百度基础架构技术发展之路

欧阳剑
百度公司

关键词：ASPLOS2014　最佳论文提名

在2014年3月召开的第19届国际体系结构对编程语言和操作系统的支持会议(International Conference on Architecture Support for Programming Languages and Operating Systems, ASPLOS)上，共收录了两篇来自中国大陆第一作者的论文，一篇是中国科学院计算技术研究所陈天石、陈云霁的论文，另一篇是百度公司的论文。更令人振奋的是，陈氏兄弟的论文被评为最佳论文，百度公司的论文获得了最佳论文提名。中国互联网公司在国际计算机系统及体系结构顶级会议上发表论文，这在国内工业界是第一次[1]。

很多同行比较好奇，想了解百度发表这篇论文的情况以及背后的故事，包括百度的研究开发体系以及相关的基础架构技术。在此感谢*CCCF*编辑部的邀请，让我们有机会和大家分享百度的混合研究之路及百度基础架构技术的现状和未来。

从ASPLOS 2014百度SDF论文说起

在ASPLOS 2014上，百度以“SDF:Software-Defined Flash for Web-Scale Internet Storage System”为题发表的论文成为此次大会录用的49篇论文之一。国内互联网企业第一次在国际顶尖水平的计算机系统和体系结构会议上发表论文，代表着国际同行对百度公司的认可，也是对我们工作的极大鼓舞。

软件定义闪存(Software-Defined Flash, SDF)，最早是由林仕鼎[2]在2011年初提出来的。当时，从产业的角度，云和端的发展趋势已经非常明显；从技术角度，数据中心以后会承载用户绝大部分的计算和存储，而传统的数据中心体系结构仍然沿用PC的体系结构，无法满足大规模系统对性能、成本、功耗以及可扩展性的要求。当时百度正在做新一代的存储系统，考虑到传统的固态硬盘(solid state disk, SSD)在性能和成本方面的诸多缺陷，如带宽利用率低、空间利用率低及性能的不可预测等，需要面向数据计算中心重新设计SSD。于是，我们开始研制SDF。SDF是一个软硬件协同系统，完全颠覆了SSD的性能。

SDF有如下几个特点：

- 底层Flash通道用户态的软件是可见的，让软件来管理数据的布局(layout)，使得硬件的并行性能得到充分发挥。

[1] 截止2013年年底，ASPLOS的Citeseer影响因子高居CCF认定的所有计算机系统与高性能计算领域会议的榜首。在CCF指定的“计算机系统与高性能计算领域”的五大A类会议中，大陆科研机构作为第一作者在ASPLOS上发表的论文只有三篇。2012年，中国科学院计算技术研究所的论文“关于数据中心上的迭代编译优化”是第一篇。2014年，有两篇入选ASPLOS，说明大陆在系统结构研究水平上有了很大提高。陈天石、陈云霁的论文介绍详见本刊2014年第5期。

[2] 百度公司前首席架构师。

- 基于层次到竖井的设计理念，实现了扁平的新文件系统和IO stack，提高了可扩展性并降低了延时。
- 与存储系统相结合，读写块的大小尽量与硬件友好。
- 资源全局利用，取消硬件通道间的异或校验，存储系统的三副本本身能保证数据的可靠性。

经过两年的努力，SDF研制成功了。在实际应用的系统上，SDF的性能达到传统商用SSD的三倍（硬件配置相同），而成本却大大降低，每GB可降低50%。

不抱任何希望，因为在大家的观念中，SSD是非常复杂的软硬件系统，很多专业做SSD的公司都有几十人甚至上百人的研发团队。而且，SDF的概念非常新，在2011年初，甚至还有点科幻色彩，很多人都觉得不可能实现。不过后来事实证明，正是SDF非常优雅直观的设计理念和架构，保证了其实现上的简单。在只有两个人的情况下，花了大概半年多的时间，就完成了SDF初始版本，总共写了不到1万行的RTL verilog代码和3000行的C语言代码，而在这期间，这两个人还兼做了很多其他项目，包括后续将要提到的ARM服务器。

使用。随后陆续上线。第一批20台，第二批100台，第三批500台，每次上线都要运行很长一段时间后下一批才会再上线。前三批稳定运行半年多，一直都非常顺利。可是真的没有问题了吗？我们有些庆幸、甚至有些怀疑，当我们以为这个项目将要顺利完成的时候，问题终于在第四批1000多台的上线中暴露出来。因为在进行硬件设计时的经验不足，对现场可编程门阵列(FPGA)的输入/输出(I/O)没有做足够的约束，导致在数据量大的时候会出现数据不可靠的问题，直接影响了线上的使用。最初我们以为是硬件逻辑问题，一直没办法

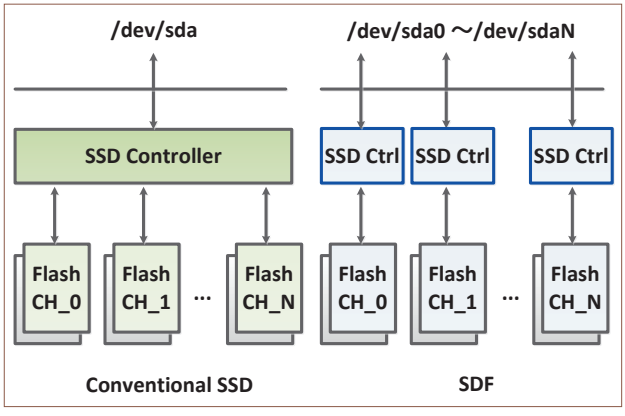

图1 SDF与传统SSD的架构区别

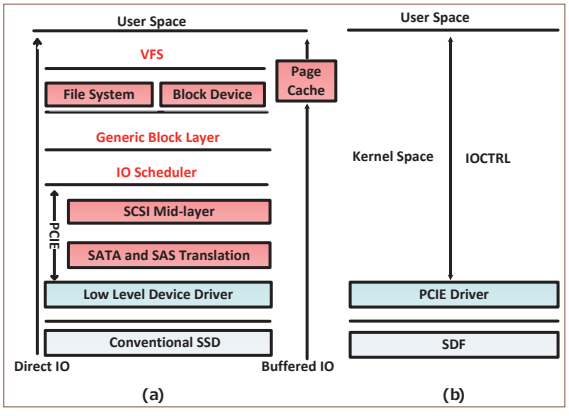

图2 SDF(b)和传统SSD(a)的IO stack区别

SDF的研发并不是一帆风顺的。实际上，当林仕鼎提出SDF的想法并打算由百度自己设计一款SDF时，百度只有一个硬件工程师和一个软件工程师负责此项目。最要命的是，这个软件工程师当时还只是个实习生，而硬件工程师从学校毕业后工作也还不到两年。当时很多人对此项目

如果是在学术界，SDF完成初始版本，可能就到此为止了。但工业界的目标是要落地、产品化，系统必须足够稳定，应用程序接口(API)足够好用。在之后的半年多时间里，我们又发布了第二个版本。经过长时间、充分的测试，第二版本还开发了轻量级的用户态文件系统以方便用户

定位，那段时间我们承受着巨大的压力。如果放弃，不仅仅是一年多的努力前功尽弃，还会给公司带来巨大的经济损失，例如硬件采购成本。经过两三个月的艰难调试，在无数次尝试之后，发现有可能是I/O约束问题，在修改约束之后，数据的可靠性大大提高，使得第四批产品也顺利上

线，并且性能数据非常好。后来我们反思这个问题时发现，最初设计时，出于稳定性的考虑，只上了一个性能仅仅够用的版本，而这个版本的性能只有最终论文上描述的 60% 左右。在 SDF 研发的过程中，我们真切地体会到，做工程和做研究其实是两条并行的道路，并没有太多的交集。一个感性的想法离发表论文可能不太远，但离实际的规模应用还有十万八千里。两者对做事情的方法要求也很不同，一个公司如果要想把这两者结合在一起，并且出成果，除了需要工程师的不懈努力外，还需要一套可行的研究及工程研发体系作保障。

SDF 做到后期，我们考虑写篇文章把我们的想法和工作发表出去。起初计划投第 11 届 USENIX 文件与存储技术会议 (The 11th USENIX Conference on File and Storage Technologies, FAST2013)，当时文章其实已经写完了，但最后一刻我们又放弃了，因为我们希望尝试一下向 ASPLOS 投稿。这又是一个工业界和学术界不同的地方。学术界追求更多的产出，所以一般先把子系统、子想法发表在一些专业性更强的会议上，然后再把整体成果发表在系统的会议上。但在工业界，发表论文是个人行为，工程师本职工作已经非常繁重，基本没有时间和精力来准备论文，所以我们想一步到位，直奔 ASPLOS 2014。

由于准备经验不足，而且投稿截止时间刚好和 SDF 上线撞车，完全是抱着试试看评委们会给什么反馈的心态而仓促上阵。没想到最后得到的反馈出乎意料的好，六位评委对 SDF 的工作都非常肯定，一个认为排名前 5%，三个认为前 25%，两个认为前 50%，主要意见都是书写和实验部分，而此时，距离 ASPLOS 截稿时间只有两周了。虽然时间紧迫，但 ASPLOS 的小试牛刀还是极大地振奋了团队，我们重写论文并补充了更多的新实验数据。这一次，我们终于得到了挑剔的 ASPLOS 评委们的青睐。

百度为什么要发表论文

和谷歌、脸谱等公司一样，百度这样的工业界公司，发论文的主要目的是建立技术品牌，扩大技术影响力，从而吸引更多优秀人才加盟百度。另一个原因是通过公开自己的技术，回馈社会。

一个公司建立自己的技术品牌，有很多方式，例如做出最酷的产品，也可以开放自己产品或者系统的源代码，或者在顶级会议上发表论文。谷歌、脸谱在这三方面做得都很不错，所以一直是美国大学生毕业求职最向往的公司之一。而国内的互联网公司，做了很多产品，也开源了不少自己的系统，但是鲜有在国际顶级学术会议，尤其是计算机系统和体系结构会议上发表文章。计算机系统和体系结构的研究对平台的依赖大，需要的资源多，一般的研究单位难以获得这样的资源和平台。而互联网公司拥有大数据和大系统，具有做系统技术研究得天独厚的优势。以百度为例，有超过 1000PB 的数据、单个分布式计算集群过万台的服务器。互联网公司的另一个优势就是有很多真实的问题、挑战和需求，这些都是与互联网用户直接相关的，基于这些问题、挑战和需求来做研究，成果也会直接反馈到用户体验上，更容易引起大家的兴趣，也更容易让人理解。很多传统的计算机系统及体系结构研究，十几年来一直基于经典的基准 (benchmark) 来做实验，或者一直基于基准中体现出来的问题来做研究，用互联网领域的流行语来讲就是不接地气。以百度的 SDF 和 ARM 服务器为例，SDF 面临的挑战是传统的 SSD 成本高，性能和容量利用率太低。而 SDF 的成果直接应用在百度的网页库上，能间接影响搜索的用户体验。百度的 ARM 服务器面临的挑战是云存储的成本太高，导致不能向用户免费赠送太大的存储空间。我们用 ARM 服务器降低了存储成本，使得百度网盘在有 1 亿多用户的情况下，仍然敢于向每个用户免费赠送 4TB 的存储空间。这些研究成果是直接惠及每一位用户的。正是由于互联网公司具有这样的优势，所以最近十年来，每年顶级的计算机系统或者体系结构会议都会有来自互联网公司的文章，而且数量还

不少。百度也希望通过论文的方式，把一些只有在大型互联网公司才能开展的研究和实践方法公开出去。

百度的混合研究发展之路

谷歌在《美国计算机学会通讯》(*CACM*) 上发表过一篇文章“谷歌的混合研究方法”[3]。该文章的核心思想是把工程和研究紧密地结合在一块，文章的核心语这样说到：“将研究与开发紧密结合，使得谷歌有能力以一种前所未有的规模进行实验，这常常为公司带来新的能力。”文章提到了五种具体的实践方法。这五种方法的核心是最终研究目标的产品化。百度的做法与之类似，其实百度的基础架构部门到目前为止还没有纯粹的研究人员，工作人员的研究工作都是穿插在日常的工程项目中。这些研究工作的唯一目标就是用更大胆、更超前的想法来把系统或者产品做得更好，这些研究最终都是要应用到实际的系统或者产品中。

百度的研究方法最大的特点就是敢为天下先。无论是 SDF、ARM，还是我们现在正在做的一些工作，都是世界范围内的先例。我们做 SDF 的时候，需要重新设计软件和硬件，硬件的架构和软硬件之间的接口都与传统的 SSD 不同，而且我们的用户又是非常关键的“网页库”，可以有很多理由让这个项目不能落地，例如“新产品不稳定”，“新的架构没经过验证”等等。而且刚开始的时候，SDF 确实出现了不稳定的情况。如果一开始我们就被困难吓倒，或者找借口不愿意承担风险，大家就不会看到今天 SDF 的成果。ARM 项目也一样，2011 年初百度开始做 ARM 服务器的时候，遇到了很多困难，例如 CPU 是 32 位的、大量代码要移植、产业链不成熟、没有参考经验等等。而且业务方是百度云，承载了一亿多的用户，绝对不能出问题。三年后，百度的 ARM 服务器已经大规模应用。前不久笔者与脸谱公司数据中心的工程师聊天时了解到，他们在 2011 年的时候也评估过 ARM 服务器，但发现 CPU 是 32 位的，需要移植大量的代码，最终就放弃了。事实上，从 x86-64 到 ARM-32 的移植并没有想象中的那么复杂。百度只靠一个工程师，花了不到半年时间就完成了一百多万行代码的基础库和存储系统的移植和测试。

百度的基础结构技术

百度的基础结构技术包括软件基础结构和硬件基础结构，涉及的领域包括计算机系统、体系结构、硬件等。百度有三个自建的大规模数据中心，年均 PUE（Power Usage Effectiveness，电源使用效率）是 1.32，最佳 PUE 达到 1.16，这些都是国内领先，世界一流的水平。在软件基础结构领域，百度实现了很多自己的系统，例如单集群过万台服务器的 MapReduce 集群、单集群过万台多服务器的 MPI（Message Passing Interface，消息传递接口）高性能计算集群等。百度的 MapReduce 系统和 MPI 系统也是在开源的方案上做了深度优化，所以当产业界只有最大 5000 台服务器的 MapReduce 集群的时候，百度的 MapReduce 集群已经过万。另外，百度还有自己开发的流式计算系统，能支持毫秒级别的实时流式计算。百度的存储系统也是完全自主开发的，我们内部称之为“新存储体系”，这个存储系统有一个统一的块设备层 (block layer) 支持表 (table)、文件 (file) 和键 - 值 (key-value) 的不同接口实现，而且该系统一开始就充分考虑了闪存的特点，支持 SDF 的接口，能充分提高硬件的效率。

百度基础技术有个特点，就是“全栈式（full stack，从硬件到软件）自主开发”，这和国内其他公司喜欢完全使用开源方案形成鲜明的对比。我们这样设计主要出于几个考虑：首先，百度业务发展很快，基础架构必须能跟上并推动业务发展，开源社区方案往往没办法做到这么快的迭

[3] 此文发表在2012年第7期*CACM*上。本刊于同年第9期刊发了这篇文章的译文。

代，例如百度的 MapReduce 可扩展性优化比开源方案早了两年；其次，如果只在开源方案上修修补补，时间长了，容易丧失做大系统的能力，毕竟维护一个开源系统并做故障修复 (bug fix) 所需要的工程团队和工程能力，和从头开发一个大系统所要求的能力是不同的；第三，全栈式自主研发能真正做到软硬件协同，从而真正达到系统最优，而开源方案一般只关注某一个领域或系统里面的某一个层次，很难做到系统最优。最明显的是，百度有基于 ARM 和 SDF 的存储系统，能做到单位存储成本最低或者单位吞吐成本最优。

本文介绍了百度的基础架构技术及其混合研究方法。百度的混合研究方法能保证研究不会偏离产品方向，也让研究更接地气。百度的大数据、大系统平台有能力让更多的创新想法得以实践和验证。虽然到目前为止，在计算机系统和体系结构方面开展的前沿研究仍然不多，但只要坚持这样的混合研发方法，随着百度的发展，研究工作就能很快接近或者达到世界一流水平。

以技术为核心竞争力的百度，基础结构技术除了服务于内部业务，将很快以开放云的方式向社会开放，支持更多的行业用户使用，让更多的计算机科技工作者借力百度的大数据大系统平台，助力我国计算系统实力的全面提升。■

欧阳剑
百度基础架构部高级架构师。主要研究方向为面向大数据的计算、存储、网络体系结构等。ouyangjian@baibu.com

CCF@U：CCF走进高校

序号	演讲人	时间	高校	演讲题目
201	周傲英 周晓方	6月4日	中国矿业大学	数据系统评测基准：回顾与展望 从数据库到大数据：应用、技术和理念
202	何万青	6月4日	闽南师范大学	确定和培养核心竞争力，加强结果导向的学习
203	何万青	6月5日	厦门理工学院	确定和培养核心竞争力，加强结果导向的学习
204	何万青	6月5日	集美大学	确定和培养核心竞争力，加强结果导向的学习
205	何万青	6月6日	厦门大学	确定和培养核心竞争力，加强结果导向的学习
206	孟小峰 高　军	6月6日	贵州大学	大数据管理：问题与思考 大图数据查询处理关键技术研究的一些进展
207	陈熙霖	6月23日	泉州师范学院	**演讲题目：**视频人脸识别
208	陈熙霖	6月23日	厦门大学	**演讲题目：**视频人脸识别
209	陈熙霖	6月24日	华侨大学	**演讲题目：**视频人脸识别

图中第二排右起第三位为CCF副秘书长陈熙霖