IC 设计中的 IT 环境(2013)

--王光辉

前言

几年前，我写过一篇文章，介绍如何在 IC 设计中选择 IT 设备和配置环境。那时候算是年少冲动吧，写的那些东西都是实际工作中遇到的，几年来，也有很多做这行的朋友看了文章后，找我咨询问题，我也尽量解答。

不过，几年过去了，IT 设备更新换代，自己的经历也更加丰富，总想提笔写一些更新，无奈杂事太多，每每无法成文。我有更新的冲动，但是也有怕误人的谨慎。在多家公司呆过之后，我发现每家公司都有自己的特点，没有经过仔细的了解，而做出的方案，基本上在后期改动的可能性很大，所以更加不敢写。我做这方面工作已经快 10 个年头了，自己思维也开始有一些老化，如果再不把自己的一些经验写出来，以后我想更难完成。

记得上一次写这方面的文章还是 2009 年初，一晃 3 年过去了，经历更多，写文章的动力却少了很多。原因无外乎生活杂事让自己无法静心。也怕到时候写的东西发出去后，被各种咨询打扰。我一向乐于 share 自己的经验，也愿意解答大家遇到的疑问，但是我很不喜欢被一些 google 都可以找到答案的问题烦恼。如果你自己不愿意去找问题答案，最简单的一个办法：给钱让人帮你解决。找不到人解决的时候，才是找人来咨询。我曾经在 QQ 上遇到过一个人，自己是某家 IC 设计公司的 IT，水平怎么样不好说，但是特别喜欢遇到问题就来问我，有次实在烦了，就说了一句，自己去 google。然后，他就说我态度不好，我真想说一句：我帮你解决问题，我有义务吗？

如果是教育机构，比如高校，需要这方面帮助的，我愿意无偿帮忙。如果是公司，你需要简单某个问题的解决，我可以电话协助你，如果需要做整体方案，不建议电话咨询，建议你花钱找人解决，如果别人解决不了，也可以找我。我想强调，这是一个整体的方案，不是 1-2 天时间就可以完全解决所有问题的，更不是电话就可以一下说清楚。我会尽量将我的经验分享在这篇文章中。

提前发一通牢骚，希望到时候能少一些基本问题的咨询。

建议大家加入 edacad QQ 群：292489873

我的联系方式:QQ 58648217　手机：13606212363

我的工作经历：

2004-2006：苏州集成电路设计中心，系统管理

2006-2009：盛科网络（苏州）有限公司 IT 管理

2009-2011：Broadcom　System Analyst

2011-now:　苏州一家 IC 设计公司

我的前一篇文章主要在盛科网络工作期间完成，和当时的情况类似，IT 环境比较适合当时的公司。后来，我去 Broadcom 工作后，我才发现，原来人家大公司的做法不太一样。理解了一个成熟的公司，IT 环境应该如何去设计。回到苏州后，我将现在这家公司的很多 IT 环境逐步往类似 Broadcom 的解决方案靠近，不过进展一直都比较缓慢，主要原因是这家公司积习难改。从此事上，可发现有一些成立时间长，人员比较多的 IC 设计公司，要做好思想准

备，改变以前的环境有很多困难。

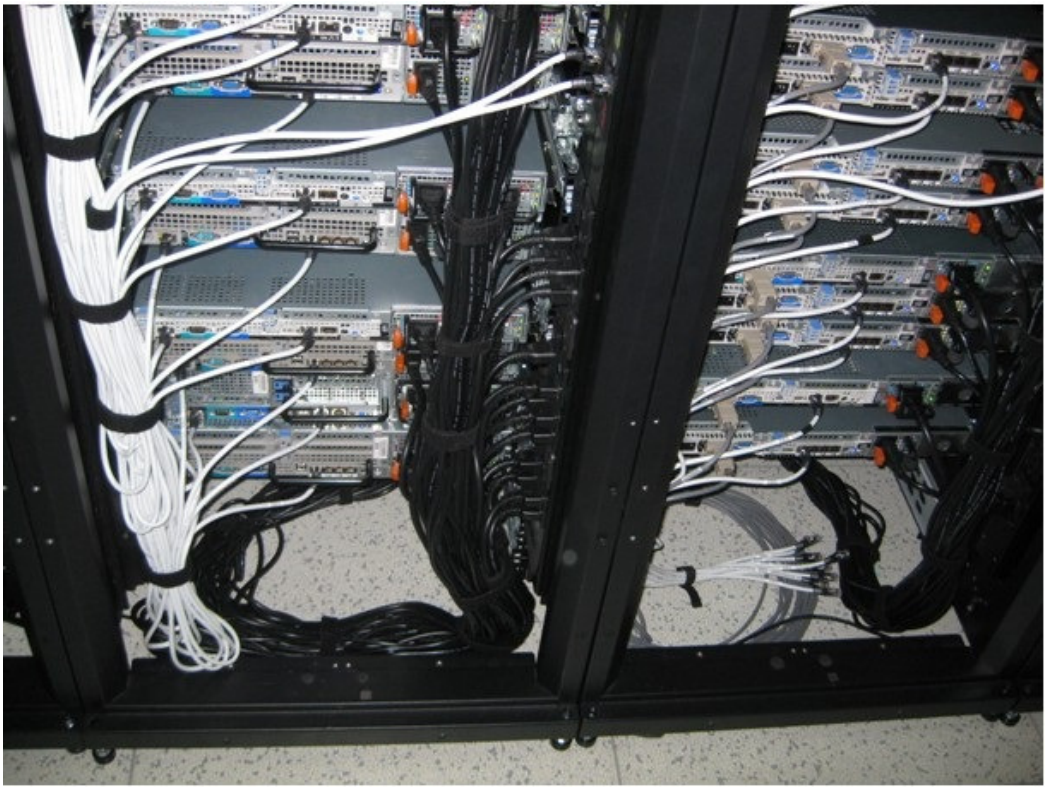
由于我现在主要集中在 IT 系统方面，所以我不过多介绍软件的一些事情，但是我尽量将 IT 整个环境，包括网络，存储，备份，EDA 设计系统等方面介绍到，给大家多一些参考，大家一起多讨论吧。目前的 IC 设计公司，大部分都是 C/S 架构,也就是说客户端 PC 远程连接到服务器使用,用户身份认证使用 NIS,配合 nfs 和 autofs 来自动挂载目录。

1. 机房基础设施（Infrastructure）
2. 网络及结构设计
3. 服务器的选择及配置
4. 操作系统
5. 存储设备的选择
6. 远程 VPN 访问
7. 集群环境搭建以及 EDA 软件加速功能
8. 版本控制及 bug 管理(CVS/SVN/VSS/Bugzilla)
9. 服务器访问方式 VNC/FreeNX/Citrix
10. 虚拟环境 vsphere
11. NFS/NIS 结合
12. 数据备份
13. FPGA 下载数据如何做到安全及归档
14. 服务器硬件的远程管理

1. 机房基础设施

一个好的 IT 环境应该是从机房 网线 电源等基础设施开始的，我见过很多公司的机房,有做得非常漂亮的，也有问题非常严重的。对于问题很严重的一些机房，要想去解决那些问题，将是非常非常困难的。我们做家庭装修都知道一点：水电无论如何不能用差的，因为一旦以后有问题，整修起来非常困难。公司的 IT 基础设置也一样，一定要在规划的时候就考虑进入，否则以后问题一大堆。

好的机房标准是什么？这可能是我们很多人都有的疑惑，每个人都有自己的理解，但是我想离不开几点：可管理 可扩展 整洁 恒温 标识清楚。



机房的基础设备主要包含：UPS 及电池，电源分配设备，空调，机柜等。

首先，我们来讲一下 UPS 及电池。为什么要使用 UPS？因为我们无法预知意外的停电事故，使用 UPS 避免服务器的异常关闭。我们购买第一台 UPS 的时候，在功率选择方面，起码要预测到最近 1 年以内可能增加的负载有多少，从而选择适当的 UPS 设备。UPS 标称功率一般只能使用到 80%。由于服务器 idle 的功率和满载功率完全不一样，所以，请做预算的时候，一定要使用服务器的满载功率，这个测量可以使用一个很简单的功率测试仪完成（淘宝上有很多，比如万方的小设备）。如果你需要测试整个已有机房的功率，可以在负载比较高的时候，使用功率钳测试火线得到安培数，然后计算出功率。

同时，由于电池容量和负载，待机时间三者之间的关系，我们必须把握好到底需要多少电池。电池是具有腐蚀性硫酸的东西，一旦泄漏可能给机房带来安全隐患，所以要注意选择正规的电池。UPS 市场上也有很多假货，所以选择正规的厂商是绝对有必要的。我买过三特电子的 UPS，结果发现市场上很多所谓山特都是假货。

其次，机房还得注意插座的分布。一般我们需要在 UPS 出来的 Output 线缆处安装一个多路的空气开关，然后每个空气开关对应机房机柜下面的 1-2 个插座，每个插座可以带多少台服务器，请注意控制。如果预算够多，可以使用 PDU 电源桥安装在机柜后面，如果没有那么多预算，只能使用插线板或者其他方式扩展了。如果有多个 UPS，考虑到冗余情况，请记住每个机柜后面的插座需要来自不同的 UPS。注意机房的布线至少考虑 4 平方以上的铜线，否则可能由于负载过高，导致线缆发热量大而引起火灾。

再次，空调的选择。空调在这里主要是制冷，所以绝对要安装独立的空调。有很多专业机房空调可供选择，那些空调一般带有恒温恒湿和除尘功能，对于小机房预算有限的情况下，显然我们不大可能考虑专业空调。但是，即使对于小型机房，你也必须考虑两点：第一，制冷量是否足够；第二，空调如果发生冷凝水泄漏，是否会给服务器带来灾难。特别是吸顶式的空调，我在 2004 年就遇到过空调冷凝水泄露，导致整个机房防静电地板上全是积水。

最后，机柜选择。一般，我们会选择 42U 的标准服务器机柜，或者特定的服务器供应商机柜，如 Dell 就有自己的机柜，这种机柜对于安装 Dell 的服务器更加方便。机柜宽度一般是 600mm，深度对于服务器来说，一般都大于 1000mm，而高度规格有 42U 及以下，比如 37U 32U 26U 22U 等。机柜的质量对于后期安装服务器有很大关系，所以绝对不要买杂牌的机柜，毕竟这个是承载服务器用的，如果质量不好，容易变形。

2. 网络及结构设计

IC 设计公司的网络其实和其他公司没太大差别，除了可能需要分内外网之外。这里的内外网络之分，不等于普通公司的内外网。一般 IC 设计公司的内网是一个更加保密的网络，除了允许的登录，禁止其他任何登录，一般只允许研发部门登录。同时，内网也禁止连接 internet。而外网，一般指的是办公网络，这个网络包括了人事、财务、行政、市场、销售等部门。

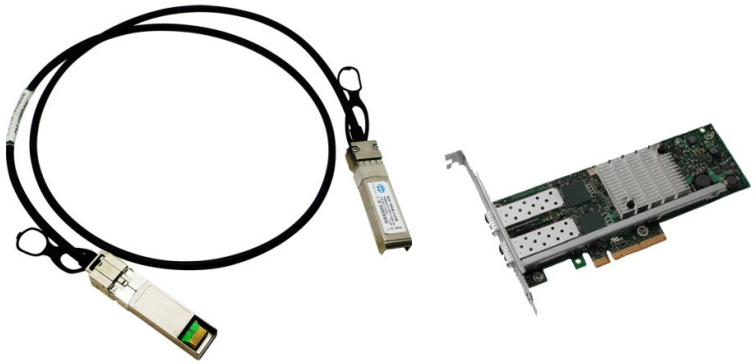
一般，我们通过划分 vlan 来实现，同时通过 3 层交换的 ACL 实现 vlan 之间的访问控制。当然，连接 internet，我们需要防火墙。防火墙的设置和其他普通公司没什么差别，我这里将不再详细的介绍。

对于做 IC 设计的内部网络，我们需要注意几个方面：

首先，服务器之间千兆及千兆以上连接。我在多个地方发现，很多公司依然在使用百兆网络连接服务器。这样的网络会使服务器之间形成孤岛，服务器资源无法共享，每台服务器承担各自独立的任务，任何一台服务器故障都将让一部分人无法工作。如果希望使用统一的存储空间，无比采用千兆及以上网络连接。

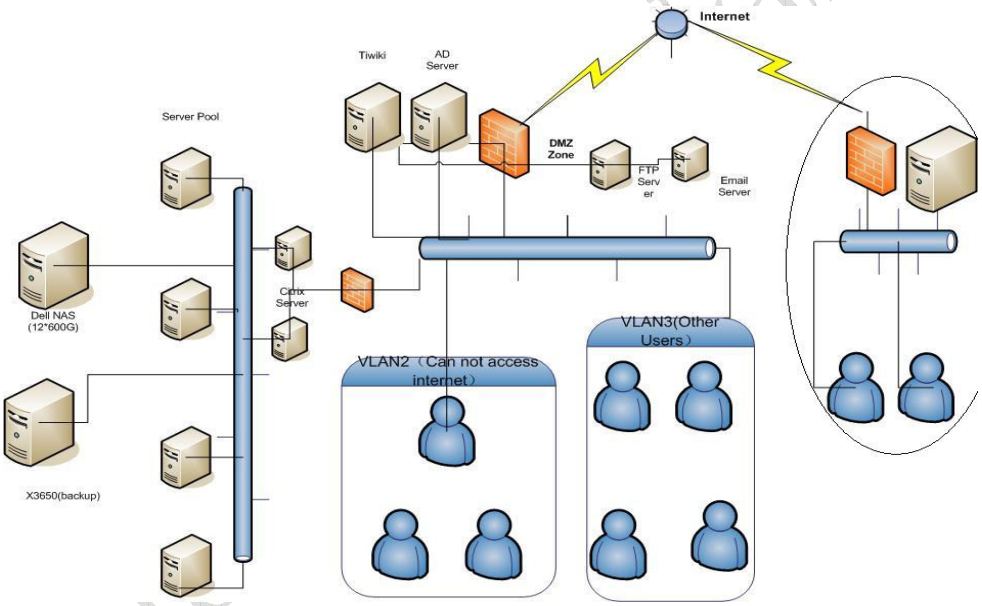
其次，存储服务器使用万兆或者多个千兆捆绑。如果系统内有专用的存储服务器，特别是在服务器互相之间 IO 数据很频繁的情况下，推荐使用万兆网络连接存储服务器，或者采用多个千兆捆绑来增加带宽。为什么我们对存储服务器特殊考虑？因为我们的存储需要为多台计算服务器提供存储功能，当多个计算节点同时读写的时候，存储服务器如果依然是千兆网口，显然会是一个瓶颈。

如果存储服务器采用了万兆网卡，那么请将服务器互联的交换机升级为带有 10G 上联端口的交换机。建议对新购交换机的，都采用 24 个千兆+2 个 10G 或者 48 千兆+4 个 10G 这样配置的交换机，10G 接口建议 SFP+，因为目前看来 SFP+ DAC 是短距离万兆的最合理方式。万兆连接可以采用 Direct Attach 线缆，成本低，延迟小。如果希望购买全 10G 端口的交换机，建议选择 10G Base-T 接口的交换机。



最后，对需要多个 vlan 的内部网络进行 ACL 控制。基于安全考虑，服务器和 Client 网络之间需要控制，那么基于 ACL 来对 vlan 之间访问控制无疑是最佳的方式。

下图是 IC 设计内部网络结构示意图：



从图中，我们可以看到，即使是 IC 设计的内部人员，他们也只能通过登录 app 服务器来使用内部资源，从而避免 IC 设计的内部人员带走服务器上的资源。而如何只允许内部各个部门的设计人员只访问 app 服务器，是通过 vlan 和 acl 实现的。

网络控制遵循的原则是：除了允许的，其他全部禁止。保证用户无法直接接触到所有设计数据。

3. 服务器的选择和配置

我在多个场合推荐过 Dell 的服务器。这是因为，Dell 服务器的硬件配置可以很方便的自我控制，同时其价格比较低廉，可以达到最好的性价比平衡，最主要的是，Dell 服务器可以做到硬件有故障的情况下 4 小时上门服务。

那么，在目前（2013 年第一季度）我会推荐采取何种服务器型号和配置呢？在这里，我主要介绍一下最近配置过的服务器。Dell R720：CPU 我选择了 2 颗 E-2690 CPU,内存方面配置建议配置 64G 以上，硬盘看用途，如果同时做存储数据使用，推荐配置一张 H710p 的 RAID 卡，同时，建议配置上 iDRAC 7 enterprise 卡用于远程管理。为什么我会推荐配 R720，我发现这台服务器最大可以配置到 24 个内存插槽，同时有 8*3.5 寸的硬盘位可供扩展。因为我们主要做计算使用，所以配置了高频率，8 核心的 E-2690CPU。我在其他公司发现，他们配置的服务器 CPU 只有一颗，内存只有可怜的 4G，这样的配置价格是很便宜，但是性价比却非常低，特别是内存，当前的内存价格已经非常便宜了，如果在计算的时候，内存足够大，是可以提供更快的计算速度的。

Dell 的 H710p 卡是一张有 1G 缓存的 raid 卡，OEM 自 LSI，其具有 Cache Cade 功能提供，可以采用 SSD 来实现更快的读访问能力。

而其 iDRAC 7 enterprise 卡，可以提供远程的 console 访问功能。只要你连接到服务器最左边的 idrac 口，即可通过 192.168.0.120 来访问，其默认用户名为 root，密码为 calvin。我们可以远程查看服务器状态，远程开关机，远程安装 OS，远程登录服务器图形界面，就和你在机房服务器旁连接一台显示器和键盘完全一样的操作。

不建议采用杂牌服务器，但是对于需要更高性价比，而对稳定性没那么高要求的用户，建议的配置方法是：购买低配置的 R720 或 R620，自己购买内存和 CPU，硬盘来增强其性能。这样可以省下很多费用，但是稳定性差一些，同时出故障后，保修方面会有一些麻烦。

服务器尽量采用同一个型号和品牌，以便管理和维护。这是我在 Broadcom 的时候，学到的非常重要的知识—标准化。

4. 操作系统

我在 2008-2009 年之间写的那篇文章中，我就强调过，操作系统选择非常重要。每个 EDA Vendor 对于其软件都会有一个操作系统的支持要求。当时，使用 RHEL3 和 RHEL4 是比较合适的选择。但是，现在我们需要使用 RHEL5U9 这样的 OS 才是比较合理的了。

这里我们要注意一个原则：OS 一定要选择最新update 的，EDA 软件要能支持这个平台。EDA 软件一般选择较新的版本，最新的可能引入了某些bug，所以谨慎使用，但是新版本一般会提升性能和加强功能。对于软件版本的选择，确实是一个非常头疼的问题，即便是非常熟悉的人员，他也得考虑你的环境才能给出建议，所以最好做好这方面的咨询。

其他的 Linux 发行版大部分情况下也是可以支持，但是一旦存在问题的时候，我们解决起来就会很麻烦。所以，我强烈推荐redhat enterprise。当然，centos 也是可以使用的，只是，某些软件会去检查是否是redhat，比如vcs 会检查/etc/redhat-release 文件，如果没有找到redhat字样，就会报错，还会报告编译方面的错误，让你花费大量时间去检查编译器和库文件问题，结果根本不是那里的问题。Redhat 的enterprise 是可以从网上download 的，只是没有购买软件授权而没有技术支持而已。

RHEL5 有 x86_64 和 i386 两个版本，我们需要选择 64bit 的版本，因为当前很多 EDA 软件使用的内存都远超过 4G 的容量限制了，特别是后端的工具，随着规模的扩大，内存容量要求也呈现出成倍增加的态势。

RHEL5 安装过程中，已经无法选择 everything 的选项了，所以我们安装过程中，尽量选择必要的软件包，特别是 development 方面的软件包非常有必要。

5. 存储设备的选择

在我所有的工作经历中，使用过好几种存储方式。Netapp 是我最满意的存储设备，但

是其性价比不是小公司所能承受的，在 2011 年之前，我无法找到一个最好的存储提供给国内大部分公司。谈到存储，很多人往往会陷入 NAS 和 SAN 的争吵中去，我感觉这种争吵完全是在浪费时间。无论你买多么高端的存储，如果你没配置和使用好，性能不会比低端的好多少。最简单的一个例子是，买一个 netapp 6200 系列，配置 14 块 SATA 磁盘和一台 netapp 2000 系列，但是配置 4*14 块 FC 盘到底谁快？但是，现在存储的规则正在被 SSD 给改变。希望各位不要过分看重某些东西，最好是实际测试一下，然后满足自己要求的就是好的存储。

在很多微型的 IC 设计公司不存在这个人问题，因为他们根本不需要考虑专门的存储设备，而直接采用了 local storage。对于微型的 IC 设计公司（数据量小于 2T，设计人员不超过 10 人），建议考虑我前面推荐的 R720+H710p raid 卡，然后搭配 4-8 块硬盘组成 raid6 即可。

而对于稍微上规模的 IC 设计公司来说，这就是一个大问题了。对于存储的容量比较大，设计人员较多的 IC 设计公司，强烈建议考虑专门的存储设备来解决数据存放的问题。

全世界 70%以上的 IC 设计大公司都采用了 netapp 的 NAS 设备。我推荐如果预算充足的情况下，尽量采用 netapp 的 NAS 设备来作为公司的主存储。

采用 netapp 存储的好处主要有：

a. snapshot 可以帮助用户自行恢复不小心删除的数据

b. COW 技术让写入的随机数据变得不随机，更像连续数据写入，提高了写入的 IOPS。

c. SIS 技术可以让数据存在多个 copy 的情况下，只存储一份 copy。从而节省存储空间。

d. Compression 技术，可以让数据在写入存储的时候被压缩存储，节省空间的同时，提高了 IO 性能。

e. 方便的 quota 技术，可以基于 user 和 group，也可以基于 qtree（类似目录）提供数据的容量管理。

f. 同时提供 NFS CIFS 和 iSCSI 访问。

目前配置一台低端的 netapp 2220，12×2T 配置的存储，价格大约 12 万多 RMB。这个价格对于国内很多 IC 设计公司来说，感觉还是难以承受。那么，在技术支持能力足够的情况下，我推荐采用基于 ZFS 的存储。

我很喜欢 netapp 的存储，尽管其有各种限制，价格也很昂贵。但是，作为管理方便来说，它绝对值这个价。但问题就是，国内的 IC 设计公司有多少会花钱去买它的各种高级 feature 来用？所以，我一直在寻求替代品。要求满足几个基本条件：带有方便的 snapshot 功能；quota 可以到 project 和 user group 同时存在；支持高级 ACL；支持 SSD；稳定和方便。

经过以上条件的筛选，我最终只找到一个答案：ZFS 文件系统。其不止带来那么多高级 feature，还多了一项可以降低 50%存储容量的功能—compression。

ZFS 是 Sun 公司开发的文件系统，其提供的功能优点类似 netapp 的 WALF 文件系统，但是在 Sun 公司被 Oracle 收购前，zfs 是基于 CDDL 开源的。我们可以使用差不多一半的价格，得到同样容量和同样功能的存储服务器，但是这要求自己有很熟悉 solaris 系统的 IT 人员。

同时，因为 zfs 具有使用 SSD 来做 L2ARC 的功能，我们可以采购一块 512G 的 MLC 的 SSD 硬盘，作为大量访问数据的缓存，提供类似 H710p RAID 卡的 Cache Cade 功能。

而对于 IO 要求特别高，特别是小文件写入要求高，比如每秒写入 500MB 以上的环境，可以考虑 LSI 的 Cache Cade 2.0 方案，或者 Adaptec 的 MaxIO 3.0 方案，提供基于 SSD 的 flash pool，read and write 缓存的功能。当然，目前 netapp 也提供了 flash pool 方案，比如可以配置 netapp FAS 2240，6×100GSSD+18×1T SATA 的方案，大约需要 30 多万 RMB。

目前 SSD 方面，如果需要写入缓存的方案，比如 LSI Cache 2.0 和 Adaptec MaxIO 3.0 中，可以选择的最具有性价比的是 Intel 的 DCS3700 系列 eMLC 的 SSD。

而对于只读缓存 SSD 来说，可以选择的就太多了，比如镁光的 M4，Intel 的 330 和 520 系列。

我来解释一下我要求的几个功能都用在什么地方：

Snapshot：我采用 snapshot 将所有的 home 和 project 都做了几个自动的 snapshot。每天晚上做一个 daily_snapshot,保留 2 个备份；白天每 4 小时做一次 snapshot，同样也保留两份。User 可以在任何时候自己动手找回 2 天内删除的数据。

看一下我做的 snapshot 的截图，这一切都是自动完成的，截图时间是 2 月 17 日上午 9 点 15。可以看到，其中保存了两个 snapshot，是 2.16 和 2.17 的晚上零时。另外 4 个是 2.17 当天隔 2 小时的 snapshot，一共 4 份，自动循环。

```
bash-3.2# cd /home/.zfs
bash-3.2# ls
snapshot
bash-3.2# cd snapshot/
bash-3.2# ls
zfs-auto-snap_daily-2012-02-16-0000     zfs-auto-snap_frequent-2012-02-17-0400
zfs-auto-snap_daily-2012-02-17-0000     zfs-auto-snap_frequent-2012-02-17-0600
zfs-auto-snap_frequent-2012-02-17-0200  zfs-auto-snap_frequent-2012-02-17-0800
bash-3.2#
```

Quota：我们经常陷入项目管理的混乱中去，linux ext3 采用的 user 和 group 的 quota 根本不能满足要求。基于 project 的 quota 让我们可以控制某个 project 只允许使用多少空间。我们再也不需要天天发 email 要求用户删除不用的数据了，用户和 project leader 会自己主动去做这个事情。

```
bash-3.2# zfs userspace zdata/home |grep 10G
POSIX User  f          4.06G    10G
POSIX User  hw         3.45G    10G
POSIX User  jl           9G     10G
POSIX User  jc         3.56G    10G
POSIX User  ll         1.50K    10G
POSIX User  mf           1G     10G
POSIX User  pf         7.51G    10G
POSIX User  sz         5.43G    10G
POSIX User  tl          423M    10G
POSIX User  wj         9.50G    10G
POSIX User  wt         9.23G    10G
POSIX User  xm         8.34G    10G
POSIX User  xm           4G     10G
POSIX User  xx         9.76G    10G
POSIX User  yh         4.66G    10G
POSIX User  zf           1G     10G
POSIX User  zh         2.81G    10G
POSIX User  zk         9.05G    10G
POSIX User  zs i          1G    10G
bash-3.2#
```

高级 ACL：我们需要对某些项目进行控制，项目组成员可能来自不同的部门，随时需要添加和删除，基于 UGO 模型是不行的，无论你是否承认，也许你正在用这个模式，但是你会发现很多限制。ZFS 的 acl 是基于 nfsv4 style 的，而不是传统的 unix 文件系统 acl。

```
bash-3.2# cd /project/
bash-3.2# ls -dV ppc476
drwx------+ 12 jlzhu    design        13 Dec 23 12:19 ppc476
         user:                :rwxp-DaARWc--s:------:allow
      user:df                 g:rwxp-DaARWc--s:------:allow
        user:s                :rwxp-DaARWc--s:------:allow
          user:               :rwxp-DaARWc--s:------:allow
        user:                 g:rwxp-DaARWc--s:------:allow
        user:z                n:rwxp-DaARWc--s:------:allow
         user:                i:rwxp-DaARWc--s:------:allow
          user:               i:rwxp-DaARWc--s:------:allow
          user:               i:rwxp-DaARWc--s:------:allow
      user:xm                 g:rwxp-DaARWc--s:------:allow
        user:x                n:rwxp-DaARWc--s:------:allow
      user:                   g:rwxp-DaARWc--s:------:allow
         user:f               n:rwxp-DaARWc--s:------:allow
        user:jl               u:rwxp-DaARWc--s:------:allow
            own                :rwxp--aARWcCos:------:allow
           group@:------a-R-c--s:------:allow
        everyone@:------a-R-c--s:------:allow
bash-3.2#
```

支持 SSD：主要是现在的磁盘实在太慢，无论如何都是不能支持到几千几万 IOPS 的，只有结合 SSD 来做缓存才能对存储是最大的帮助。在有大量读操作的时候，可以使用多个 SSD 作为 L2ARC 的缓存设备，这样读过一次的数据会先保存在 SSD 中，下次读的时候，就不会去通过磁盘的旋转来读取数据了，显然提高了读取效率。

自动压缩：文件系统支持自动压缩的功能，会节省大量的磁盘空间，更大的好处是 IO 变得更快了。重复数据删除功能也能节省大量存储空间，但是目前 zpool v30 还不支持，DEDUP 对 CPU 使用也非常厉害，建议考虑在备份的时候采用。

请看一下我们 home 空间的自动压缩情况，压缩比是 1.82。

```
bash-3.2# zfs get all zdata/home
NAME         PROPERTY                       VALUE
        SOURCE
zdata/home   type                           filesystem
        -
zdata/home   creation                       Wed Sep 21  0:50 2011
        -
zdata/home   used                           581G
        -
zdata/home   available                      219G
        -
zdata/home   referenced                     574G
        -
zdata/home   compressratio                  1.82x
        -
```

稳定和方便：如果你的存储出问题，影响所有的人，你认为有啥比稳定更重要呢？从上图可以看到我们这个文件系统已经使用了 5 个月，没有过任何问题。

如果你的环境要求 IOPS 很高，请用 SSD。如果你的存储都是大文件，请用磁盘+SSD，适当将大文件和小文件分开存储，对你有好处。

存储协议建议采用 NFS，这是目前几十台服务器中最好的方式。一定计算好自己的 IO 到底有多大。比如你有 20 台服务器都会每秒钟写 1Gbps 的文件，无论你的盘阵多强大，你的 NAS 服务器网络如果还是 10G 的，依然不能满足要求。我记得 edacad group 有人就是类

似的情况，服务器配置很高，存储配置很烂。。。瓶颈很明显。

我也听过说采用 GPFS 这种 SAN 架构的文件系统来做的，但是这个陷入不适合大部分用户。你需要仔细评估自己的需求。

我的存储服务器配置：

DellR520(12*600G SAS) + MD1200(12*600G),各自通过 H700 和 H810 连接起来，采用 zfs 文件系统。做了两个 zpool，每个 pool 的 IO 小文件通过 dd if=/dev/zero of=testfile bs=4k count=2500000 得到的结果是 300MB/s 以上。

6. 远程 vpn 访问

为什么我们会涉及到远程 VPN 范围呢？

尽管我们一再强调，我们的 IC 设计环境会和 internet 隔离，但是我们却无法避免，在很多公司，我们会希望有一些安全的接入。比如，我们有 IC 设计人员需要出差，需要远程登录；我们有合作伙伴，我们需要给对方提供一个安全接入方式，在保证自身数据安全的情况下，让对方能和我们一起工作；我们有分部，对分部可能无法完全控制，无法在分部建设一个 DataCenter，我们需要他们也能安全的使用总部资源。VPN 的目标是给远程用户提供一个安全方便的接入方式。

目前 VPN 接入的方式主要有 3 种：SSL-VPN/PPTP VPN/IPSEC VPN。这三种方式中，第一种 SSLVPN 是最方便的方式，但是偶尔会出现兼容性问题。第二种 PPTP 对于 windows 来说，不需要安装任何额外软件，只要简单的创建拨号连接即可。第三种方式，一般会要求安装一个客户端软件。

很显然，我们会首选 SSL VPN，因为这种方式一般不受端口和防火墙限制。目前市场上提供 SSL VPN 接入的厂商有很多，我这里推荐采用 Fortigate 的防火墙，它同时提供了以上的三种 VPN 接入方式。

我这里需要强调一定：无论采用何种 VPN 接入方式，一定要注意，只开放必须的内部资源给远程用户，默认关闭其他资源。

目前国内的网络存在联通和电信的差别，北方是联通的天下，而南方又是电信的天下。

所以，同时接入联通和电信才能保证公司的人员在各地都能高效接入。VPN 的接入不能只通过密码认证，最好还要有其他安全手段，比如 RSA SecurID，USB Key 等。

7. 集群环境搭建及 EDA 软件加速

大部分的外资 IC 设计公司使用 LSF 来作为集群环境的管理软件。LSF 对于 IC 设计环境来说，确实很不错，可以说是 IC 集群环境中使用的最多的集群管理软件。但是，这个软件的授权价格费用很高，一般的小公司承受不了，对于初创公司，更是一笔很大的开支。不过可以用一些开源的软件，基本能满足要求。

我在这里需要提醒大家一点：集群并不是搭建好环境后，你的 EDA 软件就能成几倍的并行增速的，这种观点在很多人的脑子里边形成固定的错误了。集群只是提供一个可以并行的基础结构，EDA 软件是否可以并行，取决于软件自身的编程。而具体到 EDA 软件上来说，就是有部分软件的部分功能可以被并行，比如 hspice 这个著名的 spice 仿真软件，我们可以

通过 hspice –mt N，这样来指定，同时 N 个 CPU Cores 一起仿真，从而提高速度。

目前可以提供并行的 eda 软件有很多，特别是后端的很多软件可以实现并行，比如 DC Calibre PT Conformal ICC 等。具体各个软件如何使用和设置，请查询文档。如可能，我也会写一些软件的并行设置介绍。

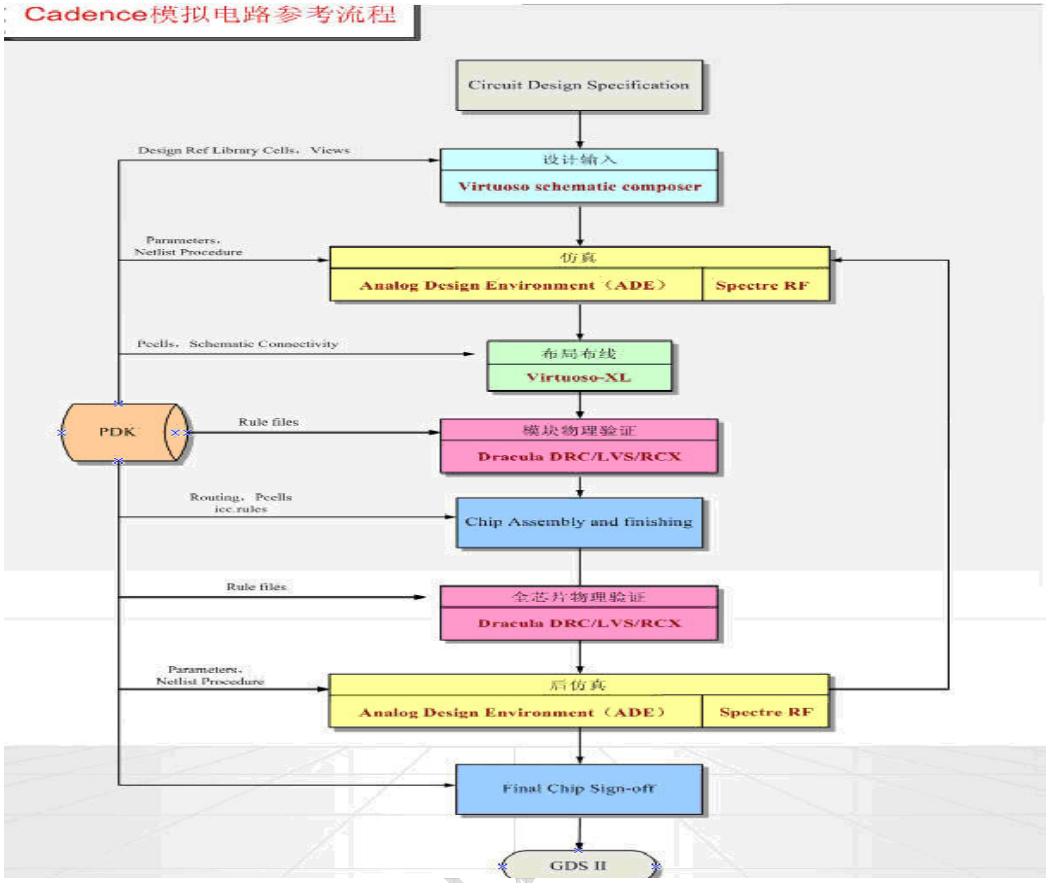
LSF 是我见过的做得最好的集群管理，当然你可以说还有 SGE。我使 LSF 用这个工具有好几年了，确实有很多优点。使用这个工具，最好是所有人都能理解这种工作模式：job 只要提交，然后就等出结果就可以了，无需去关注在哪儿跑。

我刚进入现在工作的这家公司的时候，没有任何负载管理的工具。大部分人都是随便选择一台服务器，开启一个 vnc，然后直接 run job。平时都还好，一旦遇到项目忙的时候，一台只有 4 个 CPU core 的服务器，可能同时跑 10 个 job，然后所有人敲一个 ls 都要等半天才有反应。部门经理问我为什么，我的回答就是太多 job 同时 run，而他完全无法控制。在引入 LSF 之后，我们的一切都改变了，所有的 job 都在后台，自动分配资源。服务器的利用率也有很大的提升。因为还可以对一些后端的 job 进行并行计算，导致效率提升更大。有人问我 LSF 的好处，我想了很久，也没想到好处到底有多少。我只是想说，服务器利用率高了很多，计算效率高了很多，毕竟大部分 IC 设计大公司都在采用它。你可以发现 TOP500 的超级计算机也在用它。

下面是我以前写过的关于 IC 设计中使用集群的作用，原文如下：

我们做 IC 设计中，新购买的服务器有越来越多的 CPU Core，设计也越来越复杂。很多时候，我们的服务器 CPU 利用率很低，一台 16 cores 的服务器很多时候只有 1 个 Core 被利用上了，而其他大部分闲置。另一方面，我们的设计也越来越复杂，要求我们尽快完成仿真，综合，布线等工作。有没有办法加速这一流程呢？那就是我们将计算服务器集群化。

首先，我们看一下设计流程。这里以 cadence 的工具设计模拟电路以及 synopsys 工具设计数字电路为例。



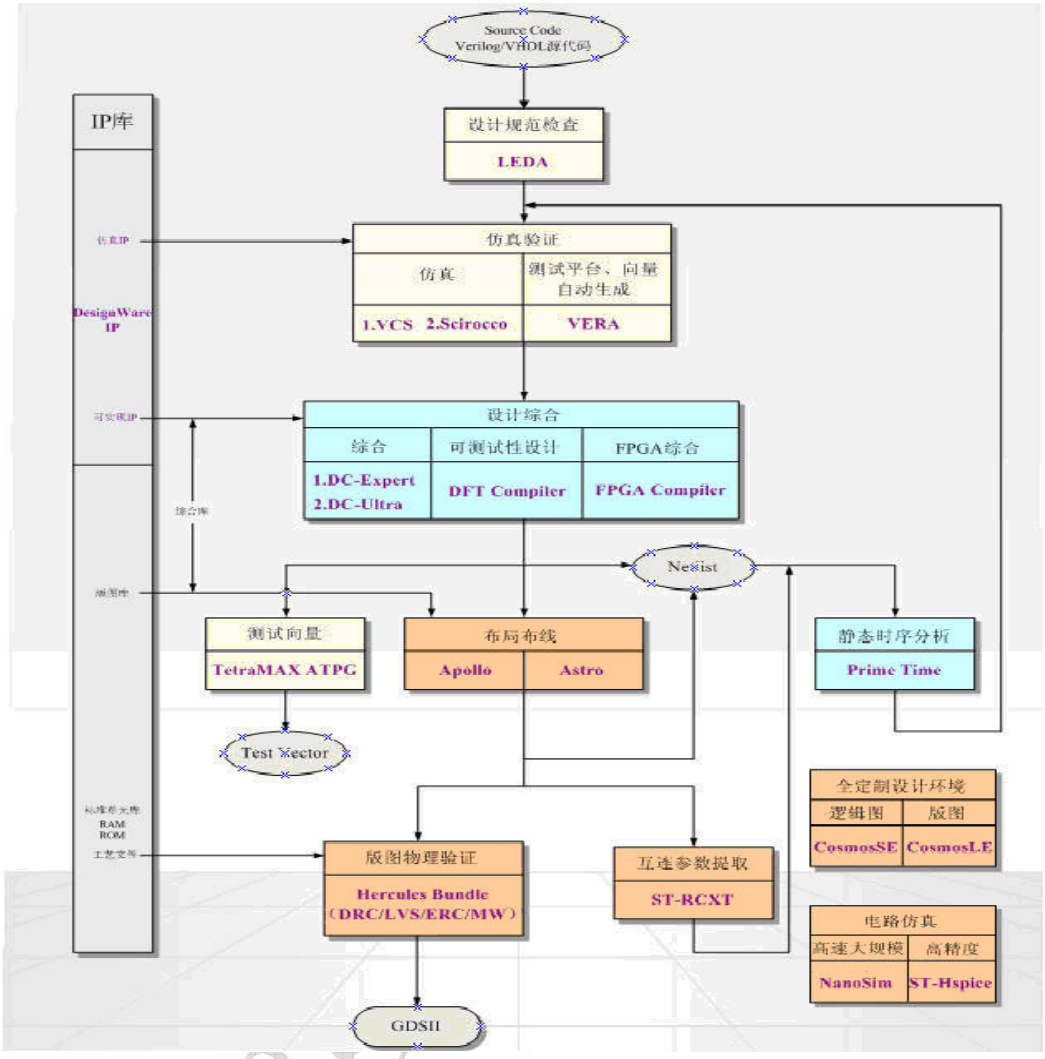
主要用到的工具有：

Virtuoso Schematic Composer， Analog Design Environment(ADE),Spectre/Spectre RF（MMSIM）,Virtuoso XL,Dracula DRC/LVS/RCX.

我们使用集群计算有何好处？

首先，我们做模拟设计中，最耗费计算资源的工作就是 Spice 仿真，这里主要的工具是 MMSIM 软件包中的 Spectre。很多大型的后仿会持续很长时间，有的后仿真甚至超过一个月，这样就给工作带来了很大的问题。我们等待这个月任何一个意外发生，都可能导致工作重来，比如停电，服务器死机，人为错误。使用集群，我们可以加速这个过程，在计算资源允许的条件下，我们可以加速 5 倍以上。显然，这可以大大优化流程。

Synopsys 数字设计流程：



Synopsys 主要用到的工具：

VCS,Leda,Design Compiler(DC),TetraMAX,Astro or IC Compiler,Prime Time,ST-RCXT,Calibre or Hercules.

使用集群的好处：在我们数字设计中，会用到很多验证，而这些验证工作会很多个模块，这就意味着我们会有很多个 Job 同时运行。如果我们有 1000 个 Jobs，只有 100 个 CPU Cores，同时提交上去，会导致系统运行的 job 互相竞争资源，系统很容易死机。而我们使用集群，可以将这些 Job 排队，前面一批 job 运行完毕后，后面一批进入，而这一切完全是自动化的过程，无需人为参与，只要工程师一次性将这些 jobs 提交到队列即可。

事实上，在 synopsys 的设计流程中，VCS 的部分功能，DC，PrimeTime，Calibre 等都可以实现并行计算，都可以加速设计流程。

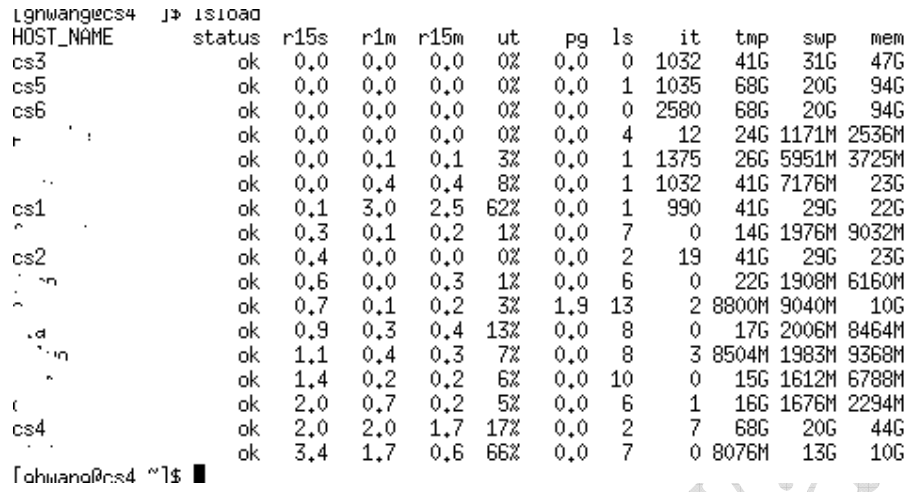
以上只是举例说明两个设计流程中适用集群可以优化的地方，目前最佳的集群环境是 Platform 的 LSF。在 IC 设计 TOP10 的公司中，几乎都采用了其作为主要的集群管理工具。

使用集群，我们可以将服务器资源利用最大化，实现 24 小时不间断运行任务。并且，永远会将最闲的服务器最先利用上。

使用集群，我们可以加速 IC 设计的很多流程，从而加快设计的完成。为我们的芯片尽快完成赢得时间。

全文完

下图为 lsload 命令所显示的结果，大家可以看到各台服务器的负载，CPU 利用率，剩余内存等信息。



LSF 会自动去调度，找出最佳的后台服务器，尽量做到负载均衡。所有的后台服务器，用户都不能直接登录去 run，这是由系统和网络结构限制的。但是，对于用户，要让所有的操作做到最简单，用户不需要去了解复杂的后台设计。

这里介绍一下 LSF 的一些使用

a. 提交 job

$bsub my_job

Job <1234> is submitted to default queue <normal>

上面输出中 1234 是分配给 my_job 的 ID, normal 即系统默认 queue

b. 提交并行 job

$ bsub -n 8 myjob

myjob 以并行 JOB 的方式执行，且需要 8 个 cpu cores。比如在脚本中，hspice 使用了-mt 8 的情况下。用上面的命令会让 lsf 帮你找到空闲的 8 个 cpu core 之后才提交给具体执行的主机。

c. 查看当前自己或者其他人的 job

$bjobs (只查询自己的) $bjobs –u all（所有人的）

然后可以得到 jobID

$ bjobs -u all

JOBID USER STAT QUEUE FROM_HOST EXEC_HOST JOB_NAME SUBMIT_TIME

1004 user1 RUN short hostA hostA job0 Dec 16 09:23

1235 user3 PEND priority hostM job1 Dec 11 13:55

1234 user2 SSUSP normal hostD hostM job3 Dec 11 10:09

1250 user1 PEND short hostA job4 Dec 11 13:59

d. Kill 掉自己的某个 job

$ bkill 1234

Job <1234> is being terminated

e. 挂起和恢复 job

$ bstop 3421

Job <3421> is being stopped

$ bresume 3421

Job <3421> is being resumed

f. 查看 job 的输出

$ bpeek 1234

<< output from stdout >>

g. 查看服务器负载

$lsload

h. 查看服务器状态

$bhosts

i. 查看 job 的详细信息

$bjobs –l 1234

8. 版本控制和 bug 管理(cvs/svn/vss/bugzilla)

版本控制在 IC 设计方面，使用 CVS 的依然比较多。优点是简单，使用方便。但是，缺点也多，比如无法很好的管理二进制文件。所以，目前使用 CVS 的升级 Subversion 的人也是越来越多，我比较推荐大家多考虑 Subversion，特别是在有二进制文档管理的时候。

VSS 主要是一个基于 windows 下的版本管理工具，被大量使用于域环境下的文档和代码管理。不适合在 linux 下使用。

Bugzilla 是一个使用 perl 写的 bug 管理工具，基于 web。可以配置邮件服务器，这样任何 bug 的状态更改，都会发 email 给用户，可以很方便的管理内部的 bug。需要注意的是，bugzilla 的安装，因为需要安装很多 perl modules，所以最好是安装服务器可以临时上网，直接通过网络安装那些 perl 模块。

本来还想详细介绍一下如何配置 cvs，svn 这些工具的，不过我发现 google 一下也可以很容易找到。我就不详细写了。

9. 服务器访问方式(Xmanager/VNC/NX/Citrix)

很多 IC 设计公司非常担心自己的数据被泄密。其中包括被离职的员工带走，同时还希望出差的时候也能工作。或者说有好几个分部，分部之间要能一起工作。

老实说，这部分内容，我在 B 公司的时候没发现他们有多少需求。员工可以 VPN 到内部，然后登录服务器。

对于服务器的访问方式，我想在这里详细介绍一下。因为我看到很多公司依然在采用缺点比较明显的访问方式。

Xmanager 或者 Exceed 都是早期访问 unix 上的 GUI 工具非常方便的访问软件，也在很多 IC 设计公司使用。但是其缺点也很明显，主要有两个缺点：一是占用带宽比较大；二是一旦 PC 端关机或者网络中断，正在做的工作将丢失，无法找回前面的 GUI 界面。特别是第二个

缺点，给我们工作带来了很大的麻烦。

VNC 是一种新的访问方式，所有一切都在服务器上。你只需要在 PC 端下载一个几百 k 的小工具即可。VNC 的优点就是，无论你的网络如何断开，或者 PC 端关机，都不会影响到你正在做的工作。VNC 的缺点是：每个用户都需要一个单独的 vncpasswd，需要用户自己设置。同时，偶尔会出现 VNC 进程挂起，大量占用服务器 CPU 的情况。

另外还可以使用 NX 工具，这类工具有几个优点（这段引用网友的介绍）：

a.通过 ssh 进行 X 界面的传输,可以直接读取 NIS 中的用户密码,方便管理.

b.可以 Keep Session,保证断开连接后会话不中断,大大增加了工作效率.

c.对图像进行压缩后进行传输,这点可能在局域网中体现不了太多的价值,但如果你的公司中有一些特殊人群需要使用 VPN 在外部进行连接,NX 的连接速度比 VNC 快的多.

d.动态显示窗口调整,如果你在公司需要连接显示器进行工作,回到家用直接用笔记本的小屏幕进行工作,NX 就非常适合你,VNC 做不到这点,VNC 需要在启动前定义窗口大小.

NX 目前有几个版本,第一个是 NX Nomachine,这个是商业软件,如果有预算建议买这个版本,NX Nomachine 也有 Free 版本的,但是只支持 2 个用户.第二个是 NeatNX,这是 google 基于 NX 开源 lib 做的，也是开源软件.第三个是 FreeNX,也是开源软件，可以支持不限个数的客户端.

FreeNX 是基于 nomachine 公司对 NX 的开源发布制作的。这个工具类似 VNC，不过不需要 vnc 那样每个用户都设置一个 vncpasswd，它只需要系统密码即可。不过客户端需要配置才能使用。由于我在使用 freenx 的过程中，出现过几次桌面系统无法使用，且无法找出原因及解决办法，所以我很少使用。但是我相信，如果你买了 NX，出现这种问题是可以得到商业支持的。

Citrix 是另外一种访问方式。我在 B 公司的时候才发现，他们居然使用这个工具来作为那么多 IC 设计工程师的访问工具。这个工具非常成熟，但是价格昂贵。只有基于 windows 和 solaris 等 unix 的版本，没有基于 linux 的版本。所以，如果我们采用，在需要访问 linux 资源的时候，需要有一个巧妙的方式将 linux 上的 app 发布到 unix 服务器上。

我们需要满足国内公司数据安全的要求。我采用的方式就是 Citrix+基于 XPe 的瘦客户端。

首先，Citrix 实现数据不能 copy 到本地来，同时禁止剪切板等。瘦客户端让用户的一切工作都是基于 Citrix 服务器的，并且用户还不会感觉到任何不方便。

其他的客户端，包括 Exceed，VNC，FreeNX，我都测试过，都无法做到隔绝数据，用户都能通过剪切等方式 copy 到本地来。

Citrix 通过 disable 剪切板，可以让用户那边访问到服务器的同时，而不能 copy 走任何数据（当然截屏是防止不了的）。这一点，对于给合作伙伴或者远程用户提供服务器访问尤其重要。

10. 虚拟环境（vsphere virtualbox 等）

在我们的设计环境中，可能有一些功能服务器，比如 AD/DNS/Mail，他们平时的负载很轻，但是可能他们的 OS 和计算服务器不兼容，比如 AD 服务器，我们希望使用虚拟机的方式来节约硬件资源。

目前虚拟技术做得最好的依然是 vmware，如果大家不担心授权问题，可以考虑 vsphere 5.0。对于作为虚拟机的服务器，内存最好多配置点儿，而 CPU 多核心即可，频率不需要太高，毕竟我们只是在上面跑轻负载应用。

另外，提醒一点就是存储方面，尽量配置带 Cache 的 raid 卡，可以对 IO 有很大帮助。

虚拟环境，不要拿来作为计算服务器使用，毕竟虚拟过一次之后，整体性能方面比物理服务器上直接跑计算还是有一定的性能损耗。

11. NFS 和 NIS

为什么 IC 设计公司，多采用 NFS 方式实现统一的环境呢，我想这是因为 nfs 作为 unix/linux 下最方便，最简单的一种方式，给大家带来了很大的方便有关。基于 nfs，我们可以将 NAS 上的存储空间 share 给多个 nfs client 使用。目前 nfs 一般都采用 v3 或者 v4。建议，nfs 和 automount 结合起来，那样就不需要在每个 client 上一开始就 mount 很多目录了。

NIS 作为一种很古老的机制，一直使用到今天，依然备受 IC 设计行业青睐，确实不容易。也许未来会被 ldap 取代，但是目前依然是比较合适的一种方式。NIS 就是方便每台加入 nis 域的服务器，都能被用户使用同一个用户名和密码登录。

提到 auto mount，我这里大概讲一下，如何在 NIS 和 NFS 结合下，实现 automount。

a. NIS Server

```
#cd /var/yp/
#vi auto_master
/project        auto_project    -rw,fg,hard,intr,suid,proto=tcp,vers=3
#vi auto_project
projectA    zfs_server_ip:/zdata/project/projectA
projectB    zfs_server_ip:/zdata/project/project
```

b. Zfs server

```
#zfs set sharenfs=rw=@192.168.0.0/24,root=@192.168.0.20/32    zdata/project/projectA
```

以上是允许 projectA 从 nfs server 上让 192.168.0.0/24 这个网段可以 read/write，同时，允许 192.168.0.20 这台管理服务器可以以 root 的名义访问数据。

c. NFS Client

```
#chkconfig –level 345 autofs on
#vi    /etc/auto.master
+auto.master
#ypcat auto.master
auto_project    -rw,fg,hard,intr,suid,proto=tcp,vers=3
```

关于如何设置 nis nfs 请自行从 google 上查询，这里不再列出。

12. 数据备份

每个公司的数据都有制定自己的备份策略，而这些策略一定要基于自己的现实情况制定。

我在大公司工作的时候，他们的备份策略非常完备。这里也简要介绍一下：

首先，他们在本地存储上，有 D2D 的策略，即任何数据，都在本地其它存储上能找到一份完整的 copy，这类备份的目的是防止主存储遭受 raid 损坏等类似硬件故障从而失去数据，或者恢复数据时间比较长而影响工作的一种措施。

然后，他们会通过磁带库，备份数据，每周都做增量备份，每月会将完整备份，复制一份发送到总部。

同时，还会通过 snapshot 保存几份数据，让用户可以随时自行恢复误删除的数据。

在这里，我谈的主要是小公司，那么我们显然不大可能有那么多资源去做以上这些备份策略，但是我们可以用较小的代价，也实现类似的数据安全。

首先，我们可以在本地通过大容量的 SATA 硬盘，通过 raid 卡组成 raid5 或 raid6.然后，通过 rsync，实现每周一次对主存储的 full copy。这样基本上也不会对主存储有太大的负担，同时，也实现了主存储故障之后，至少可以恢复到一周以前。

其次，对于重要数据，如 cvs 库 svn 库等实现每天一次全备份。做到重要数据，在机房硬件主存储和备份存储没有同时故障的情况下，还能恢复到前一天的数据备份。

最后，对于重要数据，我们还需要至少两块加密的移动硬盘。用于每个月交替备份重要数据，然后保存在银行保险柜。这是为了防止机房出现火灾等重大事故。

13. FPGA 下载数据如何做到安全及归档

做 IC 设计，经常可能会用到 FPGA 的 image 需要下载到仿真器上去，那么如何管理这部分数据呢？

我们主要有两种方式，第一个就是，所有数据通过 IT 人员，由用户填表单，主管批准。IT 负责 copy 出来交给需要的人。同时，将 copy 的数据做一份归档，以备追查。

这种方式，对工作的方便性有一定的限制，特别是有时候周末加班加点工作的时候，IT 人员可能并不在现场，需要 copy image 出来就很麻烦了。

另外一种方式，通过 rsync，每 5 分钟同步一次到专门的下载服务器。同时，将删除的文件自动备份起来。然后，在下载服务器上开一个 ftp，通过 ftp 限制下载数据的大小。这种方式的缺点是，数据安全性并不高，你无法控制用户下载的数据到底是 image 还是源代码，只能做到后期的追溯。

技术解决不了的问题，请交给规章去做，不要迷恋技术可以解决所有问题。

14. 服务器硬件的远程管理

我们知道 dell 在很早就有 BMC 实现，对于目前 Dell 29xx 和 Dell Rxx 服务器，基本都支持 iDARC Express 和 iDRAC Enterprise。对于 windows 服务器，推荐 Enterprise 版本，对于 Linux 服务器，如我们 IC 设计环境使用到的，Express 也可以。我们可以实现远程的电源管理。不过如果需要远程的技术支持人员，还是建议采用 enterprise 版本的 iDRAC 卡。

配置方式有两种：

a.系统启动的时候 Ctrl+E，进入配置，可以设置 ip 地址。当然这里的 ip 地址也要唯一，和服务器网卡 ip 地址一个段。还可以配置 default gateway。

b.安装完毕 RHEL 操作系统之后，可以/etc/init.d/ipmi start,然后通过命令行配置 ipmi 的地址。

```
ipmitool -I open lan set 1 ipaddr 192.168.1.140
ipmitool -I open lan set 1 netmask 255.255.255.0
ipmitool -I open lan set 1 access on
ipmitool -I open lan set 1 defgw ipaddr 192.168.1.254
```

以上配置为本地的 ipmi 设置 ip 信息。

我们可以使用 IPMI 管理服务器了。

a.客户端远程检查服务器状态：

ipmitool -I lan -H 192.168.1.140 -U root -a chassis power status

b.远程关机
ipmitool -I lan -U root -H 192.168.110.140 -a chassis power off

c.重启
ipmitool -I lan -U root -H 192.168.110.140 -a chassis power reset

d.远程开机
ipmitool -I lan -U root -H 192.168.110.140 -a chassis power on

e.显示系统日志
ipmitool -I lan -U root -H 192.168.110.140 -a sel list

f.改变系统 boot 方式
ipmitool -I lan –H 192.168.110.140 -U root -P xxxx chassis bootdev pxe
ipmitool -I lan -H 192.168.110.140 -U root -P xxxx chassis bootdev disk
ipmitool -I lan -H 192.168.110.140 -U root -P xxxx chassis bootdev cdrom

Dell 的默认密码是 calvin，你可以自己修改。