及到两个 CPU 指令的同步，其难度更甚于设计一款 CPU。要想将容错技术下迁至 Intel,不仅需要专利技术，而且需要强大的硬件研发设计能力和庞大的资金保证。

终于有一个公司站出来了，这就是日本 NEC。

NEC 全名为日本国家电气公司，是日本 IT 巨头，全球员工三十万，产值大过联想集团，从手机投影仪到服务器大型机，甚至连卫星都在其产品线上。NEC

的超级计算机当年曾胜过我国的银河排名世界第一。

NEC 独具慧眼，看中了对容错技术的巨大发展商机。NEC 与 Stratus 合作，终于在 2002 年研发生产出第一台基于 Intel CPU 和 Windows/Linux 平台的硬容错服务器，而且价位在人民币在十几万到几十万人民币之间，与 HA 双机系统在同一个档次上。

商家无利不起早，NEC 投入巨大资金和人员研发硬容错技术，自然有其商业考量。但是硬容错技术终于走出阳春白雪，走进千家万户，NEC 可谓厥功甚伟。

六．一枝独秀不是春

硬容错服务器的竞争

硬容错技术到此应该是柳暗花明一马平川了。

是，而且更是。

硬容错服务器面世后得以迅速在全球普及，这是不争的事实。与此同时，另一场竞争好戏又开场了，这场好戏对于用户来说可只有好处没有坏处。

根据日本 NEC 和美国容错公司（Stratus）达成的合作协议，由 NEC 生产的硬容错服务器在 NEC 直销的同时，美国容错公司也可以贴牌销售。于是出现了这样一个戏剧化的局面，一台 Made in Japan 的硬容错服务器，NEC 销售时叫做 NEC 5800/ft, 美国容错公司销售时则变成了 Stratus ft/server。

大家注意， 这可不是李逵和李鬼的故事翻版，这是 IT 业界正式的合作协议。当然，在生活中如果出现这样的事的确很难理解，比如说美国福特贴牌销售日本丰田的汽车，人们心里很就难接受福特。但这种贴牌或 OEM 的方式在 IT 界却是有过先例的。当年 PowerPC CPU 问世时，IBM、 MOTOROLA、法国 BULL 公司就曾达成协议，由其中一家生产小型机，其余两家贴牌销售。

这种合作模式一旦投放到市场上，一种有趣的格局就形成了。

美国容错公司进入中国市场较早，再加之以美国品牌宣传，在中国先期占有了较大的容错市场份额。

日本 NEC 在完成其国内容错市场的洗礼后，相继也进入了中国市场。

NEC 来的有点晚，而且正赶上中日大气候还不对，NEC 必须低调而且要拿出亮点来才行。

更优惠的价格，更妥善的售后服务。这就是 NEC 的两招杀手锏。

作为制造商， NEC 当然在成本控制上有更大优势，既然来晚了，而且来的有点不是时候，端正态度的第一招当然是更优惠的价格。同时，NEC 在中国浸淫多年， 其在华服务和物流网络已非常完善。所以 NEC 向用户承诺，由 NEC 保障客户的服务和备品备件保修，为客户带来两层保障：代理商和生产厂家。

结果一场好戏就上场了，同一台硬容错服务器，贴着两家商标，从工业自动化到高速路市场，在中国翻翻滚滚打得不亦乐乎。

天知道 NEC 和美国容错这两家当时达成了什么样的合作协议？这样妙趣很生的竞争倒是在 IT 业界很多年没有见过了。

这场竞争与 IBM、HP 之间的竞争有很大不同。同一台硬容错服务器，无所谓性能高下问题，硬容错服务器属于通用平台产品，不存在二次开发问题，所以理论上说，这两家的成功案例都可以互用，竞争只能在价格、服务和市场宣传层

面展开。目前 NEC 携其中国合作伙伴长久斯捷和慧桥等公司，美国容错携其代理中科软和海得等商家，在全国各个行业上短兵相接展开了全面的竞争。

硬容错服务器的确是关键业务的首选，NEC 和美国容错在国内的这几家代理大都曾是 HA 双机容错的大家，现在他们宁可在硬容错服务器市场上杀得天昏地暗也不愿回头再捡起 HA，说明他们很识货，更说明他们也在 HA 上也有过类似的伤痛。

科学技术是第一生产力，技术无国界，按鲁迅先生《拿来主义》的说法，只要能为我所用就行。这场竞争和当年旅顺口的日俄战争不同，鹬蚌相争，中国的用户大可悠哉悠哉地做渔翁。

在这场对中国用户有赢无输的竞争中， 中国客户只需要注意两点：更优惠的价格，更好的服务保证。

目前中国用户的具体采购都来自于 NEC 或美国容错的国内代理， 既然这场竞争是由 NEC 和 Stratus 发起的， 那就不能让这些国外厂商免责。中国用户在追求更优惠价格的同时，应该用服务框架协议的方式由 NEC 或美国容错在服务、保修、备品备件、产品升级等方面，向中国用户就具体项目提供进一步的具有长期法律意义的书面保证，而不能用一纸保修证书敷衍了事。

一枝独秀不是春,百花齐放春满园。

双桨单舟总争渡，千帆竞发渡江海。

用这两句对联来形容 NEC 和美国容错的竞争倒是很合适。按达尔文物竞天择的说法， 合理的竞争对商家也是好事，对用户更是乐见。只有合理的竞争，利益的天平才能向用户倾斜，让店大欺客的局面一去不再复返。