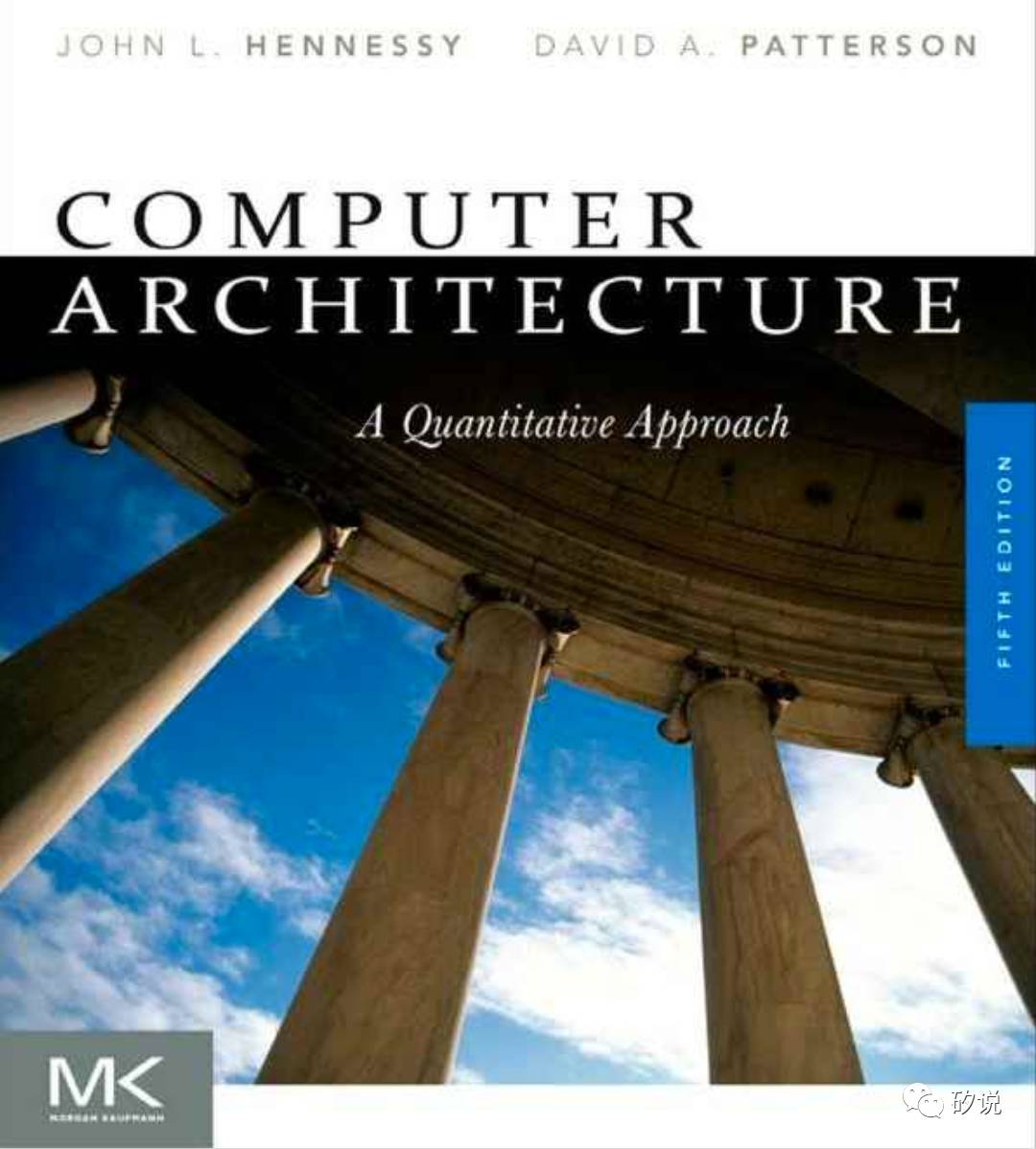
**RISC-V与DSA！ 计算机架构宗师Patterson与Hennessy 演讲实录**

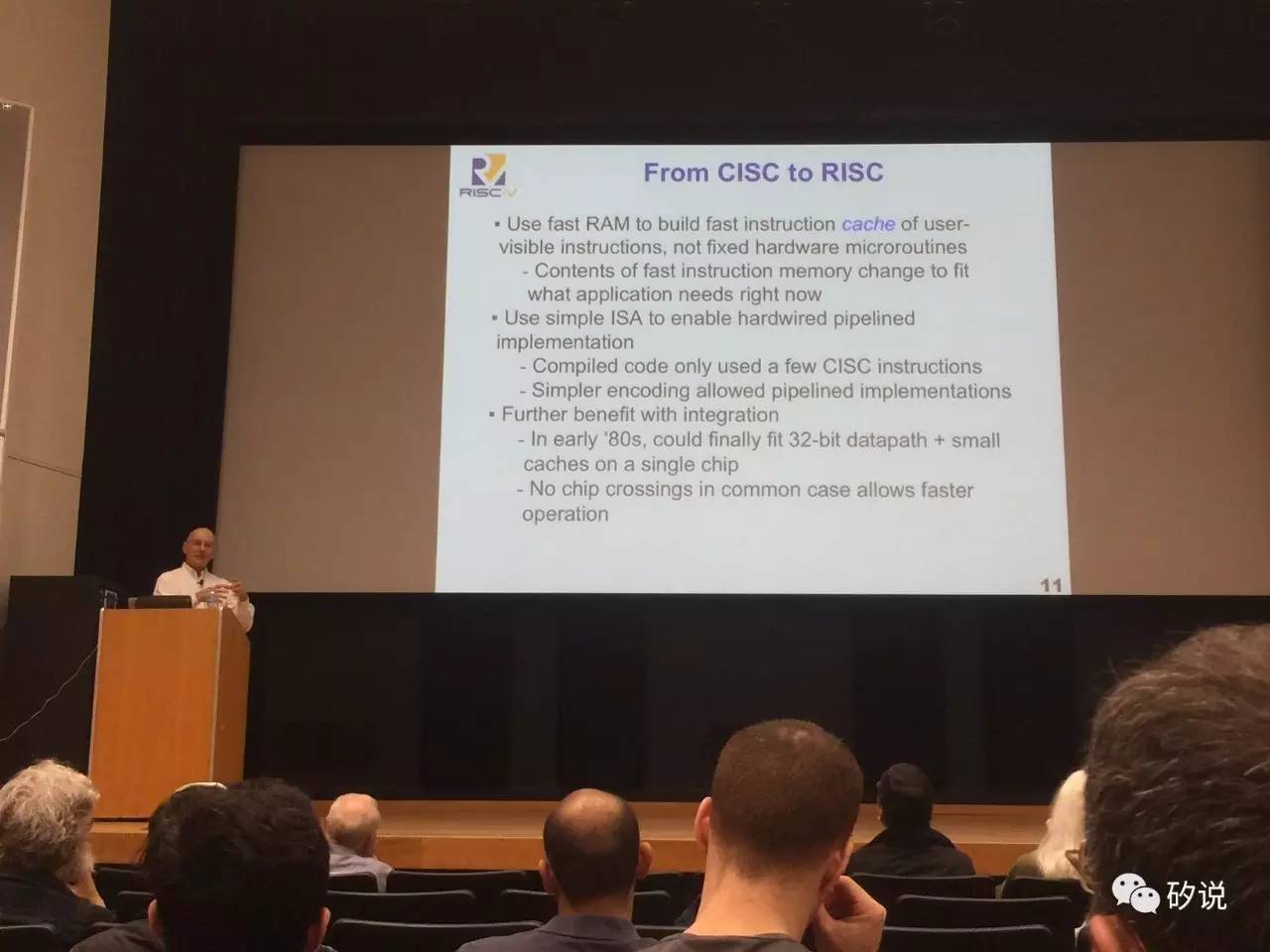
2017-03-25 09:26 [设计](http://www.sohu.com/tag/69968) /[计算机](http://www.sohu.com/tag/59959) /[移动芯片](http://www.sohu.com/tag/65771)

相信所有和计算机体系结构打过交道的朋友们都看过David Patterson与John Hennessy的煌煌巨作,《计算机体系架构：量化研究方法》。两位在计算机架构领域鼎鼎大名的教授，一个来自加州大学伯克利分校，另一个来自斯坦福。



《计算机体系架构：量化研究方法》自1990年第一版出版后，如今已经过了近30年，2011年第五版出版，目前Patterson和Hennessy正在准备第六版的书稿。计算机处理器领域沧海桑田，1990年时处理器是最热门，最前沿的科技，是无数年轻人最神往的方向，摩尔定律推动芯片时钟频率蒸蒸日上，未来一片大好。到了今天，处理器的技术含量仍然是所有科技里面最高的之一，但是早已不是最热门的科技，Intel挤牙膏式的工艺演进计划预示着摩尔定律的终结似乎就在明天，处理器的未来如何人们众说纷纭。Hennessy从斯坦福校长的位置上退了下来，Patterson也从伯克利退休。两位巨星级的人物在三月份再次在斯坦福聚首，为我们带来了一次精彩的演讲。矽说的作者有幸在现场聆听了这次演讲，记录如下。

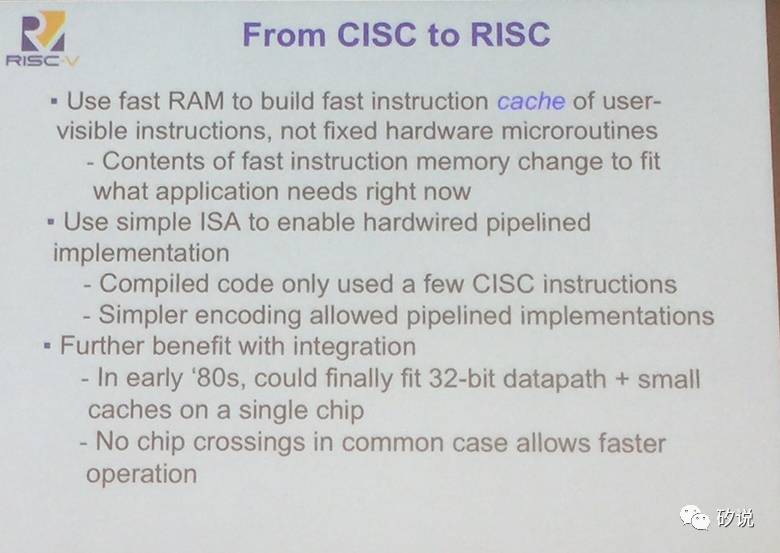
首先上场的是David Patterson，为我们带来了关于指令集架构（ISA）的回顾以及RISC-V项目的展望。



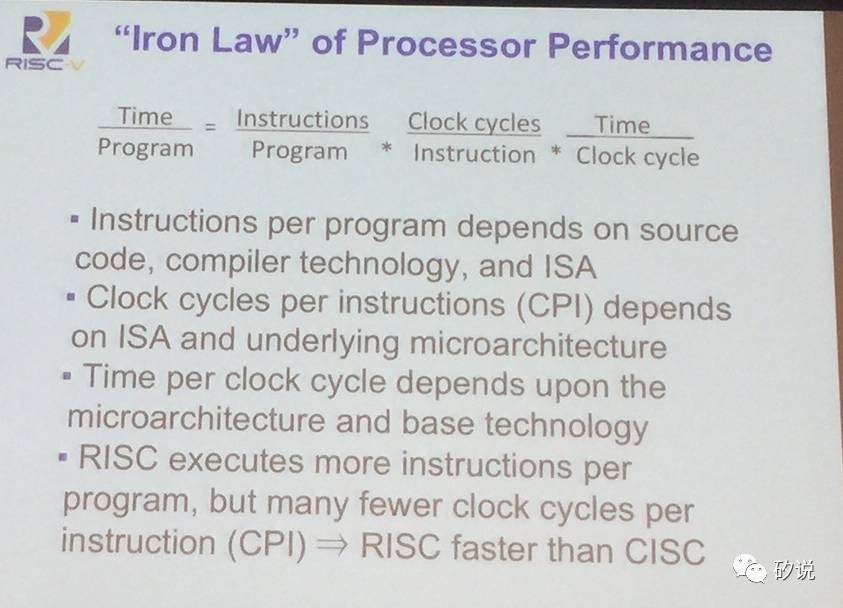
（Patterson教授在演讲中）



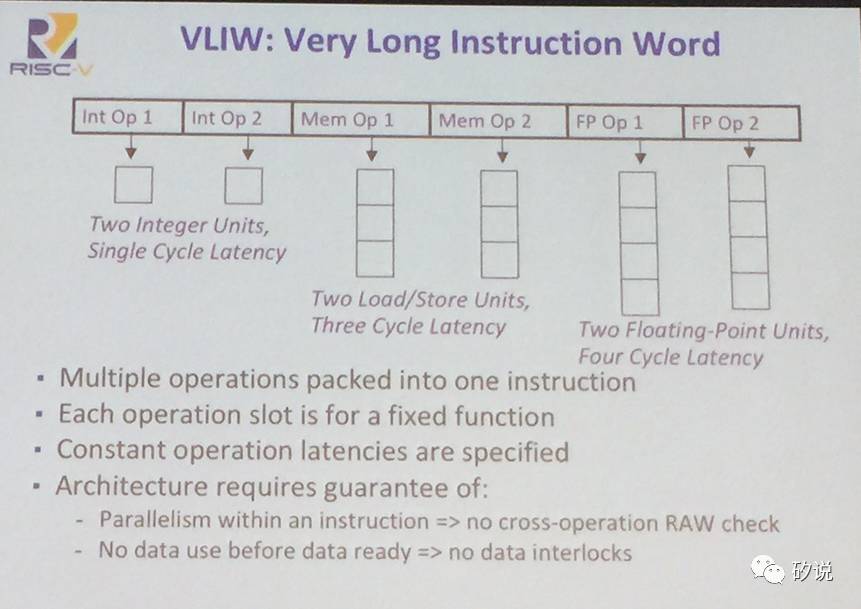
Patterson教授的演讲主题是50年来计算机体系架构（ISA为主）的回顾以及RISC-V架构。



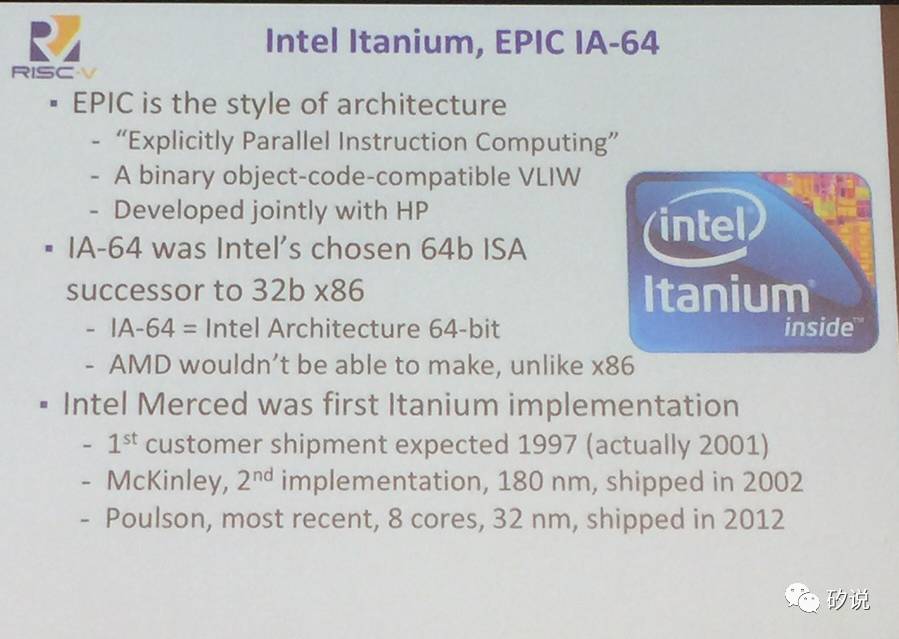
Patterson教授首先回顾了ISA的发展史。在计算机发展之初，ROM比起RAM来说更便宜而且更快，所以并不存在片上缓存（cache）这个东西。在那个时候，复杂指令集（CISC）是主流的指令集架构。然而，随着RAM技术的发展，RAM速度越来越快，成本越来越低，因此在处理器上集成指令缓存成为可能。RISC的出现可谓水到渠成。研究发现计算机执行大多数程序时CISC指令集中绝大多数指令都只在极少的时间才被用到，因此专门为这些指令设计硬件并不划算。相反，使用精简指令集（RISC）可以大大简化硬件的设计，从而使流水线设计变得简化，同时也让流水线可以运行更快。



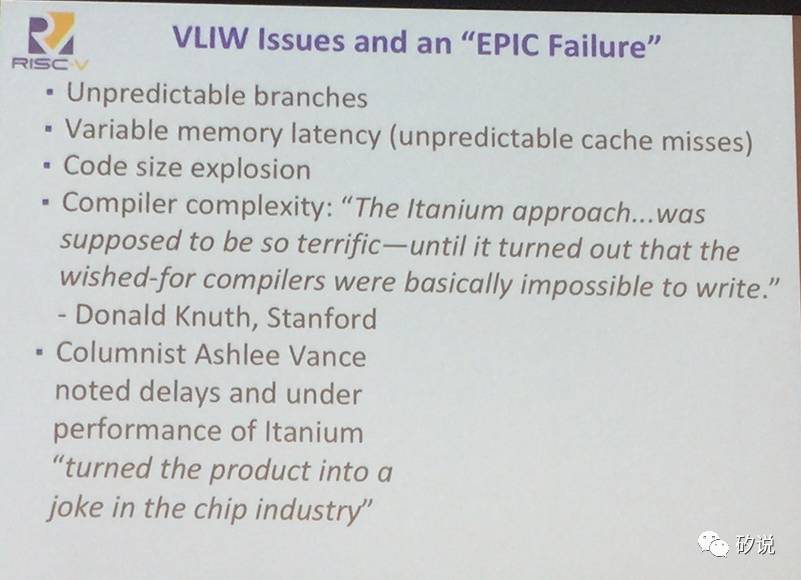
Patterson教授再次重申了评估处理器性能的指标，即程序运行时间。程序运行时间由几个因素决定，即程序指令数，平均指令执行周期数（CPI）以及时钟周期。程序指令数由程序代码，编译器以及ISA决定，CPI由ISA以及微架构决定，时钟周期由微架构以及半导体制造工艺决定。对于RISC，程序指令数较多，但是CPI远好于CISC，因此RISC比CISC更快。



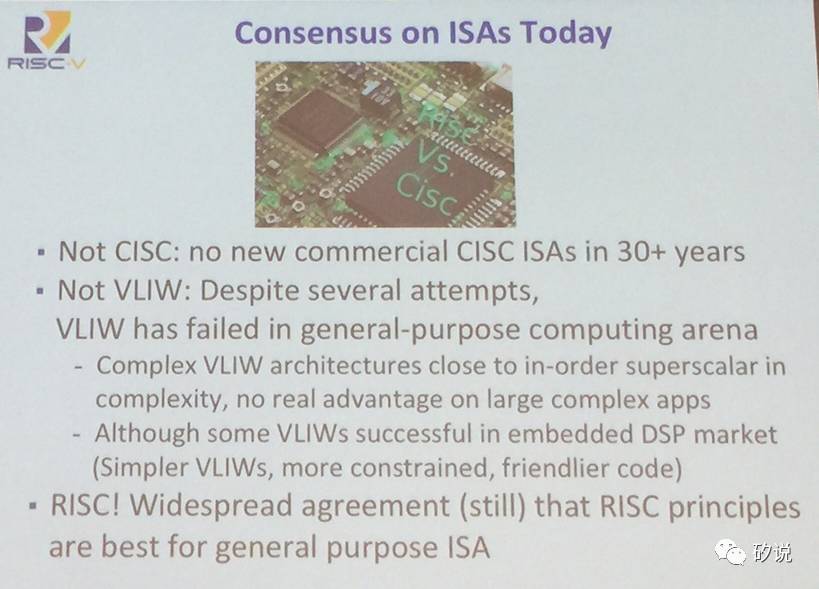
除了CISC和RISC之外，另一种流行（过）的ISA是超长指令字（VLIW）。VLIW把多个操作放在一条指令里，因此需要一条指令中的多个操作能够并行执行。



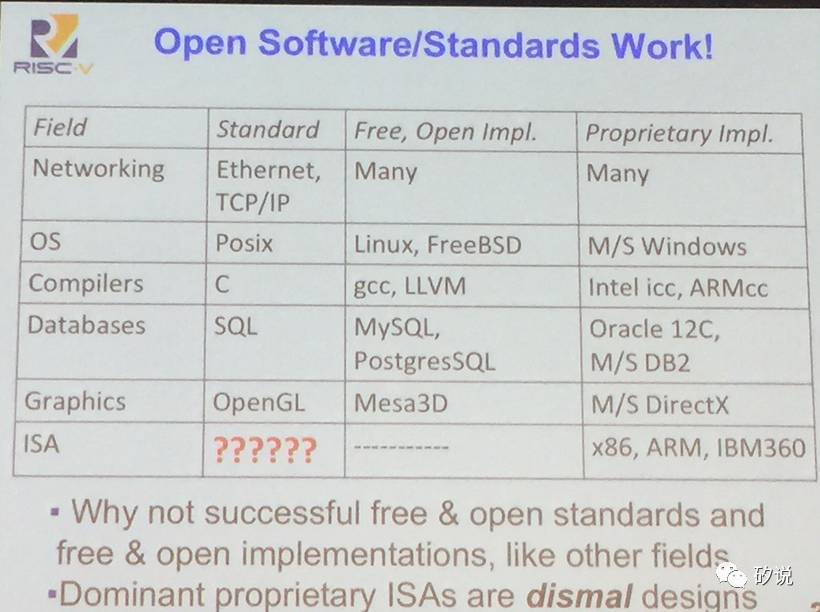
VLIW的代表是Intel Itanium（安腾），使用的架构代号是EPIC，开发的合作伙伴是惠普。安腾第一代Merced预期出货日期是1997年，实际出货时期为2001年；第二代McKinley使用180nm工艺，出货时间为2002年；第三代Poulson，也是最近的一代，8核心使用32nm工艺，2012年出货。



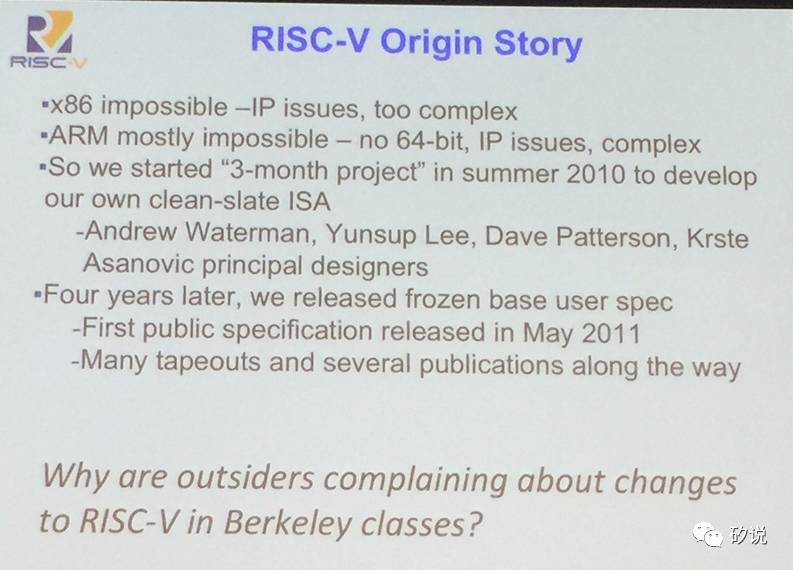
然而，VLIW架构遇到了巨大的失败。VLIW的问题，包括分支预测困难，Cache miss无法解决，代码爆炸以及最关键的，编译器过于复杂以至于无法实现。斯坦福的Donald Knuth（计算机科学领域又一位传奇人物）表示，“安腾看上去很棒，但是编译器根本没法写！”



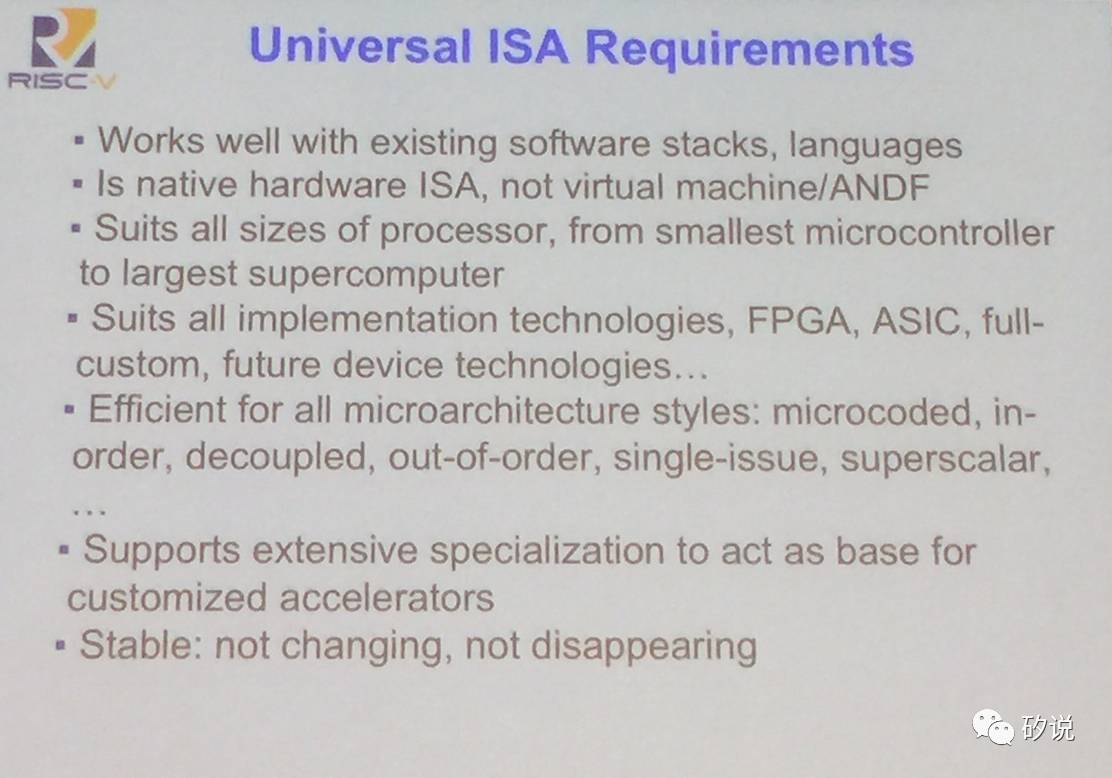
目前处理器的ISA，已经30多年没有新的CISC ISA出现（Intel x86表面用的是CISC但是内部有硬件把CISC转换成RISC再真正执行）。VLIW在一些嵌入式DSP市场获得应用，但是在其他的市场都没有获得成功。考虑到处理器的绝对数量，目前最主流的通用ISA还是RISC。



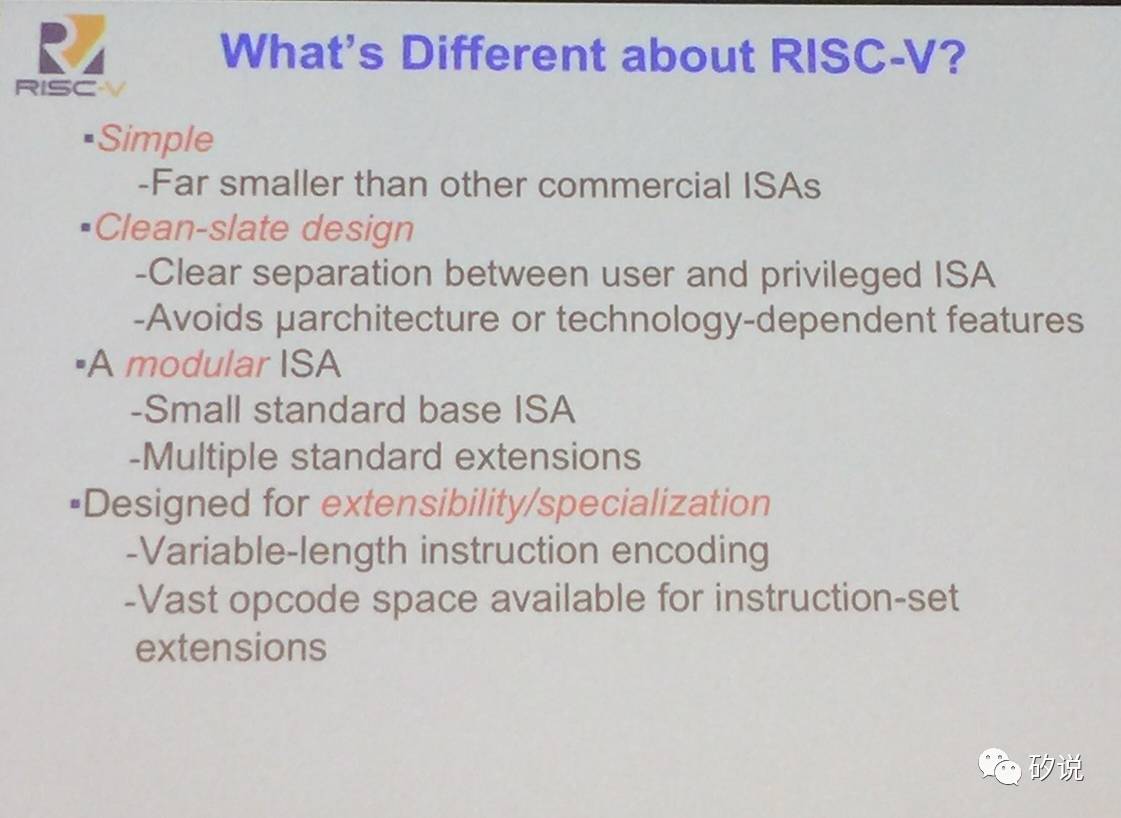
回顾完ISA的历史，再来看看目前ISA的生态。这里把ISA和网络，操作系统，数据库，图像标准库作了比较，可以看到网络，操作系统，编译器等等领域都有主流的标准，基于该标准同时有开源免费的版本，以及商用的收费标准。然而，在ISA领域，之前并没有公认的标准，也没有开源免费的ISA，仅有商用的ISA，这让整个ISA领域的生态显得死气沉沉。



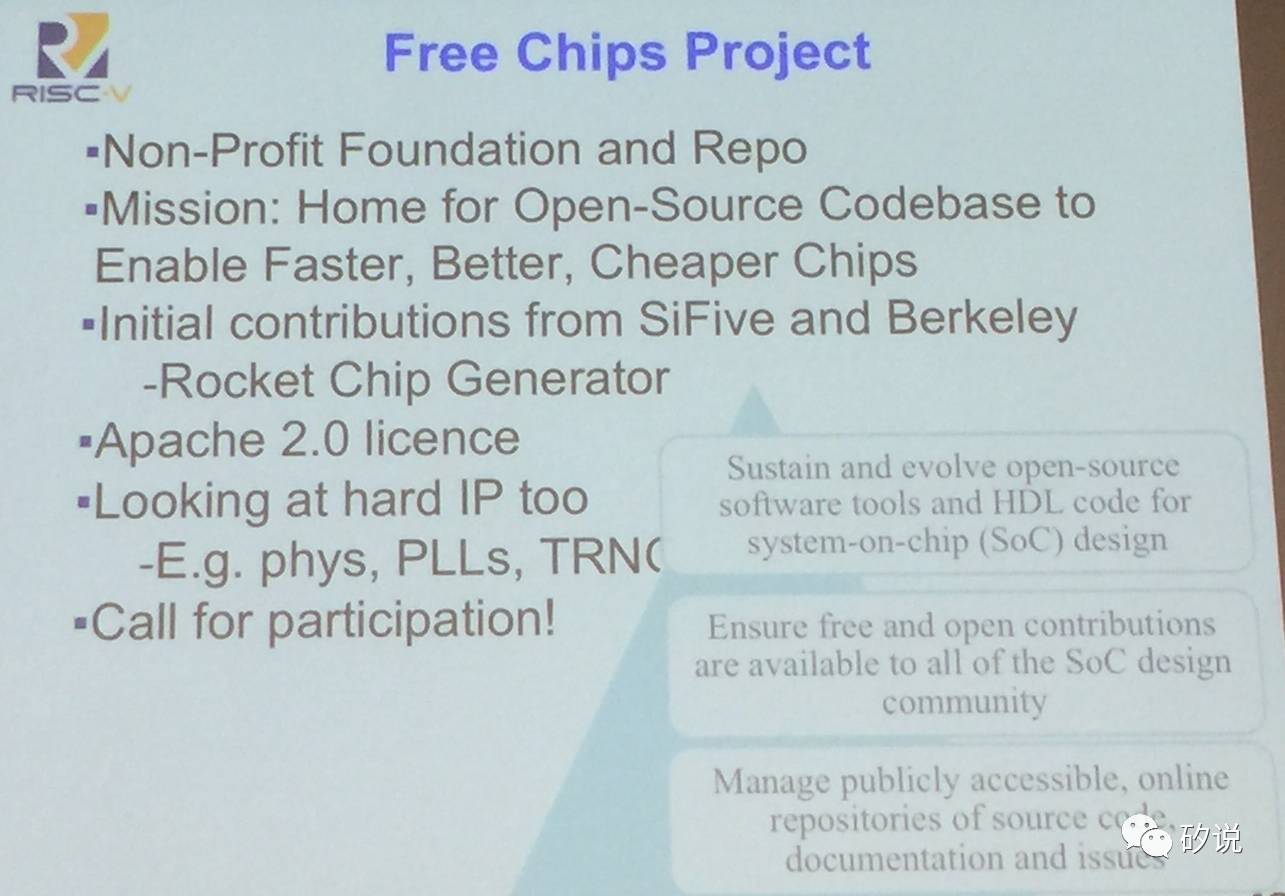
于是，RISC-V应运而生。要做开源的ISA，基于x86和ARM都几乎不可能，因为它们都太复杂，而且还存在IP的问题。在2010年夏天，Patterson教授带领团队开始从头开始设计一个干净的ISA。经历了很多年，经过多次流片验证，终于在2014年发布了最终版spec，就是RISC-V（V是第五代的意思）。



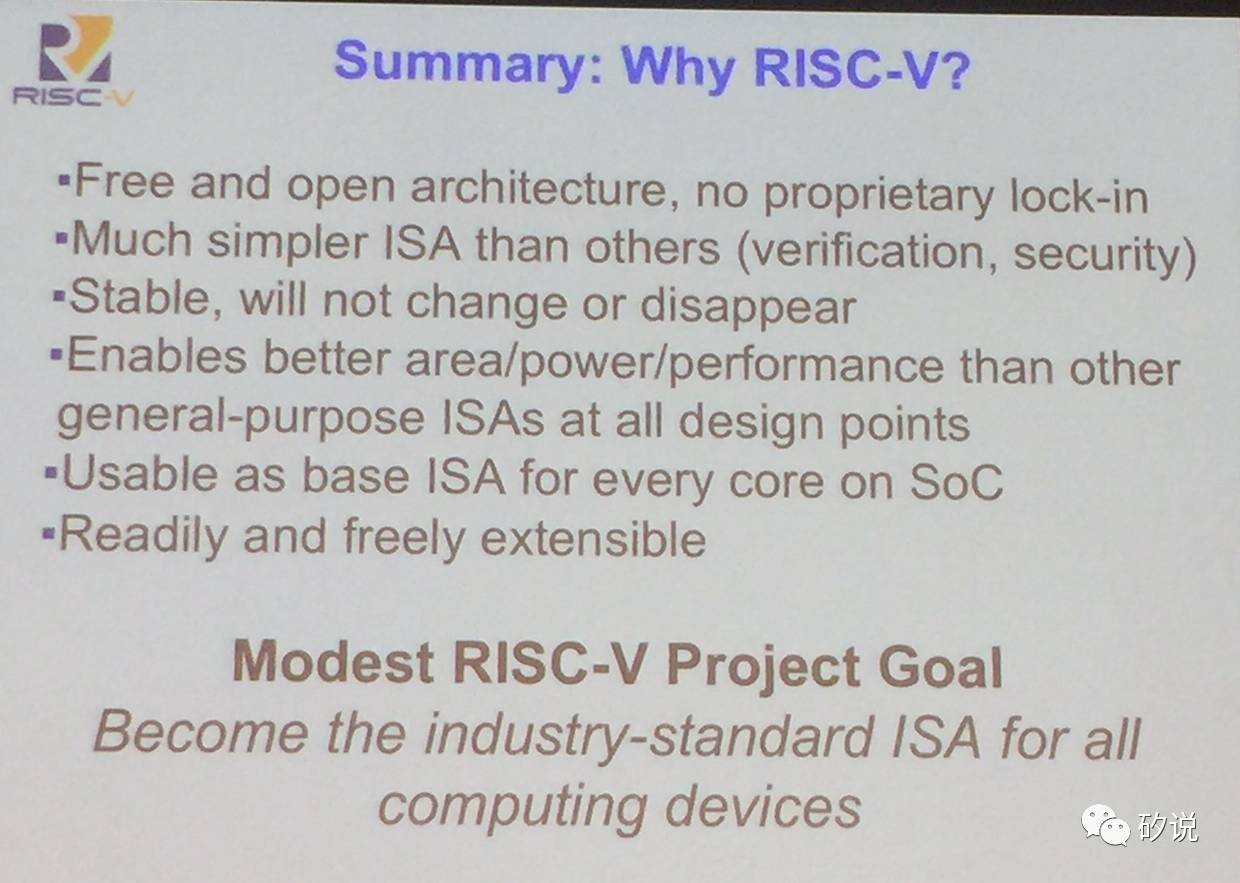
RISC-V作为一个开源ISA，首先要满足对ISA的一般要求。首先，它必须与现存的主流编程语言和软件兼容。第二，它必须有直接硬件实现，而不是一个虚拟机。第三，它必须有很好的弹性，能满足小至微控制器（MCU）大到超级计算机的需求。第四，能与各种实现方式兼容，包括FPGA，ASIC，全定制CPU，以及未来的其他实现。第四，需要与各种微架构配适，包括有序执行，无序执行，单发射，超标量等等。最后，还需要满足可扩展性（可以作为基础ISA，在特殊用途中加上额外的增强ISA），以及稳定性（不会一直变化，不会突然消失等等）。



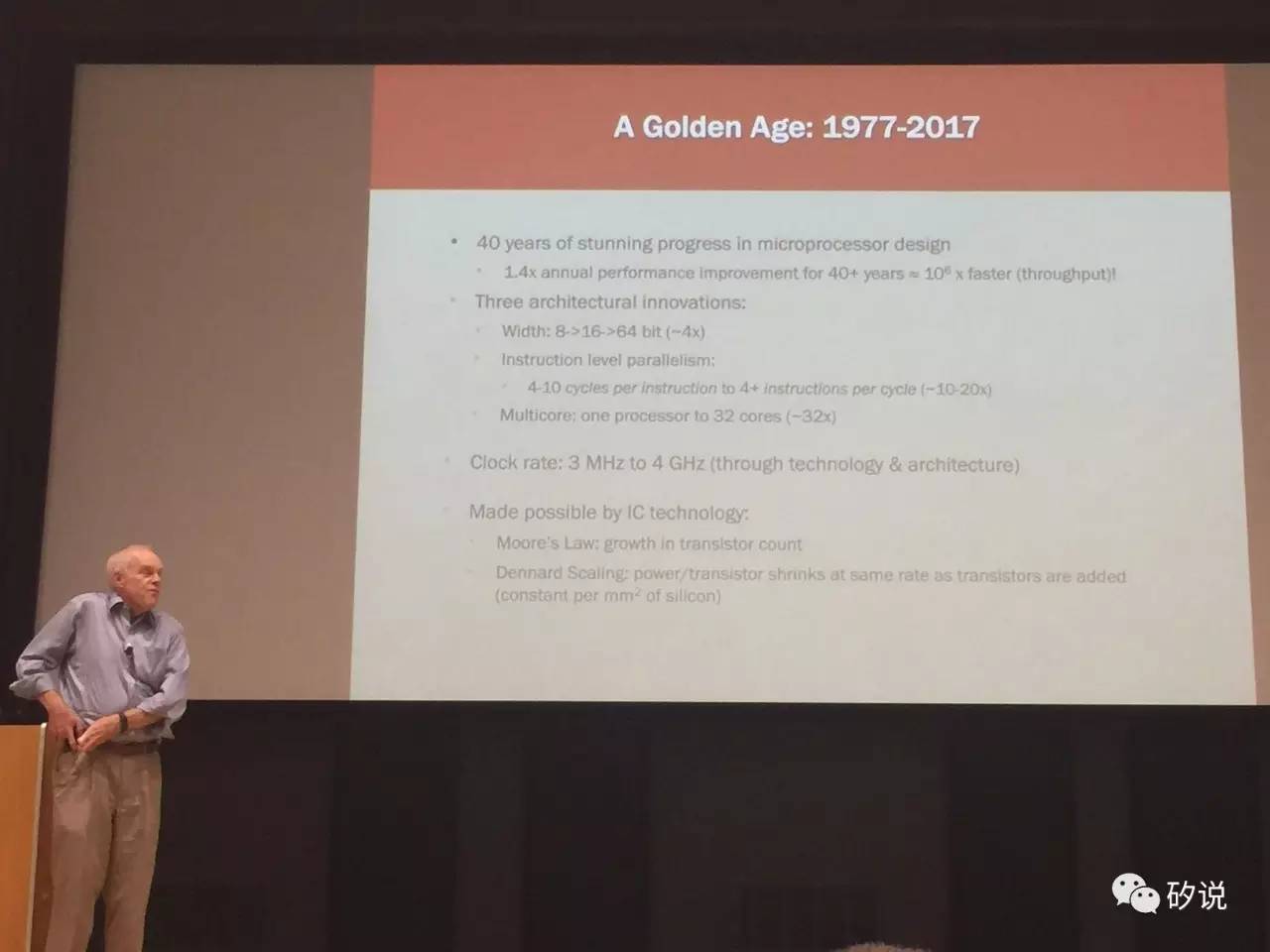
除了满足一般的需求外，RISC-V还有自己的特色。首先，它很简单，比其他的商用ISA规模都要小很多。第二，它很干净，例如在用户与特权ISA之间泾渭分明，有非常清晰的界限。另外，RISC-V中没有与微架构或实现方式有关的特性，因此具有普适性。第三，RISC-V是模块化的ISA，它的基础ISA集很小，但是可以根据用户需求去加载扩展集。最后，RISC-V特别为了可扩展性和专精化做了优化，使用了可变长度的指令编码，并且有许多空间以供指令集扩展。



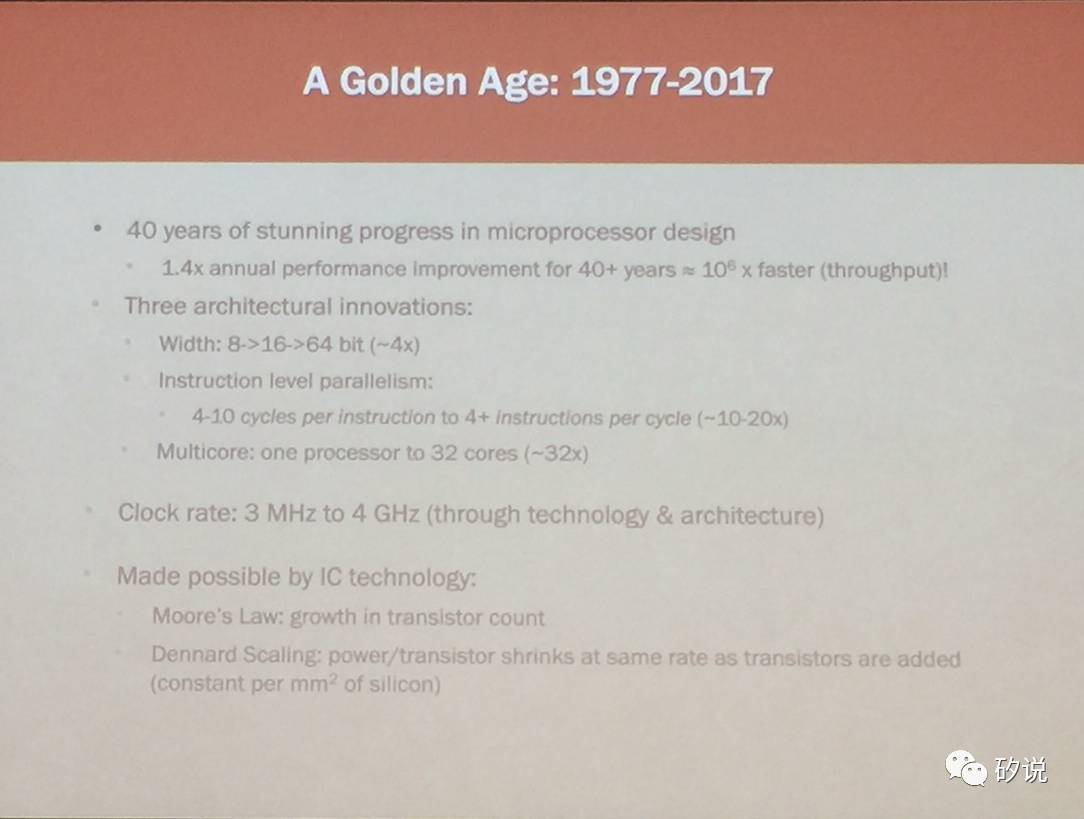
最特别的一点是，RISC-V支撑了一个开源的社区，包含了非盈利基金会以及开源代码库。RISC-V的愿景是未来各种灵活而低价处理器芯片的基础。RISC-V一开始的贡献者包括伯克利和SiFive（一家初创公司），目前在征求各类设计者加入开源社区，需要代码以及其他硬件IP（如PLL，PHY等等）。



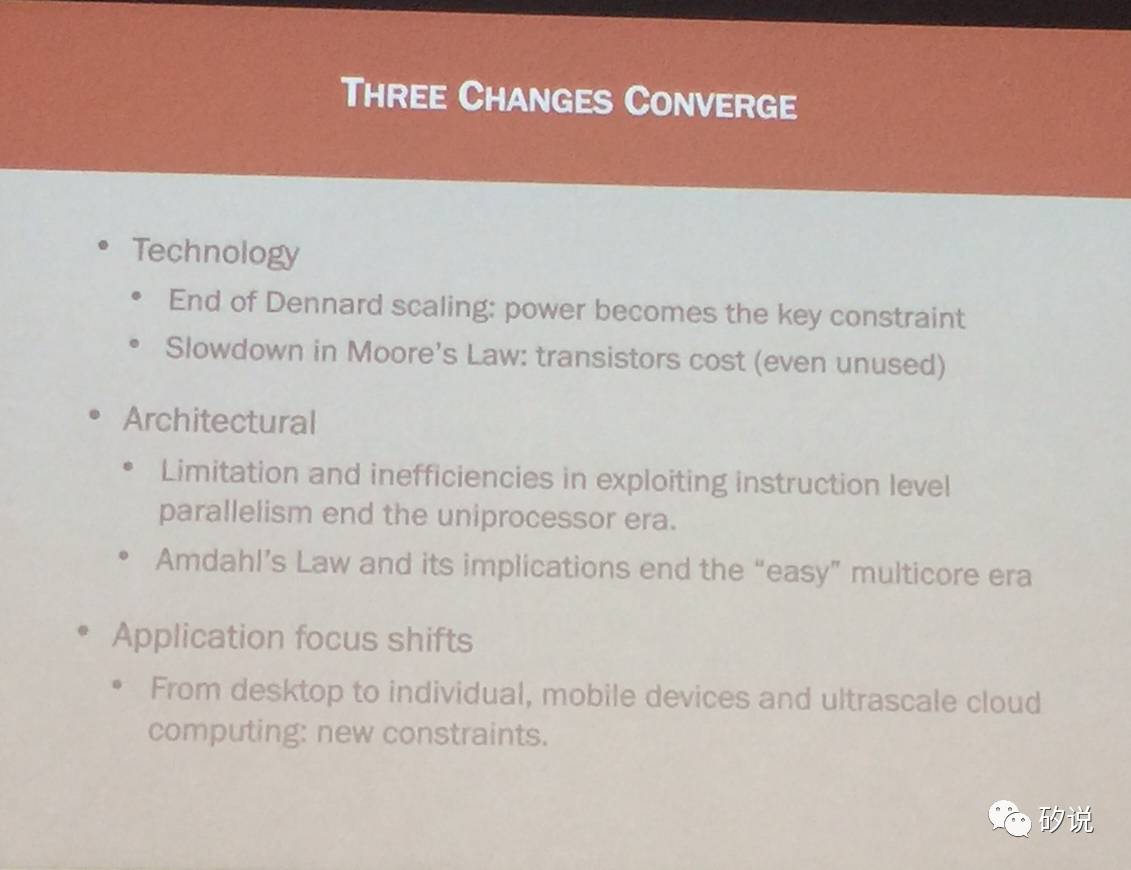
最后，总结一下几大使用RISC-V的理由。第一，RISC-V是免费开源架构，无须付费。第二，它的ISA比起其他ISA来说简单许多，因此验证起来也方便许多。第三，RISC-V很稳定，不用担心突然发生很大变化或者直接就消失。第四，RISC-V可以在各种设计中比起其他ISA更高效，面积、功耗和性能都更好。第五，RISC-V可以作为各种SoC核的基础ISA，而且第六，RISC-V具有很好的扩展性，可以随意按照需求扩展。现在RISC-V的小目标，是成为一种适合各种计算设备的业界标准ISA。



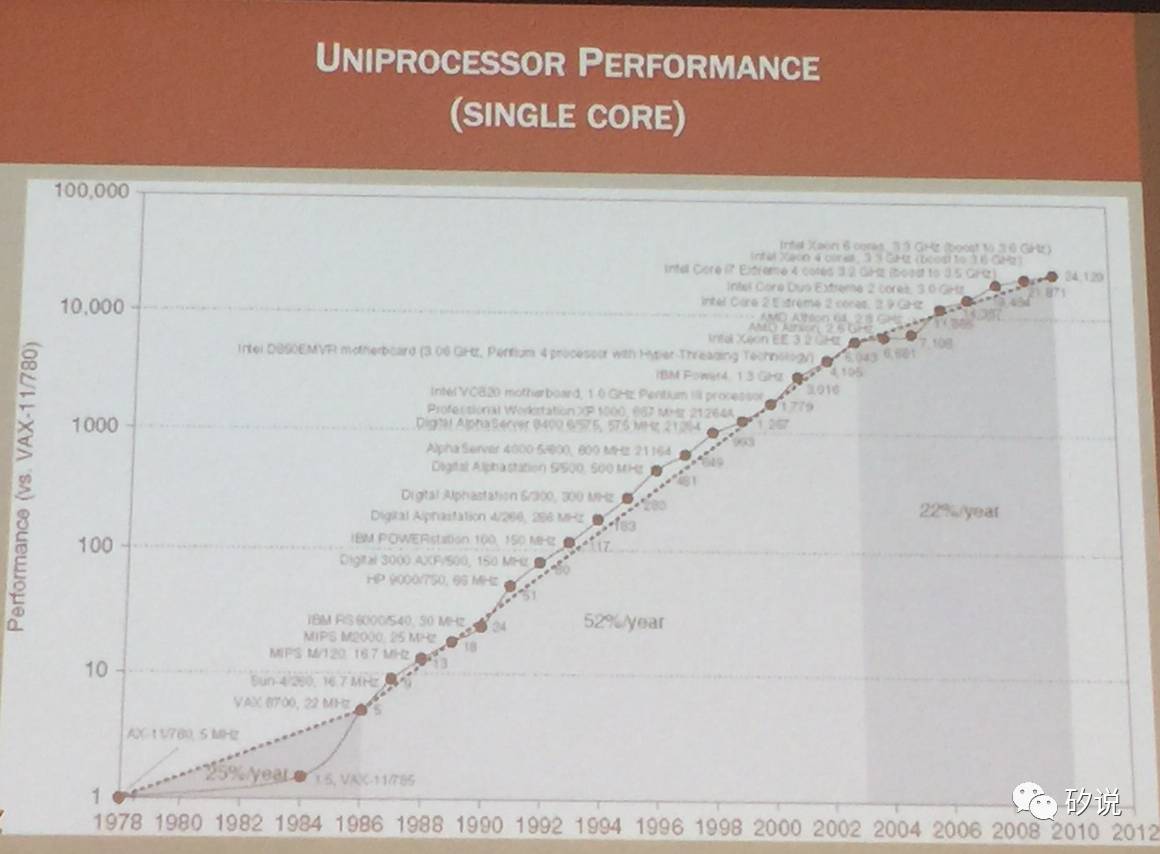
John Hennessy的演讲紧随其后。Hennessy教授的演讲在回顾了摩尔定律的发展之后，一针见血地指出了目前常规处理器演进遇到的瓶颈在于功耗，并且提出了目前处理器的新希望在于Domain Specific Architecture（DSA，即针对应用领域做优化的处理器架构，区别于通用架构）。



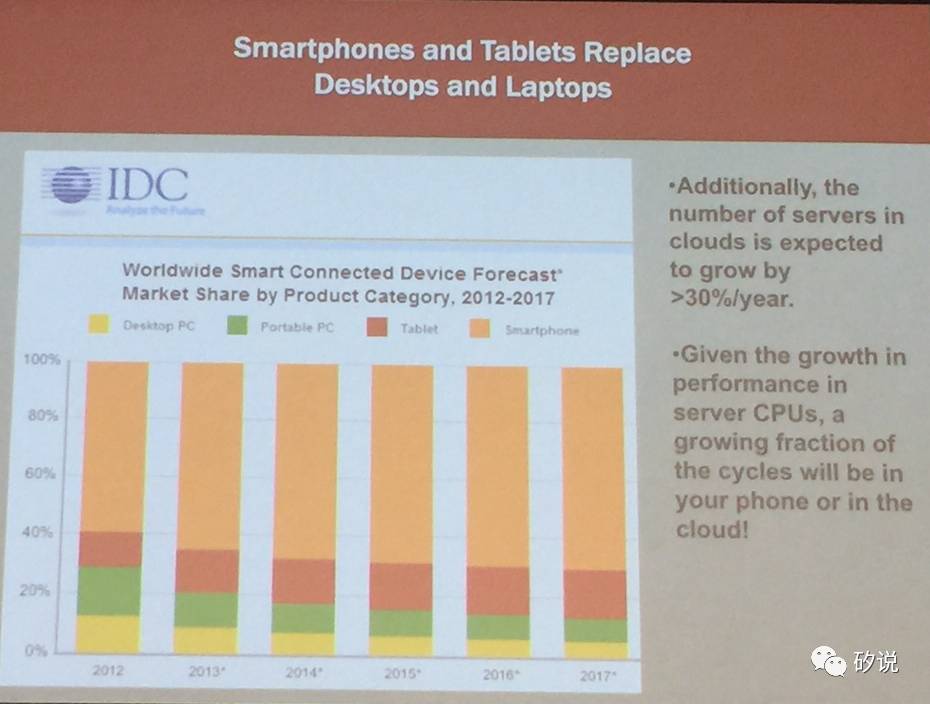
Hennessy教授首先回顾了四十年来处理器的高速发展史。四十年间，处理器性能以每年1.4倍的速度指数上升，目前性能相比于四十年前改进了约一百万倍。在处理器架构角度，最大的进步包括位宽（由八位进化到了六十四位），指令级并行度（从最初每条指令需要4-10个时钟周期执行到现在每周期可以同时执行超过4条指令，这是10-20倍的改善），以及多核架构（由单核演化到32核）。从性能角度，时钟频率从3MHz进化到4GHz。这一切都是因为集成电路生产工艺进化为基础的。摩尔定律使得处理器晶体管数持续上升，但是因为晶体管功耗和晶体管面积缩小的速度基本相同，因此在前40年间芯片单位面积的功耗基本不变。



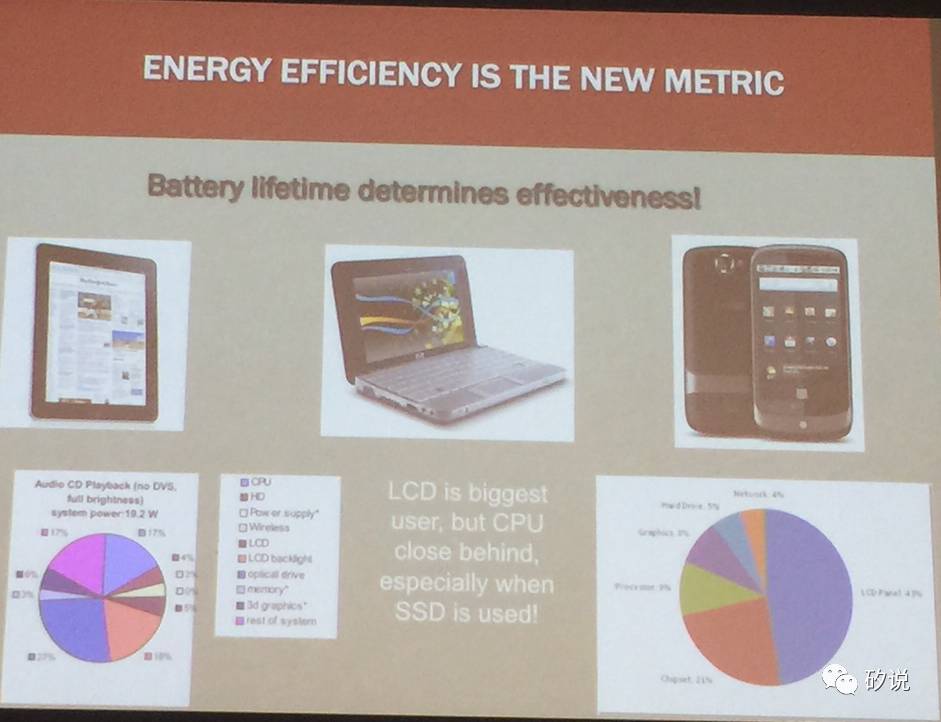
目前，三种技术趋势让传统的通用处理器演进遇到了瓶颈。半导体工艺角度，Dennard Scaling规律结束，芯片功耗急剧上升，同时摩尔定律减缓，晶体管成本不降反升。从架构角度，指令级并行已经到达极限，单核时代已告结束；而Amadahl’s Law提示多核架构的速度提升取决于程序中有多少部分无法并行执行，多核架构目前的速度提升也变得越来越慢。从应用角度，处理器的应用场景从原来的桌面电脑变成了个人移动设备和云端超大规模服务器，这也带来了新的设计约束。



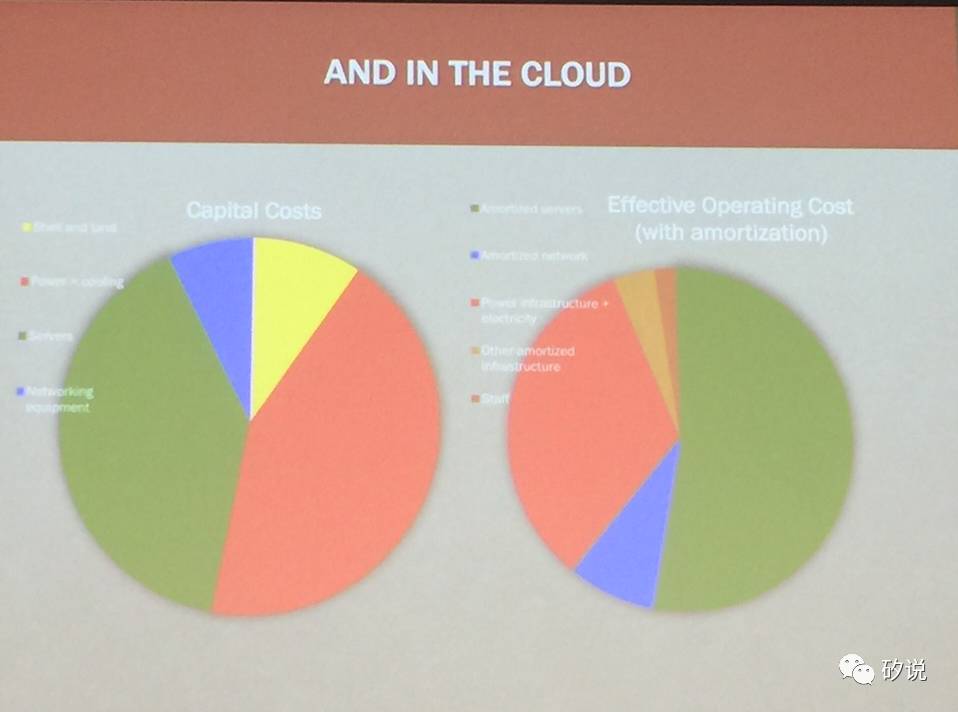
从单核处理器的速度进化趋势也可以印证之前的观点。从上世纪90年代到本世纪前五年，单核处理器的性能以每年50%以上的速度提升，而到了2005年后，但和处理器性能的提升速度降到了每年20%左右。



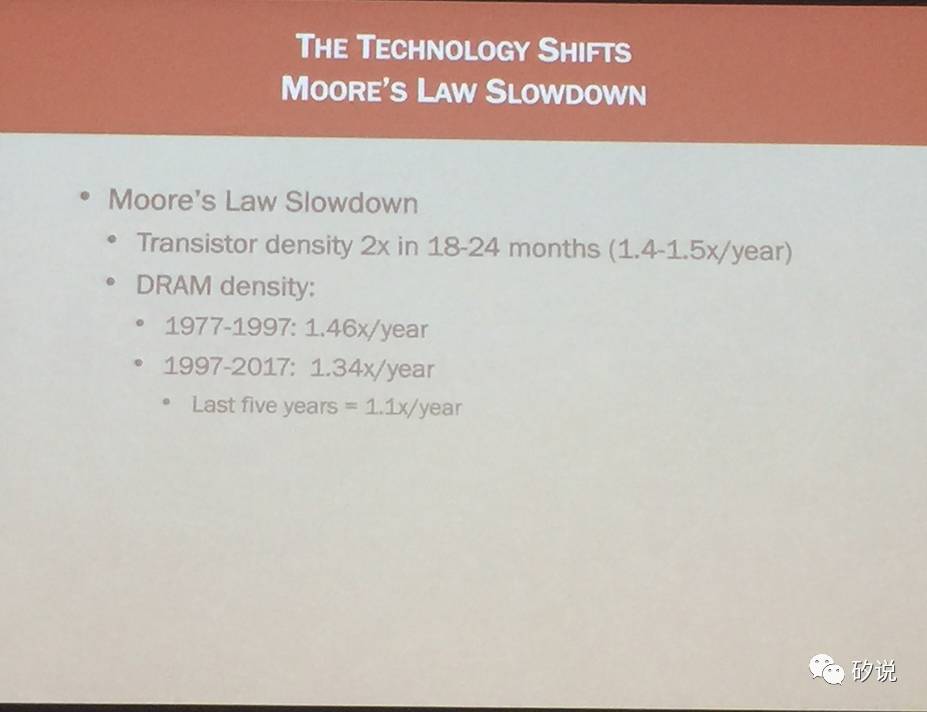
从单核处理器的速度进化趋势也可以印证之前的观点。从上世纪90年代到本世纪前五年，单核处理器的性能以每年50%以上的速度提升，而到了2005年后，但和处理器性能的提升速度降到了每年20%左右。



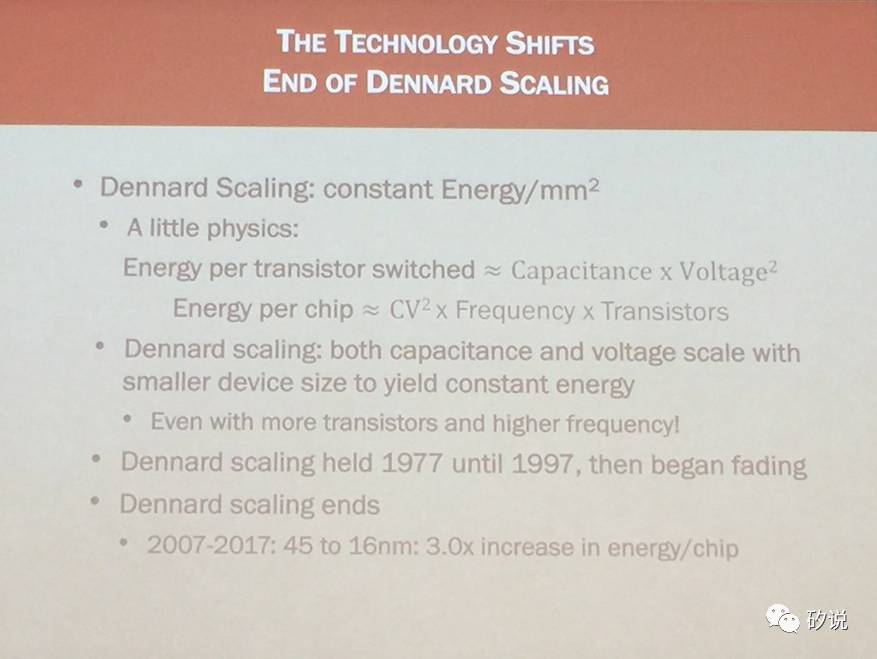
随着市场份额的变化，处理器的设计需求也发生了变化。能效比正在成为目前最重要的指标。在移动领域，由于电池容量的限制，必须注重能效比。目前，处理器在移动设备中已经成为继屏幕之后能量消耗最大的元件，因此移动设备中处理器能效比是最关键的问题。



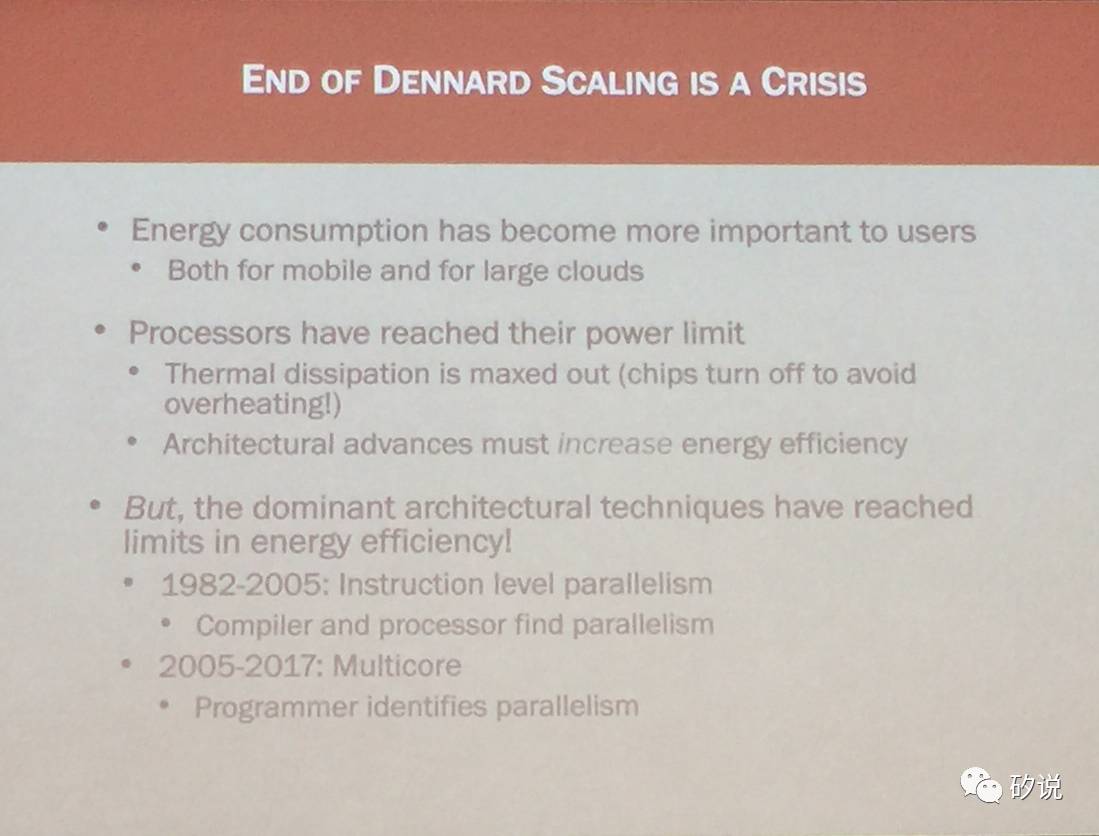
在另一个未来处理器最大市场——云端服务器市场，能效比也是最关键的指标。目前数据中心的成本中，散热已经成了最大的成本之一，为了减少成本必须考虑处理器能效比。



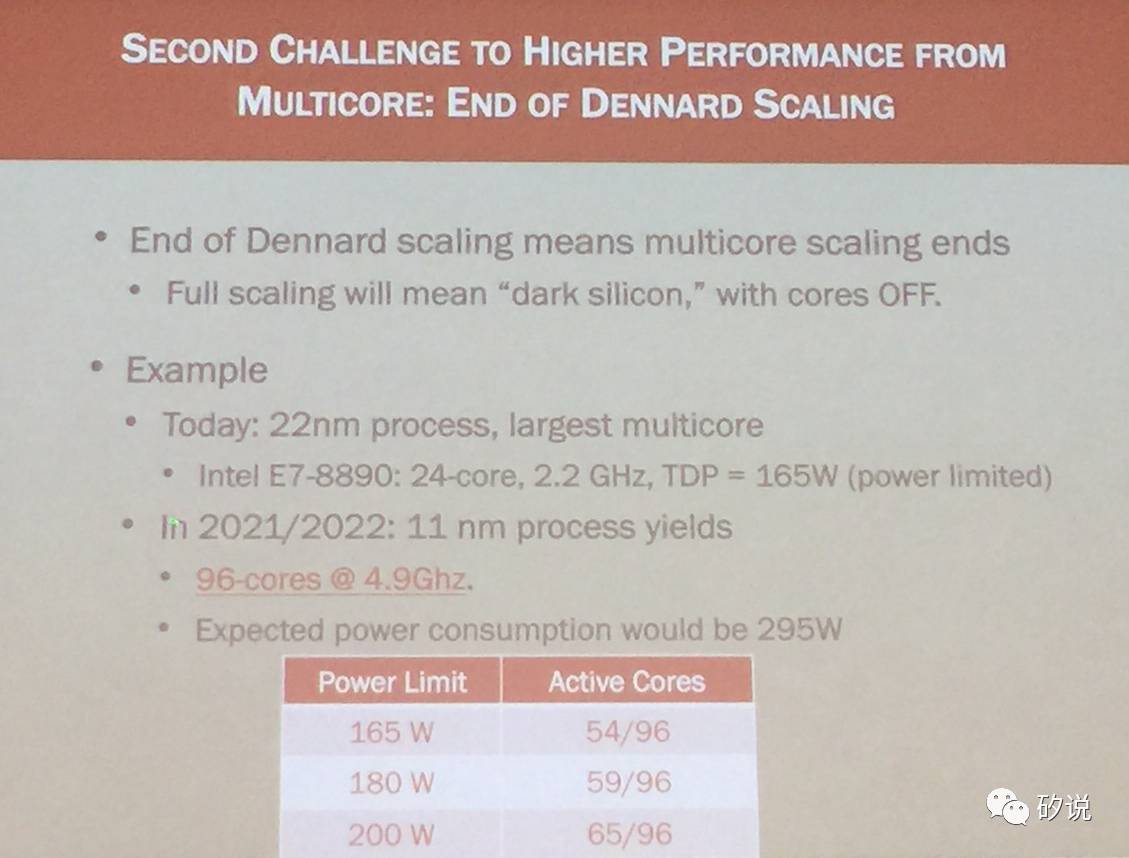
半导体工艺进化趋势的变化也很重要。摩尔定律遇到瓶颈是近年来半导体业最深刻的变化。DRAM密度变化在1977-1997年是每年1.46倍，1997-2017年平均密度变化是每年1.34倍，而在过去五年平均密度变化是每年1.1倍。一个更令人惊讶的事实是，DDR4标准DRAM的带宽虽然比DDR3大很多，但是DDR4内存的内部速度实际上比DDR3慢！这在过去半导体产业按摩尔定律蒸蒸日上的时代是难以想象的。



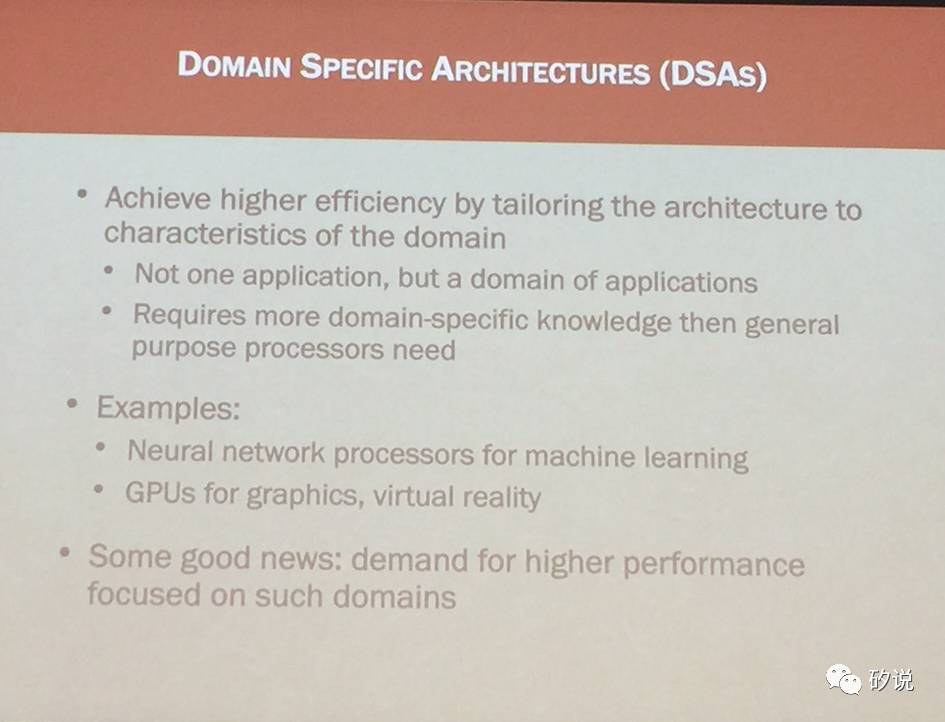
另一个关键的半导体工艺趋势变化是Dennard Scaling不再有效。Dennard Scaling是早期半导体工艺变化的规律，即将晶体管尺寸和电源电压一起变化，单位面积晶体管的总电容上升，但是电源电压在相应变小于是总体的单位面积能量消耗基本保持不变。Dennard Scaling规律从1977年保持到了1997年，在这之后慢慢失效，例如从2007年到2017年（晶体管特征尺寸由45nm缩小到16nm），每块芯片的总能耗变大了3倍。



Dennard Scaling的结束对于传统处理器设计方法来说是一个危机。能量消耗对用户来说越来越重要（无论是对移动设备还是云端服务器），而且处理器散热已经接近了极限。处理器架构必须改善能效比，但是传统通用架构设计方法的能效比已经到极限了。



Dennard Scaling的结束也意味着在能效比约束下，堆核数已经很难增加性能。再增加核数就会导致Dark Silicon，即芯片的许多核会很多时间处于待机状态，从而导致很高的成本（编注：例如目前移动处理器流行的大小核架构，通常同时只会打开高性能大核或者低功耗小核）。所有核打开时，处理器功耗非常大，会导致散热问题。举例来说，目前22nm工艺制造的最大多核处理器是Intel E7-8890，有24颗核心，运行在2.2GHz，最大功耗为165W。在2021/2022年，假设可以使用11nm工艺，96核心处理器运行在4.9 GHz，那么在165W功耗的限制下，只能打开54个核心，功耗限制放松到180W可以打开59个核，限制放松到200W可以打开65个核。但是要同时打开96个核，则功耗实在太大，很难实用。



为了提高能效比，一种很有希望的架构是针对应用领域做优化的专用领域处理器架构（DSA）。DSA的优点在于，可以为特定的一类应用（注意不是一种应用，而是一类）做架构优化从而实现更好的能效比。相对于通用（general purpose）处理器，DSA需要设计时考虑专用领域的特殊需求，也需要设计者能对该领域有深入的理解。DSA的例子包括为机器学习设计的神经网络处理器，以及为图像和虚拟现实设计的GPU。DSA设计将会成为处理器架构的新趋势。

**Q & A**

Q：请问除了之前演讲中提到的以外，还有哪些处理器领域的未来趋势是值得关注的？

A（Patterson）：我认为未来之星是深度学习领域的DSA处理器。深度学习的重要性我想在座的都很清楚。另外，使用更高级的设计描述语言，例如Chisel，来加速设计，也会成为趋势。

Q：摩尔定律的终结对于整个计算机领域的人来说，意味着什么？

A（Hennessy）：The easy ride of software is over. 这意味着软件行业的从业人员未来需要更多对硬件的理解。在之前，软件行业可以不用太关心硬件，只要把程序功能实现，就算现在的硬件不能跑，过一阵新的更强的硬件出现了一定可以跑。现在，软件必须认真考虑如何在硬件上高效执行的问题。会有更多domain-specific编程语言出现，例如CUDA。

Q：如何看待量子计算？

A（Hennessy）：量子计算就是计算机领域的核聚变（观众大笑，“核聚变”的比方是指潜力无穷但是不知道哪一天真的能用上）。目前，量子计算的问题在于规模化，一方面需要制备更多的量子位，另一方面需要在质因数分解之类问题以外找到更多应用。

Q：如何看到FPGA？

A（Hennessy）：FPGA是一个很好的技术。对于对成本不敏感的云端应用，因为FPGA芯片已经规模足够大可以装下一些处理器，因此得到了很多关注。微软在云端大规模部署FPGA，显然是在赌FPGA的可配置性在未来会有很多应用。Google则是把赌注押在了ASIC上（指TPU）。在客户端，FPGA由于功耗过大，目前仍然很难得到大规模应用。

Q：对于学生们有没有关于职业规划的建议？

A（Patterson）：在座的学生都很优秀，我想大家除了学习以外还要关注一些其他方面的能力，比如如何影响他人，如何写作表达自己，如何与家人朋友一起享受生活等等。这些都非常重要。如果你一定要问我技术有关的建议，我要说的是，你在这里（斯坦福）一定要上机器学习的课程。我感觉现在的人工智能就像之前的微处理器和互联网一样，一定会深刻地改变我们的社会。