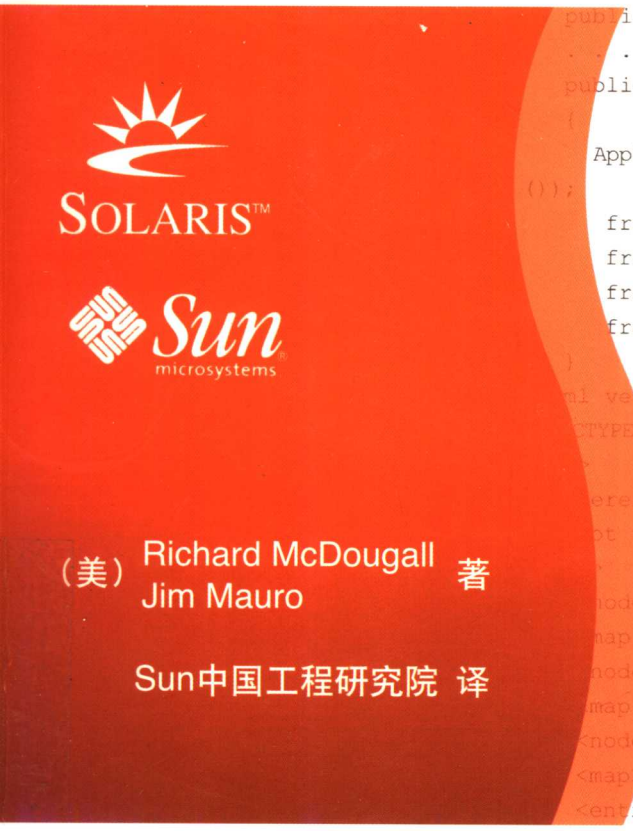
HZ BOOKS
PRENTICE HALL

Sun公司核心技术丛书

Solaris 内核结构 第2版

Solaris Internals Solaris 10 and OpenSolaris Kernel Architecture (Second Edition)



SOLARIS™

Sun microsystems

(美) Richard McDougall
Jim Mauro 著

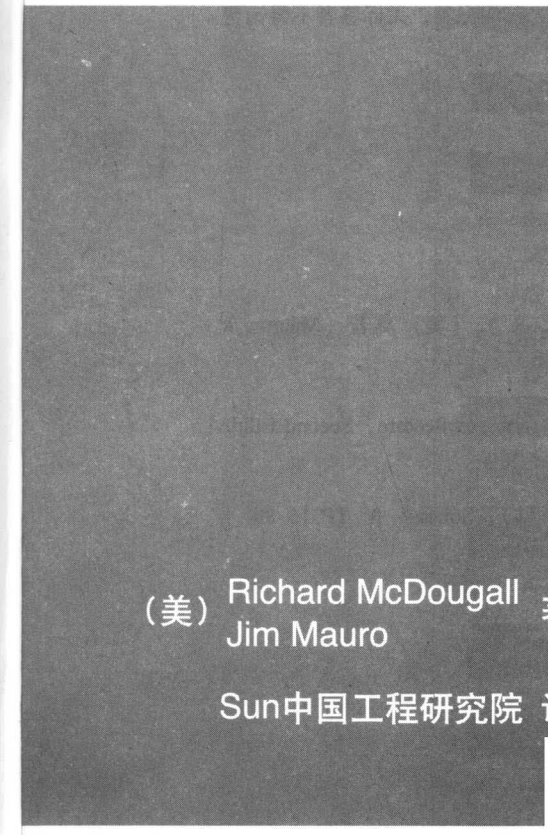
Sun中国工程研究院 译

机械工业出版社
China Machine Press

Sun公司核心技术丛书

Solaris内核结构 第2版

Solaris Internals Solaris 10 and OpenSolaris Kernel Architecture (Second Edition)



(美) Richard McDougall 著
Jim Mauro
Sun中国工程研究院 译

机械工业出版社
China Machine Press

本书与其配套出版物《Solaris 性能与工具》（该书已由机械工业出版社同步出版）共同提供了 Solaris 及 OpenSolaris 操作环境的最优秀、最全面的介绍。《Solaris 内核结构》深入探索了 Solaris 操作系统的内部原理和体系结构；《Solaris 性能与工具》阐释了大量实用工具的使用，为内核开发人员、系统程序员和系统管理员深入理解系统的行为及性能提供了系统化方法。

本书描述了 Solaris 10 和 OpenSolaris 内核中所有主要子系统的算法和数据结构，对第 1 版进行了大幅修订，加入了很多新的内容。集成的 Solaris 工具和实用程序贯穿全书，目的是让读者细致观察到 Solaris 内核的工作过程，深入理解、分析系统的性能和行为，包括内存、进程、线程、文件系统、网络 TCP/IP 实现、资源管理工具，等等。

本书适合使用 Solaris 操作系统的各类技术人员阅读。

本书版权登记号：图字：01-2006-5285

图书在版编目（CIP）数据

Solaris 内核结构：第 2 版 /（美）麦克道格（McDougall, R.），（美）莫若（Mauro, J.）著；Sun 中国工程研究院译. -北京：机械工业出版社，2007.6

（Sun 公司核心技术丛书）

书名原文：Solaris Internals: Solaris 10 and OpenSolaris Kernel Architecture, Second Edition

ISBN 978-7-111-21485-4

Ⅰ. S…　Ⅱ. ①麦…　②莫…　③S…　Ⅲ. 操作系统（软件），Solaris　Ⅳ. TP316.89

中国版本图书馆 CIP 数据核字（2007）第 068750 号

机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码　100037）

责任编辑：刘立卿

北京牛山世兴印刷厂印刷 · 新华书店北京发行所发行

2007 年 6 月第 1 版第 1 次印刷

186mm × 240mm · 38.5 印张

定价：75.00 元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线（010）68326294

中文版序

2000年5月，在Mark Himelstein橄榄枝的召唤下，我重回Sun，任Solaris内核研发总监。我发现这里可真谓人才济济，高手如云。一个小伙子，Andy Tucker，居然是高级工程师。问他做什么项目，回答很简单：虚拟化。这家伙！当我没念过书吗？计算机中90%的学问就是虚拟化。但几个月后，我明白了，他的确是在做计算机中90%的学问。Andy不到20岁就从斯坦福大学毕业了，直攻博士后，进入Sun继续研发Solaris内核技术。他先加强scheduler，让Solaris的扩展性从一个到上千个用户，可以线性扩展。之后，在Tim Marsland的鼓励下，Andy以虚拟技术为基础，继续提高Solaris的扩展性。四年后，Andy带出了Container，使Solaris的扩展性达到世界第一的水平。

我和Tim Marsland做了五年的邻居。在有名的MPK17大楼的三楼，我的办公室位于一个角落，他就在隔壁。Tim是Solaris内核的灵魂。他的办公室总是门庭若市，每天都有工程师、经理、总监来找他，还有副总裁等人找他。做他的邻居，沾光不少，也因此认识了Sun公司的各方豪杰，其中不少就在MPK17的三楼。Michael Shaparo和Bryant Cantril是多年的好友，他们在布朗大学就是一对搭档。Bryant毕业后到Sun，Michael则留下攻读硕士。一年后，Bryant跟着Michael进了同一家公司，也一起平步青云。在没有经理人的指导下，他们写出了aTrace——一个用来追踪内核中内存地址的软件工具。我问Bryant下个项目要做什么？“DTrace”他想都没想就说了。这个工具可以迅速诊断全系统的运行，也能细查小单位的优化状态，是公司运行的一大利器。

MPK17的同事们有几个好喝两杯。一天我走过Stephen Hahn的办公室，一股酒香扑鼻而来。里面的几个人或站或坐，神秘兮兮的，一见我，都不说话了。过了两天，Stephen Hahn自己到我的办公室来了，要我准他做他的“星期五项目”。至于“星期五项目”是什么，照例不必问了。这有点奇怪了。原来，他不是要我批准，而是要我保密。他也不知道我那天听到了多少，反正不论如何，我不得泄露风声。这个项目，就是Greenline。后来改名为Solaris Management Framework（SMF）。

Cindy McQuire不在那群酒友当中。她是当时少有的女性资深工程师，个性平和，做事周密，说话中肯。我有意请她做经理，她却一口回绝了。第一，她的故障管理当时做了一半，后一半不做不行；第二，软件工程比软件管理有趣多了；第三，她有两个孪生女儿，她要让她们知道，女性软件工程师不比男性差。我哑口无言，她也就去带FMA队了。Cindy善于踢男女混合室内足球，所以经常带伤上班。

另一个不在那群酒友中，但酒量却不错的是Paul Sangster。我在做内核时，需要知道一些安全方面的知识，Mark Himelstein要我和他聊聊，这一聊就是三个小时。最难得的是他在三小时内清楚地解释了什么是Single Sign-on。我后来就任网路及安全总监和Paul有很大的关系。要Single Sign-on，必须有个完整的加密系统。Solaris Encryption Framework就是这样做出来的。这套软件让客户用自己的算法，不用国家检验，不受出口管制，也不需美国知道算法。安全界的Glenn Faden和Gary Winiger，埋在Trusted Solaris里十几年。他们耐心教我其中复杂的观念，坚持Trusted 10的架构。加上Scott Rotondo，几个人把Solaris 10做成了世界第一安全的操作系统。Scott想出了一个办法，可以安全且高速地给每个文档签名。为了保险起见，他向Whit Diffie请教后才开始使用。他常说要发表，取名为Diffie-Rotondo Algorithm（外行人请搜索Diffie-Hellman Algorithm）。

当大家知道我也能喝两杯后，Sunay Tripathi带着一瓶Scotch来到我的办公室。当时是晚上六点多，他带着他的经理、三个杯子、一瓶酒出现在我门前。一看来者不善。好在我还能抵挡，不过也给他骗去了几小时，定下他CTW的蓝图。几周前，我要Sunay想想如何能改变全世界（Change the World）。他要带烈酒来配合这个大题目（通常别人只带些红酒来，诉诉苦，解解闷）。Sunay接着“Brutus”（Bruce Curtis）后，写出了

FireEngine，让 Solaris 的网络功能增强三倍以上，足以和任何操作系统抗衡。

周曙东是 IO（Input/Output）领域的突出人物，也是华人工程师的明灯。他一头钢丝乱发，身材飘逸，走在长廊里，像个大学教授。在 Solaris 10 后期，曙东临危受命，接下新的启动程序项目，重写了 Solaris 十数年没动过的启动系统。启动时间从数十分钟减到十数秒，带上漂亮的图面，大幅度提升了可用性。

Jeff Bonwick 是另一位领袖。他从 CalTech（加州技术学院）毕业，两三年时间就成为 Sun 软件工程界的著名英雄。早年他和 Roger Faulkner 合写了一个软件工具，可以极快地安装软件包。他们说是英雄就要用快刀。这“快刀”叫 bfu。他们说是 Blindingly Fast Utility，可是每个人都说是 Bonwick Faulkner Utility。Jeff 带了个团队做新一代的文档系统（ZFS），如果说宇宙中每一个原子都是一个文档，这个系统都能容纳。当然，当真如此，也没原子给我们人类了！

如果说 Jeff 是把快刀，Bart Smaalders 就是门重炮。他出身机械工程，出道时在国防工业做事，制造战车。他一头乱发，一脸大胡子，雨天时戴个宽沿皮帽，形象和软件工程师实在不符。Bart 是 Solaris 的哲学家。他坚持软件开发的程序和精神——把问题定好、数据收齐、提出方案、试用、审议、再收集数据。问题解决了，就行了，不必再啰唆。这一哲学，是 Solaris 队伍共同的信念。他的 libmicro 项目，就是这一哲学的具体实现。Cindy McQuire 坐 Bart 隔壁，她收集了些 Bart 的名言贴在门口，称之“Bart 语录”（Bartism）。

Solaris 10 经历了四位副总裁，历时四载，参与的工程师达 1500 人以上，是 Sun 有史以来最大的系统项目。2002 年，John Loiacono 决定重新进入 x86，并在北京组建百人驱动程序队伍。我在美国聘到何英，在中国和宫力合作，三个月雇了近 40 人。之后 Bev Crair 到任，一年后扩大到 90 人。2005 年 Solaris 与 JDS 队合并。北京的 Solaris 队伍已达 180 人，北京成为在美国 Menlo Park 以外最大的地区。

我有幸主持了 Solaris PAC（Product Approval Committee），在 2005 年 1 月底推出产品。一年内新用户超过 300 万，打破所有 Linux 版本的纪录。产品开发中的主要人物可比梁山泊 108 条好汉，个个本事高强、个性鲜明。

大中国区，是世界仅存的 IT 战场。从操作系统的角度，只有开放和封闭两个选择。封闭的路，会让一家公司控制整个市场。自主创新必须走开放的路。本书读者选择的就是这个战场。我相信，中国区中的读者，一个都不会选错。在开放的这边，Linux 和 Solaris 是同源亲家。Linux 这个小老弟，近年来在这方面跑得快些。可是 Solaris 路走多了，不需回头走冤枉路。读者可以两者兼得。站在巨人的肩上，迈的步又快又大。人聪明、志向大、底子厚、工具快，创新立业，有何困难？

Sun 中国工程研究院院长，王星耀

2007 年春

原　序

过去十年来，人们对操作系统普遍有一种误解，认为它已经是一个已完成、已解决的问题，虽然操作系统仍能引起人们的兴趣。产生这种想法的原因是多方面的，但其中最大的因素可能只是因为操作系统没有被理解；在很大程度上它们不是作为透明系统，而是作为专有的黑盒子发布的，甚至无法满足人们简单的好奇心。这与理解操作系统是格格不入的；如果某项事物不能被分解，即它的内部工作机制是隐藏的，那么它的复杂结构和工程上的细微差别将很难被理解。对于软件系统尤其如此，它们甚至不能在传统的意义上被分解。尽管软件的象征意义是信息而不是机器，但一个封闭的软件系统就像一个工程化的系统一样不易被理解。

这就是 Solaris 大约在 2000 年时的状况，它确实没有被很好地理解。它的内部结构和工作机制仅仅在令人费解的注释或者旧的 USENIX 论文中公开描述过，它的行为对于当时的工具是不透明的，源代码深藏于“密室”之中。从 2000 年开始，情况开始（可能是慢慢地）好转——这部分归功于 Jim Mauro 和 Richard McDougall 的《Solaris 内核结构第 1 版》所开创的先河。Jim 和 Richard 面临极大的挑战——描述一个如此复杂的系统，而事实上却没有一个人真正理解它的全部。在写作过程中，Jim 和 Richard 认识到一本书根本不能完全涵盖它。尽管内容一再压缩（例如不包括网络部分），《Solaris 内核结构第 1 版》仍然厚达 600 页以上。

《Solaris 内核结构》的出版标志着过去十年的前五年是加速变化的开始，使用和理解 Solaris 的障碍已经被清除。Solaris 成为自由软件，它的工程师们开始通过新的媒体（如博客）广泛谈论它的实现，最重要的是，Solaris 本身在 2005 年 6 月成为开源软件，成为第一个从专有转变为开放的操作系统。同时，Solaris 的机制变得更加有趣，几项革命性的新技术在 Solaris 10 中首次亮相。这些技术动摇了很多人的怀疑，并证明了操作系统仍然保持活力。当然，仍有一些困难的、重要的问题有待解决。

如果将 2000 年看作是 Solaris 变化的开始，2005 年就可以看作是这一开始的终结。到 2005 年末，曾经看起来已完成的专有产品已转化为一个令人激动的、开源的系统，系统的潜能使其更具有活力。恰巧《Solaris 内核结构第 2 版》[⊖] 迎来了这些变化。面对一项异常艰巨的任务，即如何反映五年来的大规模的工程变化，Jim 和 Richard 做出一个重要决定——他们邀请那些设计子系统和编写代码的工程师给予帮助，其中的一些章节完全是由这些工程师们在自己所设计的子系统的基础上编写的。得益于此，《Solaris 内核结构第 2 版》成为一部得到极大扩展且具有高度权威性的专著——同时很好地保持了社区开发和写作这一新的 Solaris 时代精神。

就我个人而言，我很高兴看到 Jim 和 Richard 使用了 DTrace，Mike Shapiro、Adam Leventhal 和我在 Solaris 10 中开发了这项技术。Mike、Adam 和我都曾是操作系统课的助教，我们有过一个非公开的目标，即开发一个教学工具以彻底改变操作系统的教学方式。因此，我鼓励读者不仅仅是要读《Solaris 内核结构》，还要下载 Solaris，在台式机、笔记本或虚拟机上运行它，并使用 DTrace 在自己的机器上亲自看看 Jim 和 Richard 描述的概念！

不管你是学生还是专业技术人员，是因为课程、工作还是因为好奇来读此书，我都很高兴大家能阅读这本 Solaris 内部机制的指南。享受这一过程，并记住 Solaris 并不是一件已终止的工作，而是一项不断发展的技术。如果读者有兴趣加速这一发展，或者仅仅是对于使用或理解 Solaris 有问题，请加入我们在 http://www.opensolaris.org 的很多社区。欢迎您的到来！

Bryan Cantrill
旧金山，加州
2006 年 6 月

⊖ 该书即将由机械工业出版社出版。——编者注

前　言

本书是《Solaris 内核结构第 2 版》的配套出版物，欢迎大家阅读这两本书。《Solaris 内核结构第 1 版》出版已近五年，在此期间，我们有机会与很多 Solaris 用户、软件开发者、系统管理员、数据库管理员、性能分析师，甚至偶尔的内核黑客进行沟通。我们对所有的反馈表示感谢，而且，基于读者的意见，我们对这一版的格式和内容专门做了修改。读下去就会知道两本书有哪些不同。我们期待与 Solaris 社区继续交流。

关于这两本书

这两本书讨论的是 Sun 的 Solaris 操作系统——特别是 SunOS 内核。Solaris 的其他组成部分，如桌面的窗口系统，不在本书讨论范围内。《Solaris 内核结构第 1 版》涵盖了 Solaris 2.5.1、2.6 和 Solaris 7。现在这两本书重点介绍 Solaris 10，包括 Solaris 8 和 Solaris 9 的更新信息。

在《Solaris 内核结构第 1 版》中，我们不仅想要描述使 Solaris 内核运转的内部组成单元，而且还提供实用的指导。该书的第 2 版也同样如此，并更加强调使用捆绑的（在某些情况下是非捆绑的）工具和实用程序，以用于检查和探测一个运行中的系统。我们能够使用观察工具说明更多的内核内部工作，在很大程度上得力于加入到 Solaris 10 中的革命性和创新性的技术——DTrace 这一动态的内核跟踪框架。DTrace 是 Solaris 10 中的多项新技术之一，在这两本书中有大量应用。

在《Solaris 内核结构第 2 版》写作过程中，我们得到几位朋友和同事的帮助，他们大都从事 Solaris 内核工作，他们的专业技术和指导为这两本书的质量和内容做出了巨大贡献。我们自己也不断地扩展主题，演示 `dtrace (1)`、`mdb (1)`、`kstat (1)` 以及其他捆绑工具的使用。因此，我们很早就决定要介绍这些工具，一些章节为读者提供了有关这些工具和实用程序所需要的背景信息。自此，使用工具对性能和系统行为进行分析发展成为一整章。

本书临近结稿时，我们遇到了一个小问题——书的厚度。书太厚了，这给书的出版和印刷带来了一些问题。与出版商讨论后，我们决定将书分为两册。

《Solaris 内核结构第 2 版》。这是对第 1 版的更新，包含大量新材料。包括所有主要的内核子系统：虚拟内存系统（VM）、进程与线程、内核调度程序与调度类、文件系统与虚拟文件系统（VFS）框架，以及核心内核工具。还包括新的 Solaris 资源管理工具，关于网络的新的一章。Solaris 8 和 Solaris 9 中的新特性安排在正文中恰当的地方。

《Solaris 性能与工具》中描述的用于性能和分析工作的实用程序和工具的例子也会在《Solaris 内核结构》中使用。

《Solaris 性能与工具》。描述了 Solaris 10 中捆绑的实用程序和工具：`dtrace (1)`、`mdb (1)`、`kstat (1)` 等。有些章节详细描述了如何使用这些工具分析 Solaris 系统性能和行为。

这两本书可以搭配使用，并可与位于 http://www.opensolaris.org 的 Solaris 源代码配合使用。对 Solaris 8 之前的某个版本感兴趣的读者应该继续使用《Solaris 内核结构第 1 版》作为参考。

面向的读者群

我们相信这两本书将为工作在 Solaris 操作系统上的各类技术人员提供有用的参考。

应用程序开发者能够在这两本书中找到应用编程界面之后的 Solaris 操作系统如何实现函数的信息。这些信息帮助开发者在开发 Solaris 应用程序时，理解每个界面的性能、可扩展性和实现细节。系统概览和关于调度、进程间通信、文件系统等对这类读者来说是最有用的章节。

设备驱动和内核模块开发者（负责开发驱动程序、STREAMS 模块、可装入系统调用，等等），能够在这里找到 Solaris 操作系统的总体体系结构和实现理论。这两本书的 Solaris 内核框架和实用程序部分（特别是锁和同步原语涉及的章）尤其有用。

系统管理员、系统分析师、数据库管理员和企业资源规划（ERP）经理（负责性能调优和负载规划），能够学到主要的 Solaris 子系统的行为特征。文件系统缓存和内存管理各章提供了大量 Solaris 在实际环境中行为的信息。Solaris 可调参数后面的算法在两本书中进行了深入讨论。

技术支持人员（负责诊断、调试和支持 Solaris）将发现大量关于 Solaris 实现细节的信息。每一章中提供的主要的数据结构和数据流程图可帮助调试和操作 Solaris 系统。

想知道更多关于 Solaris 内核工作的系统用户，将在每一章的开头找到高层次的概述。

除了技术用户社区之外，在学术界研究操作系统的人员将发现本书内容是很好的参考。Solaris 操作系统是一个健壮、功能丰富且大量发行的操作系统，适用于不同的工作负载，从单处理器台式机到具有大量内存和输入/输出配置的庞大多处理器系统。Solaris 操作系统为商业数据处理、Web 服务、网络服务和科学计算负载提供的健壮性和可扩展性在业界是首屈一指的。研究这一操作系统可以学到很多知识。

OpenSolaris

在 2005 年 6 月，Sun 公司推出了 OpenSolaris，由开放源代码构建的全功能的 Solaris 操作系统版本。作为 OpenSolaris 第一步的一部分，Solaris 源代码通过一个开放许可供公开使用。这对这两本书有几个明显的好处。我们可以在适当的时候将 Solaris 源代码直接包含在书中，同样可以指向全部的源代码清单。

通过 OpenSolaris（一个世界范围的开发者社区）可以访问 Solaris 源代码，开发者能够为他们感兴趣的操作系统的任何组成部分做出贡献。源代码的可访问性使我们能够组织这两本书的结构，交叉引用特定的源代码文件（可以具体到源代码树的行号）。

OpenSolaris 代表了技术专家世界的一个意义深远的里程碑；一个世界级、成熟、健壮且功能丰富的操作系统现在向所有希望使用 Solaris 的人敞开了大门，人们可以探讨它并为它的发展做出贡献。

访问 OpenSolaris 网站可以学到更多的关于 OpenSolaris 的知识：

http://www.opensolaris.org

OpenSolaris 源代码可以在 http://cvs.opensolaris.org/source 获得。这两本书中对源代码的引用是相对于上面的开始位置的。

书的组织

我们将《Solaris 内核结构》分为几个逻辑部分，将内容相关的章节组织在同一部分。我们的目标是提供一种积木式的方法，后面可以基于前面的内容深入。然而，对于熟悉操作系统设计和实现的特定读者而言，各个部分和章节可以根据需要单独使用。

《Solaris 内核结构》

第一部分　Solaris 内部结构介绍
第 1 章　介绍
第二部分　进程模型
第 2 章　Solaris 进程模型
第 3 章　调度类型和调配器
第 4 章　进程间通信
第 5 章　进程权限管理
第三部分　资源管理
第 6 章　zone

《Solaris 性能与工具》

第一部分　系统观察方法
第 1 章　系统观察工具简介
第 2 章　CPU
第 3 章　进程
第 4 章　磁盘行为与分析
第 5 章　文件系统
第 6 章　内存
第 7 章　网络
第 8 章　性能计数器

更新和相关材料

作为这两本书的补充，我们建立了一个网站，内容包括更新的材料、我们引用的工具和指向相关材料的链接。我们会定期更新网站（http://www.solarisinternals.com），及时反映这两本书和将来在 Solaris 内核方面的工作信息。网站会增加一个关于这两本书的 FAQ 论坛，以及关于 Solaris 内核结构、性能和行为的一般问题。如果在这两本书中发现了错误，我们会把勘误表放到网站上去。

作者的话

时间和精力上的巨大投入再一次为作者带来了好的回报。Sun 的 Solaris 内核开发组、Solaris 用户社区和第 1 版读者的支持使我们极其欣慰。我们相信，从为 Solaris 用户提供有价值的信息的角度说，我们在第 2 版中会更加成功。在协作过程中，我们确实丰富了我们的知识，而且，我们期待读者的反馈。

关于作者

如果 Richard McDougall 生活在 100 年以前，他会打开第一辆四冲程内燃机驱动的车辆的盖子，探索新技术以作出改进。他会寻找简单的办法解决复杂的问题，帮助创业者们理解技术是如何从他们的新经验中获得最大化收益的。现在，Richard 用技术来满足他的好奇心。他是 Sun 公司的杰出工程师，专注于操作系统技术和系统性能。

Jim Mauro 是 Sun 公司“性能、体系结构和应用工程组”的资深 Staff 工程师，他最近的工作集中在 Opteron 平台上的 Solaris 性能，特别是在文件系统和磁盘原始 IO 性能领域。Jim 的兴趣包括操作系统调度和线程支持、多线程应用程序、文件系统和操作系统观察工具。兴趣之外，他还热衷于读书和音乐——听唱盘是 Jim 的首选，并且他仍在购买和播放 12 英寸的乙烯基唱片。他与妻子和两个儿子住在新泽西。在写作和工作之余，Jim 会处理他的家人在使用家庭网络和打印机时遇到的麻烦。

致　谢

《Solaris 内核结构》的社区作者们

尽管这两本书的封面上只有三位作者，但实际上它们是社区工作的成果。我们的几个朋友超越了工作的要求，慷慨地拿出他们的时间、专业技术和精力，为本书作出了贡献。他们的付出极大地改进了书的内容，使这两本书能够覆盖更广阔的主题，也给了我们一个从特定主题专家听取意见的机会。我们真诚感谢下面这些人员。

Frank Batschulat 帮助更新了 UFS 一章。Frank 做了十年软件工程师，而且在 Sun 公司工作了七年。在 Sun，他是 Solaris 文件系统组的一员，主要专注于 UFS 和通用的 VFS/VNODE 层。

Russell Blaine 提供了 x86 系统调用信息。Russell Blaine 于 2000 年离开普林斯敦后直接加入 Sun 公司，此后一直在处理内核的不同部分。

Joe Bonasera 提供了 x64 HAT 描述。Joe 是 Solaris 内核组的工程师，主要致力于核心虚拟内存支持。Joe 的背景包括优化编译器和并行数据库引擎上的工作。他近期的工作围绕 AMD64 移植以及移植 OpenSolaris 运行在 Xen 虚拟化软件，特别是在虚拟和物理内存管理的领域及启动过程。

Jeff Bonwick 提供了 vmem 分配器的描述。Jeff 是 Solaris 内核开发的杰出工程师。他做出了大量贡献，包括最初的内核内存 slab 分配器，以及更新的内核 vmem 框架。Jeff 最近的工作是 Zeta 字节文件系统（ZFS）的体系结构、设计和实现。

Peter Boothby 提供了 kstat 的总体介绍。Peter Boothby 在 Sun 公司工作了 11 年，承担了不同的角色：系统工程师；澳大利亚和新西兰的 SAP 能力中心的经理；Sun 的性能工程师和在德国 SAP 的组经理；在苏格兰的支持欧洲 ISV 的 Solaris 和 Java 开发工作的 Staff 工程师。休假两年期间，他在法国滑雪，悉尼港赛游艇，在澳大利亚东海岸驾驶帆船，之后 Peter 返回了 Sun 的怀抱，成立了一家咨询公司，帮助 Sun 澳大利亚做大规模的合并和集成项目。

Rich Brown 为文件系统各章的文件系统界面部分提供素材。Rich Brown 在 Solaris 文件系统领域工作了十年。他现在的工作是寻找提高文件系统可观察性的方法。

Bryan Cantrill 提供了时钟子系统的总体介绍。Bryan 是 Solaris 内核工程的资深软件工程师。Bryan 的贡献包括时钟子系统、插入陷入表以收集陷入统计信息。最近，Bryan 开发了 Solaris 动态跟踪，也称 DTrace。

Jonathan Chew 在分配器 NUMA 和 CMT 各节提供了帮助。从 1995 年开始，Jonathan 就一直是 Sun 公司 Solaris 内核开发组的软件工程师。在那段时间，他的工作集中在统一的内存访问（NUMA）机器和芯片多线程。加入 Sun 之前，Jonathan 是斯坦福大学计算机系统实验室和卡内基梅隆大学计算机系的研究系统程序员。

Todd Clayton 提供了大页面体系结构变化的信息。Todd 是 Solaris 内核开发的工程师，他的工作（除其他工作之外）是虚拟内存代码和 AMD64 Solaris 移植。

Sankhyayan（Shawn）Debnath 与 Sarah、Frank、Karen 和 Dworkin 一起更新了 UFS 一章。Sankhyayan Debnath 是普度大学的学生，专业是计算机科学，曾经是 Sun 公司文件系统组的实习生。工作之余他喜欢在本地的赛道里赛车或是在镇里骑摩托。

Casper Dik 提供了产生进程权限一章的材料。Casper 是 Solaris 内核开发组的工程师，在安全和网络方面进行过广泛的工作。Casper 的许多贡献包括设计和实现 Solaris 10 的进程权限框架。

Andrei Dorofeev 提供了分配器一章的指导。Andrei 是 Sun 公司 Solaris 内核开发组的 Staff 工程师。他的兴趣包括多处理器调度、芯片多线程体系结构、资源管理和性能。Andrei 获得过俄罗斯 Novosibirsk 州立大学计

算机科学的荣誉硕士学位。

Roger Faulkner 提供了对进程一章的建议。Roger 是 Solaris 内核开发的资深 Staff 工程师。Roger 做过最初的 UNIX 系统 V 的进程文件系统实现，他做出了很多贡献，包括 Solaris 中过去和现在的线程实现和统一的进程模型。

Brendan Gregg 提供了重要的审阅贡献与对性能和调试一书的共同工作。Brendan 使用 Solaris 大约十年，曾经是程序员、系统管理员和咨询师。他是 OpenSolaris 的撰稿人，编写的软件有 DTrace 工具箱。他为 Sun 公司教授 Solaris 课程。

Phil Harman 提供了进程和线程模型描述的深入认识和建议。Phil 是 Solaris 内核开发的工程师，他的工作集中在 Solaris 内核性能上。Phil 的许多贡献包括测量系统调用性能的一般框架（称为 libMicro）。Phil 是线程和开发多线程应用程序的资深专家。

Jonathan Haslam 提供了 DTrace 一章。Jon 是 Sun 的性能组的工程师，而且是应用程序和系统性能方面的专家。Jon 是 DTrace 的非常早的使用者，为最终的实现作出了重要贡献，确定所需要的特性和改进。

Stephen Hahn 提供了用于项目、任务和资源控制各章的最早的材料。Stephen 是 Solaris 内核开发的工程师，在其他工作之外，为内核的调度代码和资源管理实现作出了重要贡献。

Sarah Jelinek 有 12 年的软件工程经验，其中 8 年是在 Sun 公司。在 Sun 期间，她的工作包括系统管理和文件系统管理，最近致力于 UFS 中的文件系统内核空间。Sarah 拥有计算机科学和应用数学的学士学位，计算机科学的硕士学位，都是在科罗拉多大学（科罗拉多温泉城）获得的。

Alexander Kolbasov 提供了任务队列的描述。Alexander 在 Solaris 内核性能组工作。兴趣包括调度程序、Solaris NUMA 实现、内核可观察性以及算法的可扩展性。

Tariq Magdon-Ismail 更新了 HAT 一章的 SPARC 部分。Tariq 是“性能、可用性和体系结构工程”组的 Staff 工程师，有超过 10 年的 Solaris 经验。他贡献的领域包括大系统性能、内核可扩展性和内存管理体系结构。Tariq 凭借在内存管理领域的工作获得 Sun 公司季度杰出奖。Tariq 拥有马里兰大学 Park 学院计算机科学的荣誉学士学位。

Stuart Maybee 提供了文件系统装配表描述的信息。Stuart 是 Sun 的内核开发组的工程师。

Dworkin Muller 提供了 UFS 在磁盘格式上的信息。Dworkin 在 Sun 的时候曾是 UFS 文件系统的开发者。

David Powell 提供了系统 V IPC 的更新。Dave 是 Solaris 内核开发的工程师，他的很多贡献包括重写系统 V IPC 工具以使用新的资源管理框架设置门限，为 Solaris 10 的服务管理设施（SMF，Service Management Facility）作出了贡献。

Karen Rochford 提供了 UFS 日志的贡献和图表。Karen Rochford 有 15 年的软件工程经验，她过去 3 年是在 Sun 工作。她专注于 I/O 领域，包括设备驱动程序、SCSI、存储控制器固件、RAID，最近致力于 UFS 和 NFS。她拥有俄亥俄州 Berea 的 Baldwin-Wallace 学院的计算机科学和数学的学士学位，科罗拉多大学（科罗拉多温泉城）的计算机科学硕士学位。在她的闲暇时间，Karen 与爱犬为伴，她有一只伯瑞犬和一只法兰德斯牧牛犬。

Eric Saxe 提供了对分配器、NUMA 和 CMT 各章的贡献。Eric Saxe 已经在 Sun 工作了 6 年，是 Solaris 内核开发组的开发工程师。Eric 的大部分工作时间用于分析和提高内核调度的性能，以及虚拟内存子系统在 NUMA，CMT 和其他大系统体系结构上的性能。

Eric Schrock 提供了系统调用的附录。Eric 是 Solaris 内核开发的工程师。他最近的工作是开发和实现 ZFS。

Michael Shapiro 撰写了 kmem 调试和 MDB 的介绍文本。Michael Shapiro 是 Solaris 内核开发组的杰出工程师和 RAS 特性的架构师。他领导了 Sun 的可预测自恢复体系结构的设计和构建，是 DTrace 的共同发明者。Michael 编写了大量工具组件，包括 DTrace、D 编程语言、内核应急（panic）子系统、`fmd (1M)`、`mdb (1M)`、`dumpadm (1M)`、`pgrep (1)`、`pkill (1)` 以及无数对 `/proc` 文件系统、核心文件、崩溃转储和硬件错误处理的增强功能。Michael 作为 Solaris 内核组的成员历经 9 年，他拥有布朗大学的计算机科学硕士学位。

Denis Sheahan 提供了工具一章的关于 Java 的信息。Denis 是 Sun 公司 UltraSPARC T1 体系结构组的资深 Staff 工程师。在 Sun 的 12 年中，Denis 的工作集中在应用软件和 Solaris 操作系统的性能上，重点是数据库、应用服务器和 Java 技术产品。他目前的主要工作是现在和将来产品的 UltraSPARC T1 性能。Denis 拥有爱尔兰都柏林三一学院的计算机科学学士学位。2003 年，他因创新工作获得了 Sun 主席奖。

Tony Shoumack 为性能一书作出了很多贡献并进行了大量审阅。Tony 在 UNIX 和 Solaris 上工作了 12 年，他是 Sun 的客户解决方案组织的工程师，专注于商务应用、数据库和高可用集群系统。

Bart Smaalders 提供了无数的好的思想，他撰写了 NUMA 一章的介绍文本。Bart 是 Solaris 内核开发的资深 Staff 工程师，主要致力于提高 Solaris 运行速度。

Sunay Tripathi 撰写了网络一章。Sunay 是 Solaris 核心技术组的资深 Staff 工程师。在过去 9 年中，他设计、开发和领导了 Sun Solaris 的内核/网络环境的主要项目，以提供新的功能，性能和可扩展性。来到 Sun 之前，Sunay 是德里印度理工学院的研究人员，并在斯坦福服务了两年，在那里，他参加了设计研究中心，创建了智能代理和部分 Mosquito 网络组，试验 IP 网络的移动性。

Andy Tucker 撰写了 zone 的介绍文本。从 2005 年开始，Andy 担任 VMware 的总工程师，工作在 VMware ESX 产品上。在那之前，他在 Sun 公司工作了 11 年，主要研究与 Solaris 操作系统相关的不同领域，特别是调度、资源管理和虚拟化。1994 年他从斯坦福获得了计算机博士学位。

审阅者

特别感谢 Dave Miller 和 Dominic Kay。作为非常出色的审稿人，在书的写作过程中，他们谨慎地审阅了大量的材料，并提供了详细的反馈和评论。

以下各位慷慨地付出了时间和专业技术，审阅了书的原稿。他们发现了错误，提供了建议和评论，这些都极大地提高了最后工作的质量。他们是 Lori Alt、Roch Bourbonnais、Rich Brown、Alan Hargreaves、Ben Humphreys、Dominic Kay、Eric Lowe、Giri Mandalika、Jim Nissen、Anton Rang、Damian Reeves、Marc Strahl、Michael Schuster、Rich Teer 和 Moriah Waterland。

Tony Shoumack 和 Allan Packer 在最后出色地帮助完成了审阅过程，并应用了几条改进。

Richard 的个人致谢

毋庸置疑，本书是真正的集体智慧的结晶。当我们看过名单时，这一版有超过 30 位作者。我喜欢和你们一起工作，现在很高兴有机会感谢你们帮助完成这两本书。

首先，我要感谢我的家人，从我的妻子 Traci 开始，感谢你在这一多年项目中非凡的支持和耐心，能让我集中精力完成这项工作，在这期间，你给了我最好的礼物，我们的新生儿，Boston。我的 4 岁的女儿 Madison 成长得如此之快，已经成为最令人惊异的小淑女。我如此为你骄傲，你对这个项目是如此感兴趣，如此自信地为本书设计了封面。是的，Madi，我们最后可以说我们完成了这本书！

感谢我们的朋友和家人在我离开忙于此书时的耐心。我欠你们几年的露营、晚餐和所有其他我应在场的社会活动！

感谢我的写作此书的伙伴，Jim Mauro。嗨，Jim，我们完成了！感谢你这样一位好朋友，在这项工作的过程中让我一直保持理智！

多谢 Phil Harman，总是在即时短信的另一头陪伴我，了解很多的主意。同样感谢很多次愉快的摄影冒险。

非常感谢 Brendan Gregg 加入我们的工作，和我们一起写了《Solaris 性能与工具》。感谢你提供的深入理解、思想和工具，没有你的加入，这本书不会成为现在的样子。

Mary Lou Nohr，我们的文字加工编辑，我最尊重的人，你的耐心陪伴我们，这个项目从 700 页增加到 1600 页，从一本书变为两本。在破记录的时间内，以不可思议的细节完成了我们发给你的每件事。没有你，本书不会有今天的成功。

感谢 Solaris 开发组，你们所作的无数创新使得 Solaris 的图书创作成为一大乐趣。感谢 Bart Smaalders，Solaris 内核性能组的组长，感谢你对于这个及其他很多项目的深入理解、评论、建议和指导。

感谢所有的提供帮助的被邀请的作者们，感谢你们的贡献，你们的深入理解和语言给了这个 Solaris 的故事一个美好的结尾。

感谢我在 Sun 公司“性能、可用性和体系结构”组的同事们。这两本书的很多内容归功于你们的艰苦努力。

感谢我的资深主管 Ganesh Ramamurthy，感谢他做这个项目的坚强后盾，以及为我们提供的充分支持和资源完成这项工作。

Richard McDougall
Menlo Park，加州
2006 年 6 月

Jim 的个人致谢

万分感谢我们在 Prentice Hall 的资深编辑 Greg Doench，感谢他为这一更新的版本多等了两年，当我们最后交给他两本书而不是一本时也欣然接受。

感谢我们的文字加工编辑 Mary Lou Nohr，在破记录的时间内完成了如此不可思议的工作。

感谢 Brendan Gregg 的非凡努力，为《Solaris 性能与工具》一书作出了大量贡献，同时为《Solaris 内核结构》一书给出了极有价值的反馈。

Marc Strahl 应该得到特别的认可。Marc 是《Solaris 内核结构第 1 版》（也是这一版）的主要审阅者。在第一版的最后完成时刻，我在致谢中漏掉了 Marc。我真心感谢他花在两个版本的时间和支持。

感谢 Solaris 内核工程组的每一个人。你们的支持和热情简直是势不可挡，并且继续创新并创造世界上最好的操作系统。万分感谢。

我的经理 Keng-Tai Ko，感谢他的支持、耐心和灵活，并感谢我的主管 Ganesh Ramamurthy，感谢他难以置信的支持。

感谢我的好朋友，Phil Harman 和 Bob Sneed，感谢他们的聆听、想法和意见，并且很多很多次在我疲劳消沉时为我加油鼓劲。

感谢我的好伙伴 Richard McDougall，感谢他的友谊、领导和远见卓识，以及在 Bay Area 的一百顿的美食和一千杯的美酒。我期望会有更多。

最后，感谢我的妻子 Donna 和我的两个儿子，Frank 和 Dominick，感谢他们的爱、支持和鼓励，以及接受两年多来我所说的“不行，我得写书”。

Jim Mauro
Green Brook，新泽西
2006 年 6 月

目　录

第三部分 资源管理