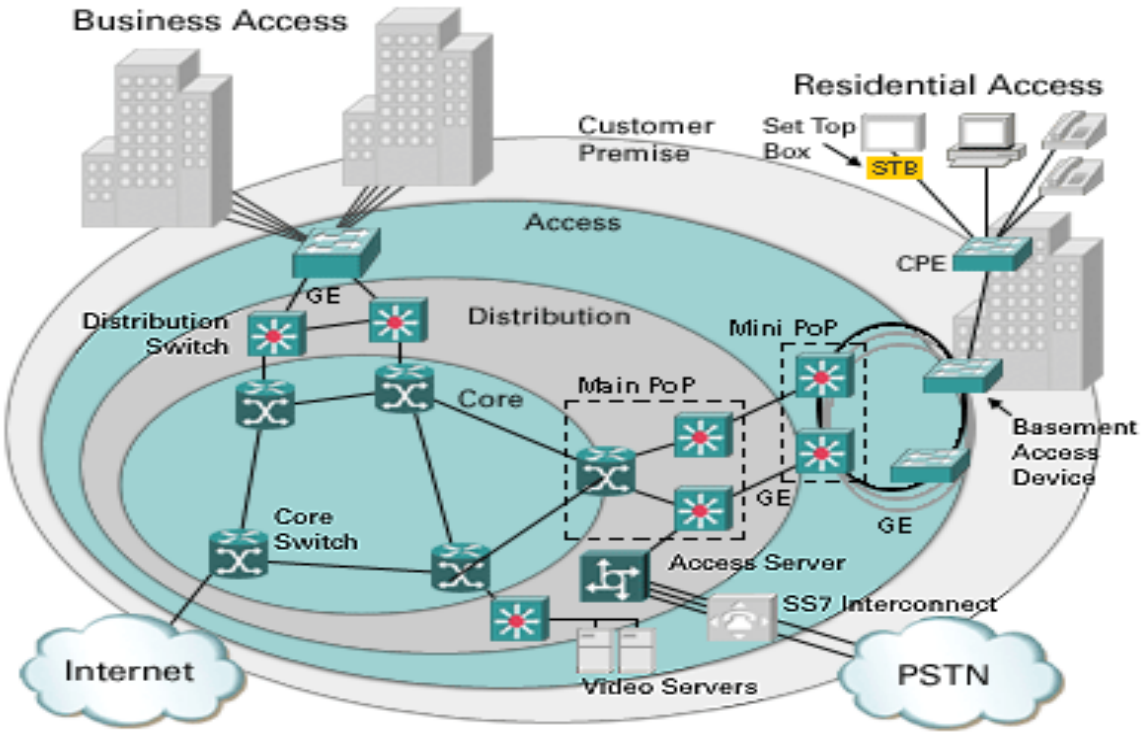
彎曲評論
科技 · 人物 · 潮流

关于城域网的思考

(中)

作者：理客
totobeing@hotmail.com
编辑：陈怀临
huailin@tektalk.org



ALU新ME的前世今生

这个系列的开头说了是为一个讨论why L3的同事的邮件而起，后来总是太ALU的新ME，似乎新ME成了主角，也只好强赶鸭子上架，所以这只烤鸭很可能是只肯德鸡，而不是前门的全聚德

VLAN/QINQ虽然很强，但还是不能是以太网拍拖传统ME方案，比如scalability问题，MAC容量就很麻烦，以致大家在实践中要求在一些场景下，比如傻瓜型管道业务，不要学习MAC，直接根据VLAN转发，以彻底根除MAC学习和容量带来的麻烦。其他的安全/QOS/OAM/NSM等大规模MAN网络需要的技术，传统ME还是没有很好的解决方案，时代呼唤英雄，思科在城域网络上的76/65的传统L3/L2方案固步自封，华为跟班思科还难以成器，AL在欧洲挺身而出，揭竿而起，开创了ME的新时代

AL能成功的推出以7750为中心的新ME也是一蹴而就的，早在2000年，AL就成和华为一起使用IBM的当时业界最好的NP芯片ranier做IP产品，据说当时华为迅速推出NE40/80系列，成功的和思科拉到最近距离，而AL似乎不太成功。但是后面的故事，确是又反了过来，AL的IP部门在成功收购了谷里的NP公司timtra后，即薄厚发，一举推出系列产品配合新ME方案，从欧洲市场开始，在ME领域开始势如破竹，而华为，在和港湾的PK惨胜中，把IP产品的错误持续到10G，本来就是follow策略，在全力纠正10G的错误的时候，即使没有这么多据都错误都未必创造新ME方案，何况当时，还哪有心情去考虑新ME，当AL的新ME 7750出来的时候，有些人立刻又蒙了。

在细说AL新ME的特点前，不得不说说思科老ME产品和方案的特点：

在没有AL7750新ME前，在传统IP城域网方案里，C76/65真的是最完美的孪生兄弟，同时在IDC和企业网市场也是霸主，太牛了，整个产品既有集中式转发，也有分布式转发，还交融许多75系列的接口卡，L2性能很好，L3/MPLS也支持，各种核心引擎可供选择和升级，真个一个万花筒，那是革命一块砖，哪里需要哪里搬，美C76/65的最美好的青春持续超过了5年，知道AL7750青年才俊重磅出山，迅速在

ME市场扩容，然后华为NE40E系列的不断发力，C76/65终于被年老珠黄，满身雀斑，暴露无遗

1、 Packet based的交换网，还容量小

2、 集中式的L3/MPLS转发引擎，性能和功能都老落伍了

3、 升级新的主流特性就得换板，比如L2到L3，不能软件搞定

4、 组播能力也差，不能做大量的IPTV

5、 C65/76同门不同价

6、 TDM/ADM等FMC特性支持差

7、 路由表容量/MAC表容量/TE容量/LSP容量/PW容量/VPLS容量等都和对手不再一个档次

….

这样一个烂柯，居然成为选美冠军近10年，是思科垄断的好，还是看客眼光拙，还是其他美女太不争气。不是金子总要收光的，美女也要夕阳红，除非天山童姥

终于爬到正题，因为这个中很长，所以至少要分3章，为了保持章节内部的连续性，这章的切割就短了些

AL7750新ME是一个从marketing到R&D都非常成功的方案和产品，在2003年左右，使用以太建设城域网已经基本得到确认，那么市场的实际需求和相应的思科基于76的产品和方案有什么问题呢？

1、 VOIP/IPTV需要HA，而C76老ME难以支持，L2不成，L3当时也没有很好的快速收敛技术

2、 商业用户企业网：在L3VPN为主的IP企业专线

3、 HSI其实一直都不是大问题，因为运营商从来都不为internet提供QOS和HA，如果你仔细观察ADSL的合同，运营商提供PIR峰值速率，但没有承诺任何CIR承诺速率，更不用说SLA，虽然大多数情况下我们上网至少都能达到几百K，但是运营商从来都不会做这个承诺，因为怕万一大不大，索赔是个麻烦事。不断对终端客户，就是运营山之间的IP连接，互相也是不做SLA承诺的

可见在城域网逐渐成为运营商各种网络的核心后，电信业务和商业业务在向这个网络迅速迁移，这二者是互相利用和促进，那么老ME具体有什么问题，再总结一下

1、 HA不行

2、 QOS不行

3、 OAM不行

4、 NSM不行

5、 Scalability不行

6、 L2VPN能力有限

这主要是针对L2的城域网的，那么上L3是不是可以解决上面的问题呢？可以部分的解决，但有下面的问题：

1、 HA有限：因为传统的IP/MPLS收敛至少是秒级

2、 QOS可以算解决

3、 OAM：MPLS L3VPN的IGP/MPLS/BGP带来的运维问题，是客户认为都是L3造成的恶果

4、 NSM：思科没有运营商的网管，是企业家的，C一直不愿意在网管上加大投入，导致基于SNMP的IP网管和IP协议的蓬勃发展太不和谐，和其他SDH等网管相差甚远，听这个名字simple network management protocol，IP这么复杂的网络，居然用一个简单网络管理协议，能和谐吗？当然IP过于复杂和变化是一个原因，但思科基于政治考虑不愿意投入也很可能是一个原因，因为思科有很好的sustainable的培训体系，不用自己花钱，并且还赚钱，为什么不用CLI，投入一个GUI的很好的NSM对思科有什么好处呢？但是这对其他供应商不一样，因为他们没有思科培训这个超强的武器，在思科的垄断下，客户有苦也没有办法，在IP的SNMP严重滞后的情况下，做业务级的网管必须要和主流厂商合作，这里最主流的是思科，思科不想玩，别人就都是太监。这里对OAM/NSM说很多，因为在发达国家，这是个大投入

5、 Scalability：L3VPN是基于BGP的，扩展性很麻烦，还要RR，用L3VPN做企业网其实是不合适，企业的路由应该企业自己感知为主，运营商需要掺和的理由有限，所以L2VPN(VLL/VPLS)应该是企业VPN的主流，要替代的L3VPN的大部分市场，而此时C76老ME在这方面的能力很有限

从以上分析可见时代需要新ME，那么在看新ME的技术基础是否已经具备：

1、 MPLS TE开始商用，通过TE FRR/HOT STANDBY，可以提供类电信级的HA

2、 L2VPN(VLL/VPLS)开始商用，可以提供更好的企业网专线

3、 POP点的BRAS还是主流，基于PPPOE为主的链接，使用L2VPN很自然

4、 10G的ASIC和NP开始商用，解决MPLS TE/L2VPN下的基本功能和性能

以上是可以利用的条件，那么没有条件的就要创造条件，比如NSM/ISSU/NSR

我们来看ALU具体怎么做的：

1、 硬件核心chipset和收购

需要一套performance、flexibility、scalability等都不错的forwarding engine chip set， 只有美国有这样的公司，ALU成功的收购了timetra，基于其FP1芯片快速推出了7750产品，在这个时候，业界的商用NP都没有这个能力，IBM的Ranier是一款非常好的2.5G能力的NP，但

是因为主流的IP核心产品供应商，除了CISCO，就是老二juniper，大家都不主力使用别人的NP，导致这部分的ROI很差，作为IBM芯片中的边缘产品，尽管技术非常成功，成功到华为使用rainier开发的NE40/80系列产品几乎接近了CISCO当时的旗舰12K，而12K用的是ASIC，所以华为市场成功的包装出第五代路由器，用NP淘汰ASIC。但是路由器不是华为的主力产品，华为也无法快速做到全球30%的市场份额，这不是技术问题，技术再先进，销售平台也要一天一天建，但是IBM不可能把NP的命运压在华为身上，去开发10G的NP，所以只有夭折掉rainier，卖给了一家叫hifen的公司，是否可以卖给华为？即使美国政治允许，以当时华为自己的芯片能力，能否保留住核心团队，迅速消化后开发出有竞争力的10G NP，很难说，从理论和技术上说可以，但是任老板是否愿意做这个投入，虽然不一定是10亿美金级别，那也是一亿美金的级别，当然，任老板和柳传志不同，任从来就不是一个不敢下注的人，所以我猜，而老板当时把宝压在了无线和3G，所以在没有公司系统完整的一套策略支撑下，是不能单独只靠收购一家芯片商就能获得路由器的大幅成功的，打仗不是这样打的，奇兵可以，但是没有整体配合的奇兵，不能取得最后的胜利，看韩国棋手李昌镐的棋，奇兵不是很多，偶然有，但李昌镐的绝大多数胜利，是建立在整体的稳定表现上，尤其是近乎完美的收官，李昌镐鼎盛时期的失败，有对手发挥太好，一直保持优势的，也有偶然自己收官不美而失利的。所以华为路由器在NE40/80的昙花一现，是有其历史原因的，如果当时老板把宝都压在路由器上，那么也可能成功，但是如何衡量是压在无线3G成功的概率更大还是压在DATACOM上成功更多，事后诸葛的看，还是无线3G更高，一个是这是华为在运营商的传统势力的优势空间，另外就是，在路由器思科垄断的市场，似乎比无线3G更容易成功，但感性的看，这是真的吗？华为真的可以干过思科吗？如果没有2002年那场不以人们意志为转移的官司，华为路由器真的在美国开拓了哪怕不大市场，从而带来全球市场的迅速增长，也许老板真的会从策略上把路由器作为和无线3G更接近的投入。但历史不是这样演的，2002年冬季的那场雪，被任老板率精英精心策划，并痛割一半数通给3COM外加市场禁令的沉重代价而化解，貌似胜利，其实是思科的成功，以致后来再想回收3COM的时候，可爱的山姆大叔死活不批，理由虽然是tipping point的问题，其实是欲加之罪，何患无辞，美国人同样不乏中国人创造困难的智慧，没有理由，创造理由也要禁止。但塞翁失马，很难说对华为的影响是好是坏，当然对dadacom肯定是坏是毋庸置疑的，看到今年美国把丰田像当年整鬼子汇率问题一样的往死里整，不难想象，如果没有当初的官司，而事情爆发到今年，美国人会高看中国人一样，放华为一马吗？我不相信，我个人的偏见，对美国人，远比对欧洲人没有好感，因为就我承认的人种论因素，美国人是欧洲的流氓传下来的，根就不好，包括宠物在内贵族似乎需要纯种，所以欧洲好不容易积累下来的贵族气质，在美国几乎荡然无存，但杂种有竞争的优势，有冒险和创新的贡献，也有劣根性的难变，很抱歉扯到了政治，可见我对美国人的偏见之深，世界最闪亮的明星大哥，需要以最苛刻的条件去要求和衡量，道德上，也许并不是那么过分。中国做老大的时候，虽然也打，但对归顺朝贡的小兄弟们，送出的回报远大于那些贡品，而小兄弟们只要认下大哥的称谓就可以了，没有任何物质文化损失，以致这种传统到中国贫困交加的时候仍然坚持，可见中华民族是一个多么勇敢善良文明的民族；现在的美国当大哥，从物质到文化都要搜刮跟班的小兄弟，不管你是有

钱银还是没钱银，刮你没商量，你武装到牙齿，军力全球过半，首富一大堆，财富流油，给受苦受难的小弟们分点汤喝咋就那么难呢？吃一点亏都要数十倍的报复回来，小米粒大的心胸老葛朗台等四大吝啬鬼也会觉得自己冤枉，做大哥的道理，我看美国人还是应该好好和中国人学学，教训中国人，美国人还没有道德上的资格，小道消息：华为也曾试图收购timetra，但是美国大哥没批准，直到现在，华为公司还动用各界游说关系，在向世界上最民主自由的美丽的国家证明老板是退役军人，但是公司其他员工是普通员工，不是妄想要解放全世界全人类包括美国鬼子的解放军。10G的NP，当时可能最好的INTELDE IXP2800，不好意思，终于把话头调回来了，不容易。这个东西性能差，配套芯片贵，也不够稳定，但其实都不是最大的问题，最大的问题是，INTEL抛弃了IXP2800，不再继续升级投入了，对于一个要爬起来的人，如果要继续使用IXP2800做产品演进，如果没有准备好其替换策略，这一棒是很恐怖的。中国人要把小命握在自己安全的手里，太难了。当一个人拼命伸出双手，开始有希望爬上有幸福希望的和平岸边的时候，突然岸边站起一个恶棍的大棒毫不留情的砸开那双可怜的手，这个世界上，是有人干这种十恶不赦的罪恶勾当的，但上帝是西方人的，这些罪恶如果是西方人干的，都是可以宽恕，并且上天堂的，所以我看过罗马的斗兽场后，更不信任基督教的哲学合理性了。这里罗嗦这么多，说明本身无关风雨的技术，在影响到奶酪问题的时候，也难以摆脱政治的纠缠，不管你叫人家大叔大哥还是爷爷，美国佬就是美国佬，叫啥都没用，这里只是中国人在争取核心技术竞争力的时候被歧视和痛贬的一个小小缩影，警醒每一代的少年在反对过分的民族主义的同时，要知耻自强，命运自握，激烈反对造假

2、 系统架构和NSR/ISSU（HA）

7750的硬件结构，从单板到chassis，虽不能说是非常优秀，但无疑是比较成功的，从紧凑的三维到很高端口密度以及散热电源，比如其最早提供的40*1GE的单板，就用了特殊的PHY使面板可以布下如此高密度的以太口，7750的设计，出奇之处本身都是为了方案，并非为了吉尼斯，因为7750无疑是一款比较昂贵的产品，因为其转发芯片，大容量查表器和TCAM，支持H-QOS的TM等在当时都属于贵族用品，所以成本要高于C76，欧洲人和美国人的差异就在这里J，但是市场没人理你是否内部用了金子还是铜子，客户在冷漠的时候只关心你要我掏多少银子，你是有足够理性的独立行为能力人，你跳不跳楼不应该由我来负责，那用什么来分担这么好的产品的成本，在当时，10G端口大家都贵，并且主要作为收敛后的上行，从方案上也不合适做高密度收敛板，所以高密的GE口收敛板是非常好的idea，AL不是疯子。

7750的软件架构从外部看也是很成功的，主要有以下几点

（1） 其扩展性好，可以很容易的做到一年一个大R版本，年内可能还有几个个RX.x版本，这在以多业务下的IP/MPLS TE为核心的路由器产品，不是那么容易做到的，当然AL没有思科那么大的历史包袱和产品系列，这个可能会更容易，但那么多新模块代码，项目开发本身并不是很难，但是如何快速的和软件平台做好集成，推出商用版本，这就需要系统架构设计要好，当然没有那么完美的东西，毕竟是新产品平台，售后出现问题多一些也是难免的，大家都如此

（2） 可靠性：AL可能是第一个做了NSR和ISSU，JUNIPER可能也很快做到了。

NSR用于主MCU故障时，系统业务不会有任何中断，基本原理就是把所有控制层的session都热备了，使邻居感受不到这个故障，这个和NSF不同，NSF也是要达到这个目的，但是邻居是能感受这个故障的，所以要提前通知邻居，我故障了，别挂电话，我方便一下很快，回来继续聊。另外一个是系统升级的时候，也是业务不中断，具体原理有点复杂，不多说了，因为我不大了解具体实现。但是需要指出的是，AL作为商人，买东西要吹个150%是正常现象，比如NSR，不是所有情况都能NSR，是有条件的，ISSU更是有很多限制，最致命的是，好像升级后24小时还是2小时内记不得了，要重启一下，知道这个隐埋的bomb后，好想笑，不是嘲笑，就是好笑。笑归笑，AL市场包装后的忽悠效果还是很可观的

3、 NSM

NSM在许多国内产商一向不甚重视，而这里放在系统架构后面的第一副领导的位置，可见发达国家市场和发展中国家的不同，也可见AL新ME方案的精心设计，这里有两个原因

（1） 发达国家人力成本昂贵，尤其是IP工程师，所以好的E2E业务的网管对TCO saving非常重要

（2） 多业务的IP/MPLS TE路由器，AL叫SR，系统配置复杂，如果没有好的NSM，这个运维的缺点就会难以掩盖

（3） 思科没有运营山级的ME方案的E2E网管，运营商本身就觉得路由器复杂，没有好的网管就更复杂了，当然思科不是没有能力做这个事，许多网络全网都是思科设备，思科做好网管不是更容易吗？可问题也就出在这里，既然我近乎垄断了，我还费劲巴力的做好网管，给谁省钱？你们和我思科一起玩培训不是让我既能赚钱，还能培养客户的loyalty吗？有何必自断财路呢。这种策略下，以致后来个别熟悉思科产品CLI的用户，不需要GUI的NSM网管，就喜欢CLI

AL是充分分析了市场形势，花了大力气完成SAM6520网管产品（也许名字我记错了），和SR一样，叫SM，业务管理，这些可是AL新ME方案innovation的主要精华之一。并且AL网管的报价方式也不错，把网管价格直接报在端口中，而不是按照网管能管理的节点数来报价，可见AL的报价方式更精细

AL的ME新网管，给似乎已经沉寂了几年的IP网管产品注入了一股新风和活力，重新激活了这部分网管市场，带动了运营商IP网络网管市场的许多供应商的新产品开发动力，AL的新网管对这里的贡献功不可没

4、 OAM

Native Ethernet本身实在是简单，也许因此很久都没人去关注OAM，在LAN的时候，尤其是在企业网，维护很简单，似乎也没什么大问题。而在电信网，OAM在传统网络中从协议到产品实现到网络设计，都占有自己的一席之地，即使到了AL新ME的时候，并没有立刻产生专门的的ETH OAM协议，但是因为新ME的核心是EoMPLS(Ethernet over MPLS/TE)，核心是让L2的的广播域通过VPLS站在在MPLS/TE的肩膀上（VLL可以看作VPLS的一种特例），从而获得MPLS VPN的隔离安全性和TE的可靠性，许多事情都不得不具有相反的两面，你获得的这些好处很难不付出代价，所以EoMPLS同时也把MPLS/TE复杂

性带了进来，所以在没有ETH OAM标准前，AL就做了私有的MAC ping/tracert，并通过网管和VCCV ping（VPLS），MPLS ping/tracert等作了还不错的关联，可以说给基于EoMPLS的新ME一个初步的OAM解决方案，虽然不甚完美，但在AL的OAM特性及配套网管的包装下，能做的这个程度，虽然不能和ATM/SDH的OAM相媲美，但也算过得去，不致让这个短木板太短

5、 H-QOS

这里不提新ME方案，是因为在AL7750前，没有产品做HQOS，QOS要做到什么程度，这个争议一直就没有停止过，而7750包装了5级H-QOS，my god, what are they doing?!直到现在，HQOS的商用并不广泛。IP QOS的研究应该早于2000年，但真正开始实现大概从2000年开始，QOS一直被认为是是否重要的事，这在非常重视质量的西方是没有疑问的，疑问在怎么做？谁来做？既然说到QOS，所以插入一些背景信息，以便理解。QOS的分类方法很多，按照时间或者怎么说，按照事先避孕或事后丸补救，划分两大类

（1） 事先避孕：术语叫CAC（Connection Admission Control），这从传统的电路交换通信时代就开始有了，这个时候QOS其实不是可选，而是MUST，因为是电路硬链接，在真正建立呼叫前，你必须通过协议把各段电路分配好了，然后才能告诉两端，OK，连好了，你们做吧，如果电路资源不足以建立这个链接，那么只好让两边等，当然，你们有YY的权利，这不可耻。所以不存在用户多了影响通话质量的问题，只是影响接通率，传统语音发达了几十年，有一整套监控质量的数据体系，什么呼损率了等等，我不是很熟，但一般来讲，只要接通了，就可以安全的做了，不比担心被抓，拥塞产生的语音通话质量问题，更多的是在VOIP刚开始流行的时候，这里面，万一资源用光了，有重要电话不能做，影响重大如何处理，不用担心，再挤的火车，也得留出一些首长专座以备不时之需，不是腐败，比如119，110等，你得给人家一直预留好资源，保持线路通畅吧。首长也一样重要，首长不舒畅，如何为人民服务？首长因得不到好的服务而生气冲动，作出错误决策，那要害多少人。在IP电信化的考虑中，这个技术是首先要被考虑的，在还没有MPLS TE的时候，就考虑利用IP HEADER中的一些保留bits来做拥塞信息的标记，通过负反馈机制告警汇聚接入点的设备，客满了，普通客人，就甭接了，但对VIP客户，当然要继续接，思科路由器还实现了一些RFC，后来这些研究也在MPLS/TE上做了一些，因为目前看不到什么实际意义，所以也忘了这些RFC的名字，技术象妓女，一旦过气，就人前冷落鞍马稀，懒得有人光顾了，人类有时候真是太不是东西了。IP的CAC，最主要的还是TE技术，现在还有一些类似的方案，比如从终端开始发起TE tunnel/LSP，从理论上当然是很好的，但是商用上有很多麻烦，TE的麻烦事很多，比如TE的带宽分配模型就很难记住，只记住一个名字最好玩的，叫俄罗斯木偶，但这个木偶是怎么工作的的，次次记，次次忘。因为TE需要路由协议扩展配合，还有各种麻烦，所以对大容量的TE tunnel/LSP的支持，也会很麻烦，也就意味着增加成本。如果传统的基于硬件电路独享CAC可以叫硬CAC，那么基于TE的CAC也可以叫做软CAC，TE的主要目的是把MPLS的面向连接的LSP提供预先资源保证（QOS保证），所以从IP(无连接)-MPLS(有链接)-TE(有保证的连接)，是IP技术发展的三大步走，但是因为IP电信化的FMC进程并没有那么快，而IP网络主要的traffic generator还是电信不会提供QOS保

证的internet，所以TE的QOS技术基本上没有什么大的用武之地，商用部署有限，具体到部署，在一个线路上，一部分要做RSVP，一部分不做，是不好处理的，具体就不说了，包括DS-TE这些麻烦的东西，也就省了说了

（2） IP QOS

前面的QOS可以算情色QOS，有人说情色不是色情，不过后面的章节熟套或流水帐可能多一些，不一定有意思。大片重映观众的高潮更多在中段，没有悬念的尾声还不如夕阳红，第二次高潮，在大片后新气象里可能会有一些。但是GMPLS/TMPLE/FOCE及之后更新的研究，因为和个人以商用为主的工作距离大一些，一时可能难以有时间去仔细一点的关注了，所以相关章节会拖的较久

6、 L2VPN

这本是AL的EoMPLS新ME的核心转发和业务支撑技术，可分为点到点的VLL和多点到多点的VPLS，MEF的称谓有所不同，并略有细化，分为E-LINE/E-LAN和E-TREE，E-TREE其实是VLL和VPLS结合的产物，也有叫SPOKE PW的。E-LINE/E-LAN的名字写起来很清楚，但读起来很麻烦，类似在做presentation中常用的cost和QOS，读起来较易混淆。

VLL是很简单实用的管道，对于点到点管道业务，很好，主要问题是

（1）CE/PE间的HA不好处理，当然可以租两条VLL，但是就怕提钱

（2）对于中大型企业网的多点需求，有点麻烦，一个是钱的问题，还有多条VLL带来的VLL资源浪费

（3）广播/组播业务难以支持，每个PW都复制的方式显然不爽

VPLS很灵活，理论上可以做任何业务，对多点业务，节省PW资源，CE双归的HA容易，支持组播，但问题如下：

（1）广播/组播/未知单播抑制：要基于接口/VSI/chassis，要基于packet和带宽，要绝对数值和percentage，都是trouble

（2）环路：VPLS的CE接入的HA方案，要解决L2环路问题，这里花样繁多，有AL的MAC flapping局部方案，有STP in VPLS的方案，还有其他厂商的私有方案，也是trouble

（3）MAC容量限制：因为VPLS要学习MAC，自然涉及到学习的性能，还有容量问题，对tier1运营商，在汇聚的中心节点，甚至有million机的MAC地址需求，这么多MAC，管理等都是trouble

所以VPLS带来的好处相比于同时带来的这么多麻烦，个人观点是尽量不用，谁用谁知道麻烦在哪？甚至还有客户为解决一些VPLS的问题引入PBB over VPLS，faint。当然，在组播场景下，如果用L2组播，还是不得不用VPLS，所以是用L3组播还是L2组播，个人倾向于前者，只是设备商也很苦，不能总是便宜运营商，所以对于便宜的ME接入和汇聚节点，L3经常要单独出来收钱

E-TREE：是很实用的模型，现实的商用网络，full meshed/half full mesh的多在核心，而在接入汇聚层面，更多的是TREE形，比如DSLAM和BRAD的关系模型；另外就是dual-homing的Y形，中文翻译成丫形，这是E-TREE的最简单的形式，可以用VLL redundancy，也可用这种spoke VPLS，简单的实现双归HA，但是protection switching的

性能有限，要提高，还需要BFD/TE FRR和MAC地址快速withdrawal配合，HA的解决，好像就很少简单过
L2VPN在一些时候需要透传所有报文，包括L2协议报文，尤其是VLL的时候，可以叫E-PIPE。L2VPN的透传需求带来一个QOS问题，就是拥塞的时候如何保证L2/L3协议报文的优先级，按说应该保证才好，但实际上好像没有这样配置，其实如果扩展起来，是不是L4以上的协议报文也要保证？那可能要DPI了。但是对于路由器本身主机的协议报文，一般是要配置一个默认的高优先级队列，否则拥塞的时候就会导致协议中断。

7、 VOIP/IPTV

补充一点H-VPLS，和H-VPN(L3VPN)有点类似但不同，H-VPN可以减少UPE上的VPN路由学习数量，H-VPLS和H-VPN拓扑有点类似，但并不能减少MAC学习的数量，我理解就是打开了普通VPLS的水平分割。
QOS的三个关键是：classification时识别深度和能力；队列数量级数和灵活性；精度
TV是推动人类文明进步的一个伟大发明，语音把文明从无声世界的传播带到了有声世界的传播，而TV则进一步把世界带到了视觉世界，黑白和彩色的区别虽然也很大，但相比于从黑暗到光明，就不大了。从此语音和视频成了人类生活的基本需求，并且及其简单，对于用户，你只需拨一个号码，或者按一下频道，就可以获得需要的声音或视频，及其方便。在视频和语音的承载网络上，传统网络比较简单，尤其是视频，像自来水一样，建立一套管道，内容在自来水厂，用户只要打开自家的水龙头就可以享受到丰富多彩的内容，有传说青岛的居民还可以在自己的水龙头里接出啤酒来。青岛的湾友可以澄清一下。
语音/视频和INTERNET本来大家井水不犯河水，相安无事，各走各的独木桥好好的，但首先是internet多媒体化，PC越来越多，那么之间打个电话是很容易的事，VOIP从此而起，有好事者把internet VOIP搞到了运营商，在早期，以电路语音为主的传统电话容量非常大，并不care VOIP有什么降低成本的好处，尤其是中国，早期的VOIP电话卡不少是假的，其实就是普通的电话包装个名字，所以你会觉得VOIP真牛，通话质量和原来一样，殊不知，本来就是一样的，如果有一天不一样，也是运营商自己加点扰，就像有idea未来防止P2P过度，故意让P2P业务在传输时质量差点一样。但是随着IP网络因为internet的洪水泛滥，导致IP网络的建设成本今天已经成为运营商的大头，所以再维护一套传统语音的成本就逐渐越来越可观了，所以几乎没有几年VOIP就替代了传统语音。世界有时候很戏剧，固定电话IP化不久，因为固定语音被移动语音的侵蚀及其迅速，导致vendor已经不愿意在去竞争固网NGN的项目，甚至到了互相送给对方的地步。
说到语音技术，在传统语音时代，大容量语音交换机可是西方封锁我国的一个核心技术，巨大中华的发展起源于解放军通讯工程学院的邬江兴教授在万门机的breakthrough，现在邬早已经是将军了，这个技术在中国的扩散虽然有不少或公或私的官司，但从结果看，是造就了中国通讯产业的大爆发，从这个角度看，邬教授功德无量，当然这里，还有89后西方对中国封锁万门机技术和产品的关键因素，可见封锁很多时候是中国发展的逼迫剂，而开放却可能扼杀中国的技术，这样看来似乎西方人傻了，既然想扼杀中国的核心技术，那为什么不开放呢？其实不傻，如果真的开放了，中国人也不傻，学得更快，更疯狂。从技

术产业角度看，知识产权的保护和垄断一定要在一个适度的范围内才能促进产业的繁荣，过度保护无论对技术产品和是老百姓生活都是利大于弊的。西方许多产业的繁荣，也是依赖于核心技术的快速传播，包括跳槽创业就是很好的方式，而竞业禁止如果真的被严格执行，那就是以在杀人。

具体到VOIP的技术，其实比传统技术复杂，有两套体系：H323和SIP，前者技术简单一些，后者复杂一些，所以传统运营商选择更多的是后者，而新VOIP运营商可能用前者多一些，我不是很熟，所以就不写了。这里更大的创新不得不提到skype，有核心的技术可以通过普通廉价的IP线路保证语音质量，曾经非常火，国内也有following的。但是语音这块蛋糕，无论如何做，在五彩斑斓的IP业务中，都没有太好的利润，所以skype被一个巨头收购，忘了名字，后来好像又分离了，看来是失败的收购，skype还是可以活的尚可，但风云靓丽了几年后，应该很难再有show time了。语音的复杂其实IP承载并不占很多，更多的体现在各种制式的互通上：传统语音和VOIP，移动之间（GSM-CDMA-WCDMA-TDD-LTE），移动和固网，H323和SIP，导致需要很多互通网关，而语音是通讯的最基本业务，保证其质量在这种复杂的互通情况下，并不是那么容易，比如接通时间和时延问题就不是那么好处理，我对这方面的不熟，只是点到而已。

说到VOIP对网络质量的要求，在刚刚开始的时候是绝对怀疑IPQOS的，所以早期一些运营商比如中移动甚至建立独立的IP网络来承载语音，甚至有国家要求必须保证传统语音要做为VOIP的应急备份以保证有灾难发生时候的通讯。随着IP带宽的迅速扩大，语音那一点点流量逐渐成为带宽的一个零头，给你一个EF队列，在高速的IP网络里，时延基本不再是一个问题，但是在需要忽悠客户的时候，对这些从语音时代走出来的电信运营商，语音对网络质量的高要求，现在仍然是一个还没过是的主题

视频是带宽的主要占用者，所以说到IPTV，就要说一下带宽的问题，在有独立带宽保证的TV传输网络，网络本身基本上不需要提供什么QOS保证，这也符合internet的IP承载网带宽充足，就不需要QOS的观点。因为2001年的IT泡沫，带来了大量的骨干网光纤资源，而WDM技术的发展，目前已经超前IP流量，所以运营商的骨干网带宽到现在都不是问题，并且自有率很高，租用也比较容易，这都是托2001年IT泡沫的造福。但是在接入网方面，情况就完全不乐观了，而TV对接入网带宽要求较高，所以接入网带宽是IPTV的瓶颈。

标清视频需要4M的带宽，考虑到同时还有HSI是频道切换时的单播加速技术，那么最小带宽需要是6M，对于最大量的铜线接入技术，需要ADSL2+来承载，同时还受到距离的限制，不是所有的的用户的ADSL都可以达到的。所以基于ADSL的IPTV，质量保证技术要求就比较苛刻

接入网带宽提升方法很多，从早期的Ethernet到户，到已经喊了很久并开始逐渐部署的FTTH技术，但是这里主要的问题是成本谁来承担的问题，因为只是增加一个IPTV，每月每个用户增加的收入也就10-20欧，那么改造这样一个用户需要多少钱呢？没具体值，但看发达国家高昂的人力成本和较低的施工效率，一定是很高的，那么结果就是ROI非常低，所以没有人愿意承建。但高速信息公路是很好的东西，怎么办呢？两个办法：

1、 国家投入模式：如日本和新加坡，FTTX很好主要是以国家投入为主，这和我党改革开

放要想富先修路的思想有点像
2、 运营商投入但要求专营权的形式：因为欧洲电信法早就通过了，所以所有驻地资源都必须开发，并且价格不能完全由卖方说了算，导致incumbent的运营商因为长期官僚和低效率的问题暴露出来，而新运营商得到不小的发展机会。但是对于FTTX，在国家模式不愿意的情况下，只有incumbent的carrier有能力建设，但他们提出，要我建可以，但要求一定年头的专营权，否则这种自己种树给大家乘凉的傻事，打死都不干
3、 像中国这样有钱有人成本低的国家，倒是可以大规模兴建FTTX，尤其是没有固网而最有钱的中移动，最应该用大量的利润快速覆盖FTTX，既能占领未来的接入网市场，又可以不用分很多利润给老外持股者，利民爱国的好事，要赶紧做呀
IPTV虽然不那么好，但却是triple play的核心，所以也是AL新ME核心，主要包括如下内容：
1、 基于VPLS/TE的IPTV承载方案：
（1）核心是环网方案，AL称为daisy chain，就是环形H-VPLS，如果环上的节点或者链路故障了，则通过TE FRR做环回，这种方式的问题是有点复杂不说，也浪费带宽，H通过PIM redundancy优化了这个方案，后来AL也follow了。
（2）树形H-VPLS方案，通过VRRP in VPLS来解决链路备份问题，同时也解决了这种拓扑下组播流量的多发问题，好像整体上还要配合一些MAC withdrawal技术，记不得了，反正很烦
2、 频道快速切换和丢帧重传
这是基本的QOE保证，始作俑者是思科和微软，一些简单分析如下：
（1）频道切换问题：这个问题的本质是MPEG等视频压缩标准+组播定速报文发送导致的，因为其基本原理是增量压缩技术，如果其基础帧，这里就I帧拿不到或者丢失，那么即使后面的B/P帧正常收到，也无法解码，这些帧就废了。等到下一个I帧到达后，图像才能开始，但这个时间因为受到组播定速发送的限制，是秒级附近，所以我们在看数字电视的时候，频道切换不如模拟电视快，很不爽。优化技术原理也很简单，就是在系统得到频道请求的同时，先用单播把最近一个I帧甚至其后的组播流量真正到来前的B/P帧给加速发过来，如果只有I帧，很多情况下就会有马赛克，如果不加速，那么就不能赶上组播的速度，和组播同步的时候还是有马赛克，而加速带来的问题就是，组播来了以后，要有一个减速，具体效果是有快进感，在技术实现细节上，可以通过修改视频帧的时间间隙参数使这个快进尽量平滑一点，这样处理后的视频，效果基本接近传统电视了
（2）视频丢帧问题：数据通讯丢包是正常的，所以在协议设计上都有相应的重传机制，而组播因为其性质决定了没有该特性，但丢包怎么办？优化方法就是增加了单播的重传机制，此时需要STB要感知到丢包并发出请求，这在早期的STB可能是不支持的，如果是基于Linux的STB，那么一般是可以通过Java做一个补丁支持该功能，否则就可能需要STB上游的设备支持一个proxy，但能否搞定不是很清楚
以上的功能需要IPTV的headend服务器提供相应的功能配合，在用户数越来越多的情况下，对集中式的服务器的压力和带宽浪费都是问题，尤其是很多人同时打开电视的时间

段，比如晚上的某段时间，或者热点节目的时候等。解决办法是可以在承载网络的一些节点上增加一些视频cache卡，这种情况下，这个节点应该是可以被STB通过L3访问到的，当然通过L2也可以，那会导致很多终端用户和这个视频存贮节点在一个大的L2 domain，这是很不好的网络设计，所以原则上这个节点应该是L3节点，而最好不用纯L2节点。上面只是一些简单的分析和描述，但也可以看出为了保证达到和传统电视一样的QOE，IPTV很麻烦

因为结构上定了上中下，只好把下写完，可能会比较鸡肋一点

补充一点TE的内容：TE因为比较复杂，所以TE LSP/TUNNEL在早期的capability是不大的，一般都建议用在CORE，而一些tier1的运营商CORE也很大，后来capability扩大了很多，但是如果要扩大到全网的TE，PE上问题未必很大，毕竟，并不需要和所有的PE都要全链接，但是P上可能有有些恐怖了，因为越是核心的P，越是有很多PE互联要经过它。所以仍然需要一些TE交换/分段PE/分层BGP等技术来解决扩展性的问题。在具体部署上，客户既希望能手动控制路由的选择，又要求能自动完成部署

曾经在这个大章的开始，痛贬C65/76，其实是有失公允，人在一定情绪和情势需求下的语言和文字虽然不免精彩，但也难免偏颇，对于C76/65就是，其实C76/65的成功是非常好的case，低价单板和高价单板共平台，在市场上有很多好的效果，一些实力有限的运营商，喜欢便宜的定西，不会一次就买一个功能很好的东西，虽然许多功能现在不用，但可能以后用。那么OK，给你便宜的单板，但后来当需要新功能的时候，只要合同没问题，那么需要物理更换成贵的单板，C合情合理很爽的又隔了一茬韭菜。类似的这种手段，在C的模式影响下，虽然大家也都在一定程度上使用，但还是C用得最娴熟老道

AL的新ME方案有了成功的开始并打下了良好的系统架构和方案架构，那么接下来的完善相对就比较容易了，下面简单介绍一些ME上用的技术

1、 ETH OAM问题：802.1ag(CFM)/802.3ah（EFM）/Y.1731，其中以Y.1731最全，但协议规定的配置方法比较复杂，尤其是802.1ag和Y.1731定义的MEP/MIP等等一系列概念最麻烦，如果完全按照协议规定的方式去做配置命令，本身就会带来一些OAM问题，所以实现的时候最好做UI上的简化处理

2、 环路保护问题：VPLS本身通过水平分割来解决环路问题，但是并不能阻止CE网络本身带来的路由环路问题为城域网的影响，尤其是在CE双归的时候，环路的几率就变大了，所以需要一些机制来处理，比如ALU的mac flapping（部分的解决），或者通过单独loop检查报文，再有就是STP in VPLS，把STP跑在VPLS里，感觉就很烦，总之这些解决方案都不是很好，最好在网络方案上避免环路，从这一点，个人不是很喜欢VPLS，纯VPLS带来的麻烦和好处可能是一样多

3、 MAC地址问题：VPLS带来大量的MAC地址，所以只好想办法解决，一个是拼命扩大MAC地址容量，比如到1M；再就是在某些场景下并不真正需要MAC学习，那么禁止VLAN内MAC学习，还有就是把PBB用在VPLS里，解决某些特殊场景下的MAC容量问题，和环路一样，这些方法都让人感觉不爽，以太网的一些基本问题，新ME的VPLS在实际网络应用中其实并没有很好的解决，所以个人不喜欢VPLS

4、 Reliability问题：SMARK LINK是思科的首创，TRUNK是标准协议，而MC-TRUNK是ALU的首创，本身其实并不复杂，SMART LINK可能是和MC-TRUNK类似的东西，但具体内容没看，目前更多的是MC-TRUNK
5、 slow protocol处理：慢协议其实就是native Ethernet上的L2协议簇，需要能灵活的定义那些透传，哪些终结
AL的在ME的是市场份额是不错的，但是在其整体的revenue中，应该是可以忽略的，AL最新在有一些做cluster向CORE扩展的roadmap，但整体上，还没有看到其全面进军数通市场的决心和策略，这一点，和华为最近的表现是有一定区别
另外提一点电力上网和cable上网，忽悠是不顶用的，电力的强弱电干扰问题和cable的共享网络结构问题，目前基本上没解，所以只有很少的电力和cable上网用户。
本来靓丽的AL新ME，用此章草草收尾，似乎预示着这个ME方案也该变一变了。

另类玩法PBT

刚刚有人“踢馆”，遇到此类事情，难免不爽，我也远非非常大度的人，所以回复里也含讥讽，想来何必呢，所以在此篇开头正式向那些同学表示道歉。
应该总结一下AL新ME的问题：
1、 技术上个人认为主要是VPLS带来的问题，具体细节前面都分析过了，不在赘述
2、 成本上主要是过高：原因是AL支持的都是海量的MAC/FIB/queue等，加上复杂的功能，不可能迅速降低成本，这也是AL把同等产品软件阉割一下推出比7750低价的7450的一个内在驱动。这还是没有把MPLS到运营商最边缘的情况下，如果到了最边缘，会如何呢？这是个大问题答复比较复杂，后面再讨论
AL新ME的创新，其实并没有脱离IP网络本质的巢穴，在增加业务支持和可靠性的同时，不可避免的带来了网络的复杂性，这在具有多年传统网络概念的电信运营商来看，总是觉得其本质有问题，需要改造，所以很早就有人提出IPTN(IP电信网)的概念，当然也会有一些新的draft、专利出来，概念本身并不复杂，IP节点傻瓜化，所有路由集中控制集中下发，IP节点只管按照管理中心下发的路由指示去转发，包括QOS参数，别的不用管，这不是很简单吗？这种idea现在已经演变到FOCE的标准簇，但仍然没有得到广泛部署，很难说是什么原因？思科因设备简化导致卖不上好价而消极？背离IP节点自智能的基本原则？标准不成熟历史网络难以改造？
在AL新ME火热的时候，其设备贵，网络复杂的弊端也自然会暴露出来，这对运营商在降低成本的压力下，是有考虑一些可能解决方案的驱动的，其中BT就是最激进的代表，BT本身在tier1运营商里排名不高，但是确是较早提出下一代网络规划的运营商，称为21CN，北电在种种不幸里，在数通领域已经没有什么顾及了，两者组合在一起，提出PBB的L2网络，基本思想和IPTN类似，只是利用了MAC IN MAC代替IP/MPLS路由，MAC IN MAC是为了解决对客户MAC学习带来的MAC容量问题，通过PMAC隔离CMAC，这个专利就像铅

笔上加一个橡皮头一样简单，所以被日本人发明了，并且成本基础专利，让喜欢专利的人很生气，还好，现在基本也没啥用了
PBB需要一套独立的网管，来学习所有的MAC地址，然后下发到各个L2节点，各个L2节点是各种size的傻瓜，除了和网管的通讯协议，别的什么协议都没有，是个傀儡，傀儡能卖多少钱？很明显，思科是绝对不会傻到大力支持这个方案的，其实除了一无所有的北电，借椽登槐的华为，没人真的喜欢PBB。PBT在BT表面的火热中，已经演进到了PBT(PBB TE)，有一些技术问题比如组播等还没有解决，当然，随着网络的真正商用，也许会有更多的需求，但未必是大问题。在风风火火了几年后，BT终结了PBT，几乎在一夜之间转向了AL的7750，并且一路扩展。想来是BT高层对PBT迟迟不能商用全面部署影响了这几年业务的迅速拓展十分不满，所以很快勒令下马，迅速开始AL 7750，这个结果让一些人真的很伤心，陪太子读书很多年，结果成了曹雪芹，不知道有没有红楼梦可以塞翁失马，聊作安慰。
目前还有类似思路的有T-MPLS/G-MPLS等，都不是IP vendor真正喜欢的，具体没有仔细研究。
顺便提一个城域网传输中的一朵小昙花：LDMS，无线光网，这种无线也就是微波类似的技术，无论怎么做，带宽都是有限的，所有偶尔有些特殊的情况用一下，没有任何规模化的可能