多语种域名及中文域名技术最新进展

中国互联网络信息中心 毛 伟 邓 翔

域名是互联网层次体系结构中识别和定位计算机等网络设备的字符标识，它与 IP 地址相对应。域名技术是不断发展的，在互联网刚刚诞生时，并没有域名技术，在访问计算机资源时直接使用 IP 地址，但由于不便于记忆，在七十年代引入了主机表技术（Hostname）。随着互联网的发展，连入互联网的计算机数量呈现迅速膨胀的趋势，主机表技术的集中式管理模式已经不能满足需求，因此八十年代出现域名技术。现在看来，可以说域名技术是互联网上最成功的技术之一，是互联网的基础性技术，几乎任何一种互联网应用都跟域名有关。随着互联网在非英语国家的迅猛发展，九十年代末期国际互联网界提出了将原本只能使用 63 个 ASCII 字符（“a-z”，“A-Z”，“0-9”，“-”）的域名，采用本地语言文字来表示，也就是出现了对多语种域名的需求。本文将从多语种域名相关组织、多语种域名及中文域名的技术难点和解决方案等方面，全面介绍多语种域名及中文域名技术最新进展。

一、多语种域名相关组织

互联网上最主要的技术标准组织是 IETF（The Internet Engineering Task Force），在 1999 年 11 月，第 46 届 IETF 会议上成立了一个关于多语种域名的兴趣小组（BOF），在 2000 年 3 月，第 47 届 IETF 会议上正式成立了多语种域名工作组（IDN Working Group）来研究制定多语种域名的相关技术标准。

由于中文有一些特殊的需求，在 2000 年 5 月份，海峡两岸四地的互联网信息中心（CNNIC、TWNIC、HKNIC、MONIC）在北京发起成立了中文域名协调联合会（CDNC），CNNIC 工作委员会副主任委员钱华林教授担任主席，成立 CDNC 的目的是协调、研究和解决在多语种域名中面临的中文域名特殊技术问题。由于汉字不只是中国人在用，日本和韩国也在用，所以 CDNC 又联合日本和韩国的互联网信息中心于 2000 年 7 月成立了联合技术工作组（JET，Joint Engineer Team），大家一起来协商讨论域名中汉字方面的问题。

多语种域名不仅涉及技术问题，还涉及管理方面的问题。国际上的域名管理机构是 ICANN（The Internet Corporation for Assigned Names and Numbers），它负

责互联网的 IP 地址空间分配、域名系统管理、协议参数分配、根域名服务器的管理，ICANN 也于 2001 年成立了多语种域名委员会（IDN Committee）来研究多语种域名的管理问题。CNNIC 工作委员会主任委员胡启恒院士是这个委员会的成员。

二、实现多语种域名及中文域名需要解决的技术难点

1、多语种域名系统与 ASCII 域名系统的兼容性问题

在探讨多语种域名的实现过程中，无法回避的一个重要问题是：现行的、主机数量庞大的 ASCII 域名系统是整个互联网的基础，并且它本身稳定、高效、可靠地运行着。对现行 ASCII 域名系统的任何损害，都可能危及整个互联网络，进而导致多语种域名系统技术的彻底失败。

2、多语种域名的编码表示问题

全世界有几百种语言和文字，同一种语言文字可能有多种不同的编码方式(如：汉字编码 GB2312、GB18030、GB13000)。目前 Unicode 成为唯一的一种可供选择的全球统一编码方式，但它是为了解决文字显示问题而设计的，与多语种域名体系设计需求不完全统一。

3、多语种域名的技术实现问题

域名系统包括两个分离的部分：Resolver（解析器）、Server（DNS 服务器），它们分别位于用户操作系统和域名服务器上。由于域名系统在网络应用中的基础性地位，这决定了：让 Resolver(解析器)能够识别多语种域名就意味着要求包括操作系统在内的“所有”网络应用程序必须升级；让 Server（DNS 服务器）能够识别多语种域名就意味着要求“所有”域名服务器都必须升级。这两个“所有”无疑是实现多语种域名时无法回避的技术瓶颈。

4、多语种域名的最终部署问题

现行 ASCII 域名系统在互联网络中应用的巨大成功，成为解决多语种域名问题的重要障碍。对现行域名系统的任何一点修改，对整个互联网络来说，都是巨大的、难以忽视的，甚至是致命的。

由于中文汉字的特殊性，在中文域名的实现过程中，除了上述的技术难点之外，还有一些特殊的问题：

1、中文域名繁简汉字等效问题

《简化字总表》（国家语委 1986）中的繁简体汉字对有 2237 对，经过测试，CNNIC 中文域名数据库中 83.6%的域名都会有这种繁简对照问题。如果包括 1 对多，多对 1 的 238 个高频字以及部分未收入到总表中的繁简字，繁简对照问题就覆盖了中文域名数据库中 90%以上的域名，因此繁简汉字等效是中文域名中的一个普遍问题。

如果不能很好解决繁简汉字等效，用户将不得不注册多个中文域名，在域名系统分层解析过程中也会出现不一致的问题。比如要注册“清华大学”这个中文域名，由于“清”、“华”、“学”字都有繁体字“清”、“華”、“學”，那么用户必须注册 8 个繁简体中文域名，才能真正保护自己的中文域名。这不仅仅是用户经济利益的问题，而且用户可能还将面临无休止的域名争议和法律纠纷。

如果不处理中文域名中的繁体汉字和简体汉字等效问题，中文域名互联网将被分割为两个无法通讯的部分：简体中文域名区域(中国大陆)和繁体中文域名区域(台湾、香港、澳门及海外华人)，不能实现两岸四地互联网在中文域名体系下的互通。

2、中文句号“。”与英文句号“.”在中文域名中的等效问题

域名中采用英文句号“.”作为每一级的分隔符，但在中文中句号是“。”，因此希望“。”和“.”在中文域名中等效，都是中文域名的分隔符号。这可能是中文域名的一个比较独特的需求。

三、多语种域名及中文域名的技术解决方案

一个合适的技术解决方案必须满足下面一些要求：不损害现行 ASCII 域名体系、保障网络整体安全、多语种域名体系必须兼容现有的 ASCII 域名系统。

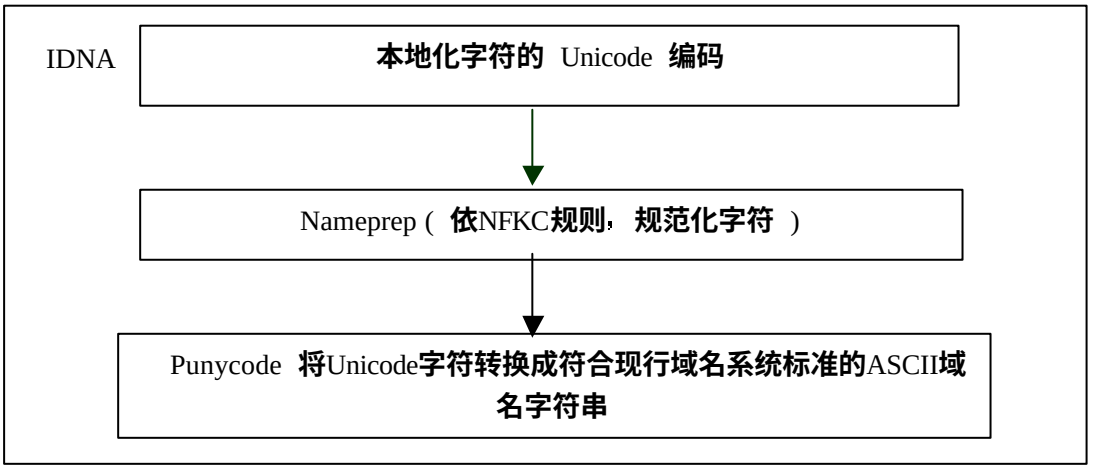
多语种域名的技术解决方案在这三年多来也在不断的发展，先后提出了下面三种解决方案：

1、采用 DNS 扩展技术。扩展 DNS 协议，以特定的数据包格式来区分 ASCII 域名和多语种域名，由于需要全面升级现行的域名服务器和网络应用程序，实施和部署起来难度太大而被否决。

2、采用 UTF-8 编码表示多语种域名。建立新的多语种域名协议，以支持多语种域名的使用，同样由于要求全面升级现行的域名服务器和网络应用程序而被否决。

3、采用 ACE 编码方式。将 Unicode 表示的多语种域名编码成 ASCII 表示的域名，不改变现有 ASCII 域名协议，不损害现行 ASCII 域名体系、兼容现有的 ASCII 域名系统。在实施上只要求升级现行的网络应用程序，即升级位于用户端的域名解析器（Resolver 端）。

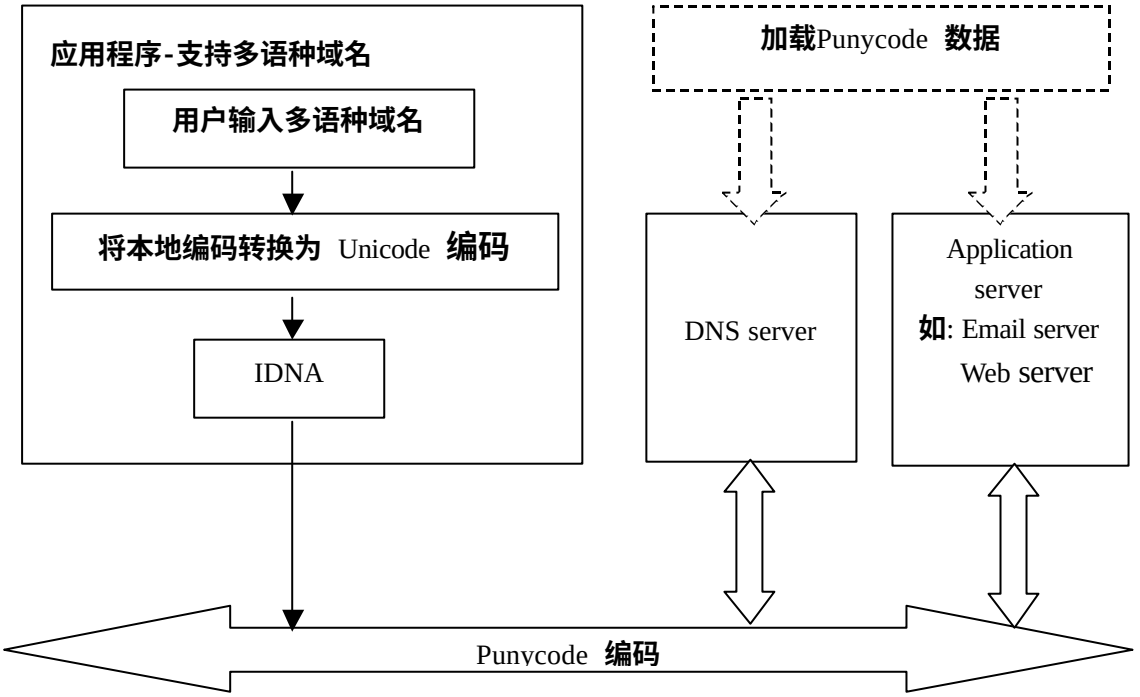
在 2001 年 8 月的第 51 届 IETF 会议上，大家普遍倾向于采用上述的最后一种方案，即 ACE 的编码方案，先把本地编码多语种域名翻译成 Unicode 编码，再通过 Nameprep 处理（它主要做一些规范化的处理，包括：字符映射、字符归一化和字符禁止等）。经过 CDNC 的共同努力，Nameprep 中规定禁止在多语种域名中出现“。”，使在中文域名分隔符中将“。”映射为“.”成为可能。然后再把经过 Nameprep 处理的 Unicode 编码翻译成 ACE 编码，再进行域名解析，上述处理全部在用户端来完成。我们对 IETF IDN 工作组中提出的所有 ACE 编码算法进行了压缩效率的评估（draft-guonian-idn-ace-eval-cn-00.txt），发现 AMC-ACE-Z (Punycode)算法对中文域名的编码效率和编码长度最好。经过我们的努力和协调，AMC-ACE-Z 算法成为 IDN 工作组的推荐 ACE 编码技术方案。多语种域名的协议处理框架（IDNA）如图 1 所示。



（图 1 多语种域名的协议处理框架）

现在整个多语种域名的技术框架就是这样，在用户端，用户输入一个多语种域名，系统将本地编码翻译成 Unicode 编码，通过多语种域名的协议处理框架

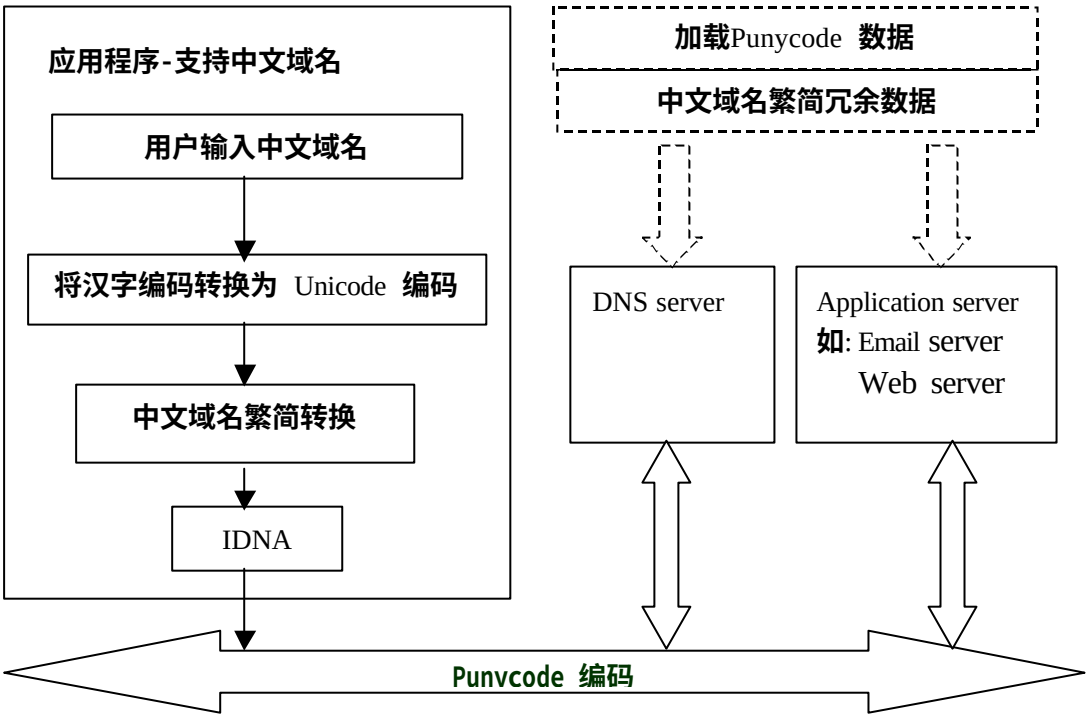
（IDNA），提交到域名服务器进行解析，域名服务器的解析数据库（Zone File）中存储的是 ACE 编码的数据。与域名相关的其它一些应用，比如说 Web Server、Email Server 无需任何修改，但在使用多语种域名数据时，必须先进行 ACE 编码，再作为服务器的加载数据，例如 Server 的名称在系统中并不是多语种的，而是 ACE 编码的 ASCII 字符，这样在使用多语种域名的时候，最好有一些 ACE 转码工具软件支持，使得用户使用这个中文域名或者多语种域名的时候比较方便。多语种域名的总体技术实现框架如图 2 所示。



（图 2 多语种域名的总体技术实现框架）

下面介绍关于中文域名繁简汉字等效的解决方案，解决方案可以有三种，一是在用户端（Client）来对繁简进行转换，另外是考虑在服务器端（Server）来对繁简进行转换，还有一种方法是在进行域名注册的时候来解决，比如说一个域名可以自动给它注册多笔，把繁体的和简体的都同时注册下来。

目前我们倾向的解决方案是采用用户端方案来解决一对一的繁简转换（draft-deng-idn-icdn-00.txt），采用注册的方案来解决一对多等与语意有关的繁简转换（draft-jseng-idn-admin-00.txt）。技术实现框架如图 3 所示。



（图 3 中文域名繁简等效技术实现框架）

目前 CNNIC 联合 CDNC 和 JET 已经先后向 IETF 提交了 7 项多语种域名的技术建议，有些在上面已经有了介绍：draft-guonian-idn-ace-eval-cn-00.txt（第 50 届 IETF 会议）、draft-ietf-idn-tsconv-00.txt（第 50 届 IETF 会议，CDNC 合作提出）、draft-deng-idn-icdn-00.txt、 draft-deng-idn-tsmodule-00.txt、draft-ietf-idn-tsconv-02.txt（第 51 届 IETF 会议，CDNC 合作提出）、draft-ietf-idn-tsconv-03.txt（第 52 届 IETF 会议，CDNC 合作提出）、draft-jseng-idn-admin-00.txt（JET 合作提出）。

经过两年来的讨论，目前 IETF 关于多语种域名的技术标准已经基本上定型，有望近期获得通过，这对多语种域名在全球的应用和普及将会起到极大的推动作用。当然，用户最终要能够使用多语种域名还需要众多软件厂商的支持。

多语种域名除了技术上的问题，还有管理上的问题。在国际上，关于如何管理多语种域名也在积极的进行讨论。我们认为发展多语种域名不仅应该保证当前域名系统的稳定和兼容性，还应当保证各种语言使用者的利益，尊重各地的政策机制，包括政治的、经济的、法律的和文化的等等；多语种域名不仅仅是一个技术问题，更多的是一个管理问题。基于引入多语种域名的基本目的是服务于非英语使用者的需要，因此多语种域名的管理不应当完全控制在商业利益手中，在制定多语种域名的管理政策过程中，应该尊重相应语言使用者的意见；由相关语言使用者组成的合作组织应当在相应的多语种域名管理中发挥主要作用。我们把相应的观点向 ICANN 的多语种域名工作委员会（IDN Committee）做了提交，最后在 ICANN 的报告里面也反映出这些观点。

经信息产业部批准 CNNIC 从 2000 年开始提供中文域名注册试验服务，目前在注册数据库中大约有 50 万条记录，作为试验性的服务，CNNIC 没有向用户收费。中文域名在注册体系上采用了与国际接轨的管理模式，分为三级管理：注册管理机构、注册服务机构和注册争议解决机构。可以预见，随着多语种域名在管理和技术上等一系列问题的解决，中文域名在我国也必将得到进一步的发展。