豆腐卖了肉价钱

-也谈中国容错市场之怪象

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

豆腐卖了肉价钱

- 也谈中国容错市场之怪象

一． 容错初兴， 市场产品鱼目混珠

1. 双机容错系统和硬容错服务器：豆腐和肉

“豆腐卖了肉价钱”是一句民间俚语，常用于购物后悔的场合，意即花了冤枉钱。

在生活中我们不会用买肉的价钱去买豆腐，我们更不会用买汽车的钱去买自行车，这是因为我们的经验和知识让我们对货值比较心里有数。但是在今天的中国容错市场，“豆腐卖了肉价钱”的事却时有发生，因为有一些用户的确是用买容错服务器的钱去买了 HA 双机容错系统。

容错技术在国外已比较成熟和普及，在国内正处于初兴阶段。国内一些关键业务，如道路交通管理、生产过程控制、医疗卫生银行证券等行业，对作业连续性和数据安全性要求有着特殊的高要求，所以在服务器的选择上率先采用了容错产品。

正因为容错技术处于初兴阶段，大量的新名词、新概念、新产品充斥其中。MTBF，HA，双机容错，双机冗余，硬容错，还有所谓几个 9 的可用性等，看

得让人莫名就里眼花缭乱，其实这里也有些故弄玄虚的市场宣传成分。用户本身都在担负着自己的科研或生产管理工作，没有很多闲暇去一一研究鉴定，所以目前容错技术在产品和市场宣传上时有鱼目混珠的事情发生。

2. 从容错角度来看服务器的选择

容错技术主要是指服务器容错。一般是指在服务器软硬件上做了部件冗余配置或故障时的软件切换机制，以备在故障时可以自动转移运行。

从容错角度出发，服务器的选择一般有以下三种：

1） 单机服务器，如 IBM/HP/联想等品牌

这类服务器基本没有容错功能，最多就是在电源、硬盘或板卡的热插拔上做了一些基础的改进，但对核心软硬件故障，如 CPU、内存、背板或操作系统内核等故障不存在任何容错能力；

厂家一般喜欢用 MTBF(平均无故障时间)来表征单机服务器的可靠性。

2） HA 双机容错服务器，一般为台湾或国内的一些自研产品，在这个档次上

尚没有国际主流品牌。

这个层次最是眼花缭乱。HA 双机服务器，冗余服务器，双工服务器等各种称谓莫衷一是。

3） 硬容错服务器

目前全球范围只有两家硬容错制造商，三家品牌销售商。

制造商：日本 NEC，美国惠普。

品牌销售商：日本 NEC/Stratus（美国容错）， 美国惠普

日本 NEC 主要制造基于 Intel CPU 的通用容错服务器，操作系统一般为 Windows，VMvare 或 Linux。

日本 NEC 制造的容错服务器除了 NEC 直接销售外， Stratus（美国容错）根据与 NEC 达成的协议，也贴牌自行销售。

HP 的 NonStop 服务器是在 HP 收购传统的容错公司 Tandem 后发展出来的一款容错产品，也具有硬容错服务器的特点，如采用 Lockstep技术。但是 NonStop 主要面向 HPC 高性能计算（High Performance Computing），CPU 最多可扩展到 4000 颗以上，操作系统为 HP 专有，属于高端计算产品。

3. 从容错角度比较三种服务器

从容错的角度来看，普通服务器、HA 双机服务器和硬容错服务器这三者之间不是从低到高的排列，而是三个时代的产品。

1995 年之前，普通服务器时代；

1996-2005：双机容错服务器年代；

2006 年以后：硬容错服务器年代。

作为用户，如果要辨别三者之间的区别其实也很简单：

普通服务器和双机容错的区别：

看看是否有两套 CPU/内存配置。

这种区别是表象的，所以很容易了解。

双机容错系统和硬容错服务器的区别：

看看是否能做到 CPU 指令同步和内存数据切换。

了解这一点需要一些技术背景，这也因此成了是“豆腐卖了肉价钱”的症结所在。

其实用一个生活中简单的实例比较可以直观地看出三者的性能。

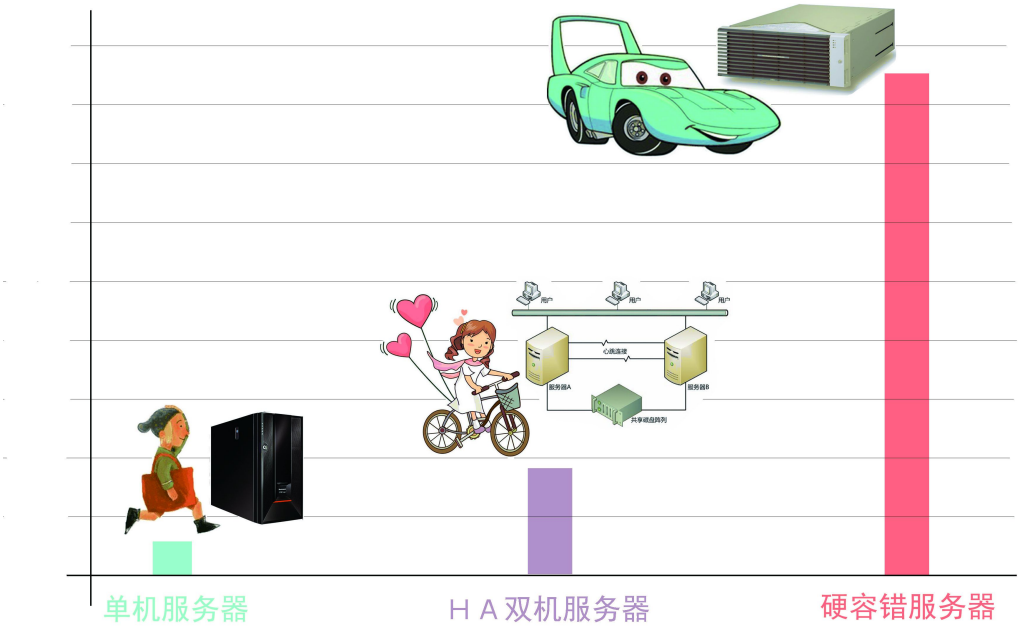
我们把容错性能比做行进速度，

普通服务器相当于步行，(5 公里/小时)，

双机容错服务器相当于自行车，(25 公里/小时)

硬容错服务器相当于汽车。(120 公里/小时)

容错性能



自行车可以装上电瓶改为电动自行车，甚至再加上车篷变身助力车，可是没有气缸发动机，自行车不可能成为机动车。而这里的汽缸发动机就是指硬容错技术中的 CPU 指令同步和内存切换技术，有此两点，硬容错服务器才能做到无间隙的零秒切换。而双机容错系统（或称双机冗余系统）不管标称其可靠性有多少个 9 组成，因为无法做到 CPU 同步和内存数据保护,所以只能是自行车而不是汽车。

关于这三者更详细的技术讨论请见下文分解。

二．天下三分，性能价格泾渭分明

1. 单机服务器和 MTBF

MTBF,即平均故障间隔时间，英文全称是“Mean Time Between Failure”。

MTBF 是衡量一个产品（尤其是电器产品）的可靠性指标，单位为“小时”。

我国对 MTBF 平均无故障工作时间的规定是 4000 小时。联想曾发布过其商用 PC 天启系列商用 PC 的 MTBF 为 45000 小时，达到当时世界领先水平，因为那时国内 MTBF 的先进水平也不过 15000 小时左右。

我们就以联想天启服务器 45000 小时 MTBF 时间为例，解读一下普通服务器的可靠性。

45000 小时相当于五年。 45000 小时的 MTBF 时间并不是说该种服务器每台均能工作 5 年不出故障。而应该理解为 5 年里至少会出一次故障，换言之，即该服务器的平均年故障率为 20%，换言之，一年内，平均 100 台服务器有 20 台会出故障。

45000 小时的平均无故障时间看似很大， 但是折算成 20%的故障率后就触目惊心了。

即便是触目惊心，这在 IT 制造中也是正常的现象。部件的 MTBF 值一般都会高于主机，因为主机中有电源、风扇、主板、插卡、接口等诸多组成部分，所以

作为一个组合体，其可靠性相对于部件来说是极大下滑的。

我们都听说过一颗铆钉、一根空速管就有可能造成飞机失事罹难，其实同样的道理也适用于服务器。容错的目的就是不怕一万就怕万一。

另一个关于 MTBF 更尴尬的话题就是 MTBF 数据的来源。

一个 IT 产品最新发布的时候，其 MTBF 数值往往也同时公布了。而此时该产品不过是零岁而已。MTBF 不是经过实际考核而是厂家经过测试推算的，所以可信度还要再打一个折扣。

从容错的角度来看，一般服务器因为没有解决单点故障的能力，所以可靠性才如此偏低。所以关键业务如果选用普通服务器，在可靠性上要冒很大风险。

也正因为如此。像银行证券等对可靠性要求极高的行业，鲜有选择一台普通服务器作为主机的案例。

2. HA 双机容错系统和那些“9”

单机在可靠性上的先天性不足早已为业界了解，所以才有了后来的 HA 双机服务器。

HA 双机系统瞄准了单机服务器最大的软肋-无法解决单点故障，转而采用双机容错的方式来解决单点硬件故障时的系统程序运行能力。

对 HA 双机容错最简单的理解是，一台服务器故障时，作业自动转移到另一台服务器上继续运行。

双机容错服务器还有很多名称，双机冗余服务器，HA 双机系统，双机双工系统，双机热备系统等等。

双机容错的可靠性指标从上文的 MTBF 变成了用百分比表示的“可用性”，这样，在不同的双机产品中就出现了 99.99%, 99.999%, 99.9999%等华丽的指标。

这类的指标和 MTBF 一样，看似很技术化其实也很空洞。用户在容错需求上特意增加了投资，防的不是一万，而是万一。MTBF 和这些 9 说来说去还是不能避免万一的发生。而且这些指标是内部评估为主，客观性没有经过时间检验。最重要的是，CPU 的指令不同步，故障时内存数据不能切换，稍有 IT 知识的人都知道，这种容错是不可能彻底的。就如上文所说，你即便把自行车加上电瓶搭上车篷，但没有气缸和发动机，永远只能是助力车而不是机动车。

双机容错，看似两台服务器再加一套 HA 容错软件就解决了可靠性的问题。

双机容错，看上去很美，其实是一种心酸无奈的选择，其原因有二：

一是当时（1995 至 2005 年这十年）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。

HA 双机系统在这两点上革命都不彻底。

- 运行不间断：

  在服务器发生故障时（如掉电或 CPU 故障）一定要做到 CPU 指令级别同步。

- 数据不丢失：

  服务器发生故障时一定要做到内存数据不丢失。

这两点 HA 双机系统都做不到。当然，HA 双机系统也不可能做到。因为 HA 双机系统是用一套软件来完成侦测切换的。这套软件是运行在操作系统之上的，CPU 故障必然导致操作系统瘫痪，所以故障服务器上的 HA 软件此时已无法运行，此时根本不可能管理 CPU 指令和内存数据的同步。

HA 双机容错系统中的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 USB、RS232 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。

首先，HA 双机系统相互侦测是不连续的，间隔一般在秒级，通常是 1-5 秒。

一秒代表什么？对于现在的 CPU，一秒就可以执行数亿条指令。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。这样在故障时，故障主机已执行的数亿条指令是不能在备用服务器上复现的。

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。这种时间天窗对作业连续运行要求极高的应用也是一种灾难。

如果说切换时间还可根据不同的 HA 软件逐步优化，内存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。

3. 换个角度看硬容错服务器

硬容错服务器好在哪里？为什么说它比单机服务器和双机容错发生了质的飞跃？关于这个问题，我们换一个角度去谈反而会看得更明白：硬容错服务器为什么价格较双机系统要高？

上文提过，容错的最高境界是两台服务器中指令同步，内存不丢失。

单机服务器因为没有处理冗余单元，所以也谈不到容错能力。

HA 双机系统因为基于应用软件来解决容错，所以也不可能追踪到 CPU 指令或内存管理。

真正意义上的容错是 100%无间隙的容错，这只有通过硬件完成。

必须开发出一套硬件 ,包括 ASIC (Application Specific Integrated Circuit)芯片组以及配套电路，当然也包括相应的底层软件，使之可以在线监控两套系统的 CPU 和内存的运行。

编一个程序是容易的，几个月就能做出原型；开发一个芯片是困难的，一般都需要数年的时间。看看咱们国家的自主 CPU 芯片“龙芯”艰难的诞生历程就会明白。

更困难的是，当你开发容错芯片时，你还不能闭门造车。

业界流行的 CPU 是 Intel 的，业界流行的操作系统是微软的，容错硬件的开发要涉及 Intel CPU 和微软的 Windows 的诸多技术，这些技术都是非开放的专利。购买专利昂贵的代价是造成容错系统成本增加的主要原因，目前容错服务器在全球销量还远低于普通服务器，造成开发成本摊薄缓慢，这是容错服务器降价缓慢的另一个原因。

日本 NEC 和美国容错（STRATUS）的容错服务器都是由 NEC 公司生产的，其中 GeminiEngine 芯片就是这样的专用的 ASIC 容错芯片，配套以相应控制电路和处理软件，就构成了容错中的核心技术“lockstep”。”lockstep“顾名思义就是逐步锁定，步步跟随，这是在 CPU 指令级别实现的同步容错，真正达到了 CPU 指令流和内存数据的 100%无隙切换。

从成本构成和售价上来说，

单台服务器一般在几千到数万元人民币左右。

HA 双机容错系统是两套单台服务器加一套 HA 容错软件，HA 容错软件十年前就已成型，市场上以美国和台湾产品为主，最好的大致在 1-2 万人民币左右，所以 HA 双机系统的成本构成在 5 万元人民币以内，售价在 8 万元左右。

容错服务器除了两套处理单元（在硬件上相当于两套服务器）之外，还有上文提到的容错芯片，控制电路以及相应的处理软件，价位大致在在十几万到几十万人民币范围。

另外值得一提的是，从软件配置上说，一般容错服务器上的只需配置一套软

件即可，如操作系统，数据库等，而在 HA 双机系统中这些软件都需要双份配置，这样综合比较，HA 双机容错系统在价格上的些微优势几乎已不复存在。

另外再考虑到容错服务器的百分之百自动无隙切换优势，容错服务器比 HA 双机系统的性价比可谓压倒性的。

三． 殊途同归 硬件容错大道通天

久合必分，久分必合。天下三分，三国归晋。

当年的 BP 机去哪儿了？

作为一种移动终端设备，BP 机第一次将人们从有线电话的禁锢中解脱出来，当时真是人手一机风靡一时。

当手机普及后，BP 机也就完成了它的历史使命，很快就退出历史舞台销声匿迹了。 随之而去的还有大哥大、二哥大等通讯终端。

这不是“既生瑜何生亮”的感慨，而是长江后浪推前浪的历史宿命。

BP 机的致命缺点是单向收发，移动通讯的革命不彻底。HA 双机系统不能做到零秒无间隙切换，在容错上也没有将革命进行到底。

手机的生命力来自于它的双向收发以及数字制式。同理，容错服务器的生命力也来自于它对 CPU 指令的同步和内存数据切换的保护。

俗话说王奶奶和玉奶奶只差一点，也就是这一点，容错的可靠性跨越 MTBF 的数万小时和 HA 的一堆 9 一跃登顶为 100%。

量变到质变总是从形式上看好像都很简单，但是找寻到最后一根稻草很难。

从容错性能来说，单机服务器无法和容错服务器相比。

从容错的彻底性来说，HA 双机系统和硬容错服务器是两代产品。

从价格比来说，今天的硬容错服务器还略高于 HA 商机容错系统，但从性能价格比综合考量，硬容错服务器已远高于 HA 双机容错而且硬容错服务器价格会随着普及度进一步趋低。

作为使用者，不管购买什么产品，最重要的是心中有数物有所值。硬容错技术在欧美发达国家已成燎原之势，在国内也正蓬勃发展。在这个时候，花一点点时间了解比较一下硬容错技术和产品应该是值得的。

四． 五湖四海 IT 建设各负其责

IT 工程建设引入第三方监理，这已是大型项目中常见的模式。

五湖：业主、设计院、承包商、制造商和监理公司。

四海：投资、设计、承建，监理。

监理不能流于形式，只浮在流程和文档上，而是要切实对业主负起责任，除了过程管理之外，还要保证设备按合同型号指标到货验收，对培训、验收，保修维修等工作也需要有完备的评测。目前在国内，监理单位多数情况下还处于弱势的一方，这有其历史原因，因为监理模式在我国兴起的时间和欧美相比还很短，有时也不受重视。

本文在开始时谈及的“豆腐卖了肉价钱”的现象，工程监理就是一个杜绝此类问题的重要环节。 业主既然已花了容错服务器的钱，就应该享受容错服务器提供的“零切换”的 100%的可靠性以及“零维护”的方便性。

此时如果以容错服务器的价格换装双机容错系统，则不仅是以次充好以旧代新的经济损失，而且因为双机容错系统其固有的痼疾，在容错切换上存在指令不同步内存数据无法保留的重大缺陷，给今后的实际使用带来了很大的风险。除此以外，因为 HA 双机容错系统在配置上很是繁琐复杂，对切换的作业需要进行设置和脚本编写，在现场维护上也存在很大困难。

而最关键的一点是，既然容错技术已更新换代到容错服务器时代，长江后浪推前浪，就像手机超越了 BP 机，汽车超越了自行车，高新技术的使用和普及是人间正道。监理公司作为独立的工程监管单位，此时更应该高瞻远瞩火眼金睛，不能再在让业主拿买肉的钱去买豆腐，这不仅是经济和技术原因，更是责任使然。