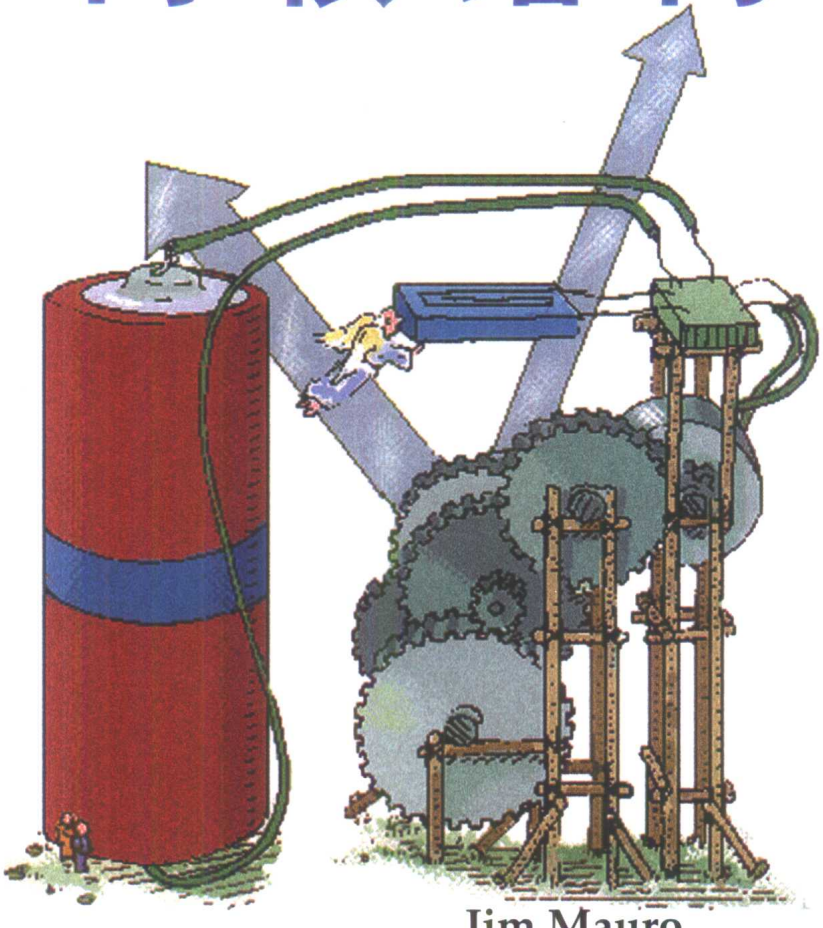
HZ BOOKS

Solaris Internals Core Kernel Architecture

Sun公司核心技术丛书

Sun microsystems

Solaris 内核结构



(美) Jim Mauro Richard McDougall 著

冯锐 张晓亮 过晓冰 陆丽娜 等译

机械工业出版社 China Machine Press

PH PTR

Sun公司核心技术丛书

Solaris内核结构

（原书第2版）

（美） Jim Mauro Richard McDougall 著

冯 锐 张晓亮 过晓冰 陆丽娜 等译

机械工业出版社
China Machine Press

本书介绍 Solaris 操作系统的内核，提供了内核结构以及该操作系统中使用的主要数据结构和算法的大量信息。主要内容包括：Solaris 内核概述、Solaris 存储系统线程、进程和 IPC、文件和文件系统。本书还讲解了该系统的实际应用，用大量实例阐述了应用方法和技巧。本书对于使用 Solaris 操作系统的各类技术人员来讲是一本极具参考价值的专著。

本书版权登记号：图字：01－2000－3395

图书在版编目（CIP）数据

Solaris 内核结构/（美）麦欧（Mauro, J.），（美）麦可古欧（McDougall, R.）著；冯锐等译. －北京：机械工业出版社，2001.9

（Sun 公司核心技术丛书）

书名原文：Solaris Internals Core Kernel Architecture

ISBN 7－111－09292－9

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

中国版本图书馆 CIP 数据核字（2001）第 055710 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：宋 宏 张鸿斌

北京第二外国语学院印刷厂印刷·新华书店北京发行所发行

2001 年 9 月第1版第1次印刷

787mm×1092mm1/16 · 31.25 印张

印数：0 001-5 000 册

定价：55.00 元

译者序

Solaris 素来就以其强大的功能和健壮的稳定性而深受企业用户的青睐；而深入 Solaris 的内核，是众多技术爱好者所竭力追求的境界。在 Linux 如火如荼的公开源码的运动的影响下， Sun 公司终于在 2000 年 2 月公开了 Solaris 源码。但是在浩如烟海的源码中漫游毕竟不是件容易的事情，本书就是为读者点亮的一盏导航明灯。

有谁比父母更了解自己的孩子呢？本书的作者 Jim Mauro 和 Richard McDougall 先生就供职于 Sun 公司，在编写本书的过程中得到了 Sun 公司的很多工程师的大力协助，他们是最有资格编写本书的人员。

本书以 Solaris 7 为基准，同时也涵盖了 Solaris 2.5.1、2.6 的相关信息。本书从四个方面对 Solaris 进行了深入的介绍，第一部分介绍了 Solaris 内核服务、内核同步原语以及内核初始化和启动，第二部分介绍了 Solaris 的内存管理，第三部分介绍了 Solaris 进程、线程的调度机制以及 IPC 机制，最后一部分介绍 Solaris 的文件系统。

本书是迄今为止介绍 Solaris 内核最好的一本书，从一出版以来就受到广大读者的青睐；《Understanding the LINUX Kernel: From I/O Ports to Process Management》（O'Reilly & Associates 出版，Daniel Pierre Bovet、Marco Cesati 著，中国电力出版社已经引进）和本书在 http://www.amazon.com连续数月蝉联同类书籍销量的前两名就是一个明证。

就 Solaris 本身的完整性来讲，本书也还欠缺了一些内容，例如 I/O 子系统、设备驱动程序、网络、流设备等等，幸运的是现在已经有一些书籍介绍相关或相似的内容，例如《Linux 设备驱动程序》（Alessandro Ruibini 著，中国电力出版社引进）、《UNIX 网络编程》（两卷本，W. Richard Stevens 著，清华大学出版社引进）、《Linux IP 协议栈源代码分析》（Satchell, S.T.、Clifford, H.B.J.著，机械工业出版社引进），读者可以根据需要自行选择参考。

本书第 1、2、11、12、13、14、15 章及附录 A 由张晓亮翻译，第 3、4 章由刘敏翻译，第 5、6、7 章及附录 B 由过晓冰翻译，第 8、9、10 章由冯锐翻译，全书由陆丽娜教授校对并统稿，伍卫国副教授、陈革飞、辛炜和吴昊也参与了本书部分章节的校对工作。

虽然译者在翻译本书的过程中力求尊重原著，但是由于中外语言习惯的差异以及译者水平有限，有些内容难免会出现偏差，恳请广大读者不吝赐教，将反馈意见发送至 feng_rui@263.net，译者将不胜感激。

冯锐

2001.4 于西安

前　言

关于 UNIX 内核的参考资料已经很多，大部分是由 Goodheart 和 Cox [10]，以及 Bach [1]、McKusick [19]、Vahalia [39] 等人写作的。这些资料成为想要更好地理解 UNIX 内核的人员所经常使用的参考资料。然而，特意为 Solaris 内核所写的资料就很少。

Solaris 专有资料的缺乏，促使我们编写自己的参考材料。我们以往通过技术白皮书、杂志专栏和培训教程出版过一些这方面的材料，有相当多的人员对这种材料很感兴趣，这也激励我们编写一本专门讨论 Solaris 的完整专著。

关于本书

本书专门介绍 Sun 公司 Solaris 操作系统的内核。Solaris 的快速成长培养了一大批用户、软件开发人员、系统管理员、性能分析员，以及其他一些技术团体的成员，他们都要求对其所用的工作环境有更深层次的了解。

由于本书的重点在于对 Solaris 内核的介绍，所以本书将提供内核结构以及该操作系统中所使用的主要数据结构和算法的大量信息。然而，我们不会从纯学术的角度来阐述这些主题，而是主要着眼于本书内容的实际应用。因此，我们会将重点放在从 Solaris 系统中提取某些信息的方法和工具上，这些信息难以通过标准的捆绑命令和工具访问到。我们将会根据您的工作和兴趣以有意义的方式阐述如何应用这些知识。

为了最大限度地发挥本书的用途，本书还包含有 Solaris 2.5.1、2.6 和 Solaris 7 的相关信息。本书覆盖了主要的 Solaris 子系统，包括存储管理、进程管理、线程、文件和文件系统。书中不涉及底层的 I/O、设备驱动程序、流（stream）以及网络。有关这些主题的参考材料请参见“Writing Device Drivers”[28]、“STREAMS Programming Guide”[29] 和“UNIX Network Programming”[32]。

虽然我们尽可能地在开始讨论一个主题前先介绍一些相关的概念性背景知识，但本书内容不是入门级的，我们假设您已经比较熟悉操作系统的概念，并且使用过基于 UNIX 的操作系统。有一些 C 语言的编程知识会更有用，但并非必需。

由于 Solairs 可以运行于多种硬件平台，所以讨论这些不同的处理器和体系结构等底层细节是不实际的，我们在一些必要的硬件细节上会以 UltraSPARC 为中心。这种方法将比较有效，因为 UltraSPARC 能够代表当今的技术并被大量地安装使用。通常情况下，需要讨论应用到其它支持的处理器和平台上的细节时，我们会提出这个概念。不同之处在于特定的实现细节上，例如每个处理器的硬件寄存器（per-processor hardware registers）。

全书通过描述各种代码段流程时所用的名字来引用特定的内核函数。这些函数属于操作系统内核，不能与 Solaris 系列产品的公用接口——系统调用和库接口相混淆。本书中所引用的函数，除非特别指明，都是内核专有函数，它们是不能由应用程序调用或使用的。

本书的读者对象

我们希望本书能够成为对使用 Solaris 操作系统环境的各类技术人员有用的参考书。具体说来，本书适合以下读者：

- **应用开发人员** 可以在本书中了解到 Solaris 在应用程序编程接口后如何实现各种功能。这能帮助开发人员在开发 Solaris 应用程序时理解每个接口的性能、扩展性及其实现细节。对于他们来说，有关进程调度、进程间通信和文件系统性能的章节是最有用的。
- **设备驱动程序和内核模块开发人员** 设备驱动程序和驱动程序、流模块以及可加载的系统调用的内核模块开发人员可以在本书中学习 Solaris 操作环境的通用体系和实现理论。本书中有关 Solaris 内核框架和功能的部分（尤其是加锁和同步的原语章节）比较重要。
- **系统管理员、系统分析员、数据库管理员和 ERP 管理员** 这类读者负责性能调控和容量规划，可以从本书中学习主要的 Solaris 子系统的运作特性。文件系统缓存和存储管理等章节提供了大量 Solaris 在实际使用环境中如何运行的信息。Solaris 可调参数（详见附录 A）的运算方法贯穿全书。
- **技术支持人员** 负责 Solaris 的诊断、调试和支持的人员将从本书中学习到大量的 Solaris 实现细节的信息。每一章中都提供了主要的数据结构和数据流程图，有助于对 Solaris 系统的调试和导航。
- **只想进一步了解 Solaris 内核如何工作的系统用户** 在每章的开始都提供了高度概括的描述内容。

除了以上列出的各种技术人员，我们相信学术界成员在研究一个产品化内核如何实现主要的子系统、解决操作系统开发中固有问题时，也能够在本书中找到有价值的信息。

本书的结构

本书按逻辑结构分为几个部分，每个部分都包括含有相关内容的数章。我们的目的是提供一个系统的构造模块，即后面的部分是建立在前面部分的基础之上的。然而，对那些已经熟悉了操作系统的某些特定部分的读者来说，可以根据自己的需要选择阅读。

- 第一部分：Solaris 内核简介
  - 第 1 章——Solaris 入门
  - 第 2 章——内核服务
  - 第 3 章——内核同步原语
  - 第 4 章——内核引导和初始化
- 第二部分：Solaris 内存管理
  - 第 5 章——Solaris 内存体系结构
  - 第 6 章——内核内存
  - 第 7 章——内存的监控
- 第三部分：线程、进程和 IPC

- 第 8 章——Solaris 多线程的进程体系结构
- 第 9 章——Solaris 内核调度程序
- 第 10 章——进程间通信

• 第四部分：文件和文件系统

- 第 11 章——Solaris 文件和文件 I/O
- 第 12 章——文件系统综述
- 第 13 章——文件系统框架
- 第 14 章——UNIX 文件系统
- 第 15 章——Solaris 文件系统缓存

Solaris 源代码

2000 年 2 月，Sun 公司公布了 Solaris 源代码。本书提供了必要的 Solaris 源代码参考，可以作为学习 Solaris 内核框架和体系的引导。

应该注意的是 Sun 公开的是 Solaris 8 的源代码。虽然本书仅涵盖了 Solaris 7 及其以前的版本，但几乎所有的资料都和 Solaris 8 相对应。

资料升级以及相关资料

为了完成本书，我们建立了一个 Web 站点以提供我们所引用的升级的资料和工具，并含有相关主题资料的链接。站点地址是：

http：//www.solarisinternals.com

我们将用本书的有关信息和未来的关于 Solaris 内核的作品定期更新此站点。同时将在那里提供有关 Solaris 7 和 8 之间差异的信息，并张贴当前版本中所遇到的错误，共享读者反馈和建议以及其它相关内容。

原书书名：Solaris Internals Core Kernel Architecture
原书书号：0 - 13 - 022496 - 0

目　录

第二部分 Solaris 内存管理

第三部分 线程、进程和 IPC

第四部分 文件和文件系统

第五部分 附 录

第一部分　Solaris 内核简介

- Solaris 入门
- 内核服务
- 内核同步原语
- 内核引导和初始化

第 1 章　Solaris 入门

UNIX 系统非常成功。在写出这些字的时候，全世界已经有超过 3000 个 UNIX 系统在运行使用。

—S.R.Bourne，The UNIX Sytem，1983

自从 1982 年 Sun-1 工作站诞生以来，Sun 系统就已经开始运行基于 UNIX 的操作系统。上面 Steve Bourne 的引言表明那时的 UNIX 市场是多么小。而今，已有成千上万的 UNIX 系统运行在各个应用层次，从单用户系统、实时控制系统，到关键的军事、商业环境，而 Solaris 在其中占有很大的比重。

Solaris 从最初安装开始就快速增长。它可以运行在 Sun 和 OEM 的 SPARC 处理器体系上，也可以运行在标准的基于 Intel 的系统上。Solaris 的适用范围从单处理器系统到 64 个处理器的 Sun Enterprise 10000 系统。

1.1　Solaris 简史

Sun 的 UNIX 操作环境最开始是针对 Sun-1 工作站提供的 BSD UNIX 的一个移植。Sun 的 UNIX 的最早版本叫做 SunOS，现在它是 Solaris 的核心操作系统组件的名字。

SunOS 1.0 是基于 Berkeley 实验室 1982 年的 BSD 4.1 的一个移植。那时 SunOS 是在基于 Motorola 68000 的单处理机 Sun 工作站上实现的。SunOS 小而紧凑，这个工作站只有一个 MIPS（每秒百万条指令）的处理器和大约 1M 的内存。

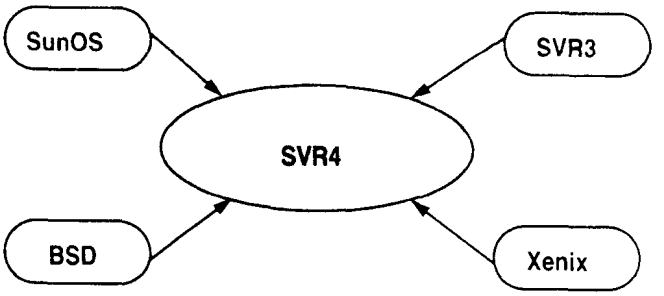
在 20 世纪 80 年代中早期，网络化的 UNIX 系统正在流行；网络化无处不在，同时也是 SUN 的计算战略的重要部分。SUN 在开发技术方面发明了具有重大意义的方法，它能够支持分布式、基于网络的计算。这些技术包括构建分布式应用程序的接口（远程过程调用，或叫 RPC），通过网络共享数据的操作系统程序（网络信息系统，NIS），以及分布式计算文件系统 NFS。在 SunOS 中加入了远程文件共享，要求操作系统作出很大的修改。1984 年，SunOS

2.0 提供了虚拟文件系统结构以实现多种文件系统类型，从而支持 NFS 文件系统。网络文件系统资源公开授权，随后被移植到现存的几乎所有现代操作系统平台上。

在 Sun 平台上运行的应用程序的数量稳步增加，每种新的应用程序都对系统提出了更多的要求，也激励着系统不断创新进步。应用程序需要更好的用于数据共享和可执行对象的软件。对于共享程序库、内存映射文件和共享内存的需要一起导致了 SunOS 虚拟存储系统的重大体系重构。SunOS 4.0 引入了新的虚拟存储系统，将多种设备和对象抽象为虚拟内存，方便了将文件映射、共享内存和硬件设备映射到进程。

在 20 世纪 80 年代，对于处理能力的需求超出了工业界处理器速度的不断增长。为了满足这种需求，新型系统被开发出来，多个处理器共享同一系统内存和输入/输出（I/O）底层结构，这种进步要求操作系统也要做出进一步的改进。首先在 SunOS 4.1 中出现了对异步多处理器的实现，它的内核可以每次仅在一个处理器上运行，而同时能够在可用的任意处理器上调度用户处理程序。多进程的工作任务经常能够在拥有多个处理器的系统上获得更大的吞吐量。异步多处理器的实现是一个很大的进步。然而，当加入更多的处理器时，扩展性就会迅速下降。很明显我们需要一个更好的多处理器实现。

这时，Sun 正在和 AT&T 进行一个联合开发项目，SunOS 虚拟文件系统结构和虚拟存储系统成为 UNIX System V Release 4（SVR4）的核心。SVR4 UNIX 加入了来自 SunOS、SVR3、BSD UNIX 和 Xenix 的特性，如下图所示。国际计算机有限公司（ICL）将新的 SVR4 UNIX 移植到 SPARC 处理器体系上，并且发布了 SPARC 上 SVR4 的参考资料。



随着多处理器系统市场按照预期的不断增长，Sun 将开发的重点放在解决多处理器扩展性的新操作系统内核上。新的内核允许执行多线程，并提供了在进程（应用程序）级线程化的功能。这个新的内核通过细粒度（fine-grained）的加锁策略，形成了如今的 Solaris 所具有的扩展性的基础。新内核和 SVR4 操作环境由此成为 Solaris 2.0 的基础。

基本操作系统的变化伴随着一个新的命名的出现：Solaris 这个名称被用来表示这种操作环境，而 SunOS 是其中的一个子集。因此，老的 SunOS 保留了 SunOS 4.X 的版本，并使用 Solaris 1.X 作为其操作环境版本。基于 SVR4 的环境采用一个 SunOS 5.X 的版本（从 SunOS 5.0 开始发布），称做 Solaris 2.X 操作环境。该命名使得许多人将 SVR4 以前发布的版本称为 SunOS，而基于 SVR4 的称为 Solaris。表 1-1 描述了 Solaris 从它的最早版本到版本 7 的发展

过程。

新的 Solaris 2.0 操作环境采用了模块化的方法，使得它在多种使用不同指令体系的平台上都可以实现。1993 年 Solaris 能够用于基于 Intel PC 的体系，并因此而极大地扩展了 Solaris 的使用范围。1999 年 10 月，Sun 宣布支持 Intel Itanium 处理器上的 Solaris。

下一个主要的里程碑是 Solaris 7 中引入 64 位的实现。完全 64 位的支持使得内核和进程能够访问更大的地址空间和使用扩展 64 位数据类型。Solaris 7 也提供了对现有的 32 位应用程序的完全兼容，支持执行当前的 32 位和 64 位的应用程序。

表 1-1 Solaris 版本历史

日期	版 本	注 释
1982	Sun UNIX 0.7	• Sun 的第一版 UNIX，基于 Unisoft 的 4.BSD • 与 Sun-1 绑定，Sun-1 是 Sun 的基于 Motorola 68000 处理器的第一个工作站；Sun Windows GUI
1983	SunOS 1.0	• Sun-2 工作站，基于 68010
1985	SunOS 2.0	• 虚拟文件系统（VFS）和 vnode 结构允许多个并发的文件系统类型。 • 用 VFS/vnode 结构实现了 NFS
1988	SunOS 4.0	• 将文件系统高速缓存和存储系统整合起来的新的虚拟存储系统 • 增加动态链接 • 第一个基于 SPARC 的 Sun 工作站——Sun 4。支持基于 Intel 的 Sun 386i
1990	SunOS 4.1	• 支持 SPARCstation1 + 、IPC、SLC • OpenWindows 图形环境
1992	SunOS 4.1.3	• Sun4m 系统（SPARCstation-10 和-600 系列 MP（多处理器）服务器）的异步多处理（ASMP）
1992	Solaris 2.0	• Solaris 2.x 诞生，移植自 System V Release 4.0 • VFS/vnode、VM 系统，与 SunOS 提出的近似的共享内存 • 只有单处理机 • 首次发布 Solaris 2，Solaris 2.0 仅是一个桌面开发版
1992	Solaris 2.1	• 4 通道的对称多处理（SMP）
1993	Solaris 2.2	• 支持大文件系统（大于 2G） • SPARCserver 1000 和 SPARCcenter 2000（sun4d 体系结构）
1993	Solaris 2.1-x86	• 移植到 Intel i386 体系结构的 Solaris
1993	Solaris 2.3	• 8 通道的 SMP • 增加设备电源管理和系统挂起/重启功能 • 新的目录名检索高速缓存
1994	Solaris 2.4	• 20 通道的 SMP • 用新的内核内存分配符（slab 分配符）代替 SVR4 的 buddy 分配符 • 高速缓存文件系统（cachefs） • CDE 视窗系统

（续）

日期	版　本	注　释
1995	Solaris 2.5	• 内核和 System V 共享内存的大页面支持 • 加入快速本地进程间通信（Doors） • NFS Version 3 • 支持 sun4u（UltraSPARC）体系。引入了基于 UltraSPARC-I 的产品——Ultra-1 工作站
1996	Solaris 2.5.1	• 支持多处理器基于 UltraSPARC 系统的第一版 • 64 通道的 SMP • 引入 Ultra-Enterprise 3000-6000 服务器
1996	Solaris 2.6	• 增加了对大文件的支持（大于 2G 的文件） • 动态处理器集 • 基于内核的 TCP socket • 锁统计 • UFS 直接 I/O • 动态重配置
1998	Solaris 7	• 64 位内核以及进程地址空间 • 集成日志 UFS • 带优先级的内存分页算法

表 1-1 中的信息显示了每个主要的 Solaris 版本的加入的重要新特性。所有这些特性的细节信息在 Solaris 版本 What's New 文档中可以找到，它是随 Solaris 提供的文档的一部分。

1.2　关键的不同之处

在 20 世纪 90 年代，Solaris 的开发进度仍旧飞速前进。有一个关键的特性使 Solaris 区别于早期的 UNIX 实现。

- 对称多处理——Solaris 的实现范围从单处理器系统到 64 个处理器的对称多处理器服务器。Solaris 提供了线性的扩展，直到现在最多支持 64 个处理器。
- 64 位内核和进程地址空间——一个支持 64 位平台的 64 位内核提供了 LP64 执行环境（LP64 指的是数据模型：长型（long）和指针（pointer）数据类型是 64 位）。同时还提供了 32 位的应用程序环境以便 32 位的二进制执行文件能和 64 位应用程序一起在 64 位的 Solaris 内核上执行。
- 多平台支持——Solaris 支持范围广泛的 SPARC 和 Intel x86 基于微处理器的体系。分层的体系结构使得百分之九十的 Solaris 源程序是与平台无关的。
- 模块化二进制内核——Solaris 内核使用动态链接和动态模块来将内核分到模块化的二进制文件中。有一个核心内核二进制文件包含中心程序；而设备驱动程序、文件系统、调度程序和一些系统调用则作为动态可加载模块来实现。因此，Solaris 内核以二进制形式交付，而不是源文件和目标文件，并且在修改参数或添加新功能时不要求重新编译内核。
- 多线程进程的执行——一个进程可以含有不止一个执行线程，这些线程可以并发运行在一个或多个处理器上。因此，单个进程可以使用多个处理器来处理并发线程，

从而使多处理器平台的运行效率更高。

- 多线程内核——Solaris 内核使用线程作为调度和执行的实体：内核将中断和内核服务作为常规的内核线程来调度。此关键特性为中断提供了可扩展性和低延迟的中断响应。

  以前的 UNIX 实现通过控制处理器优先级级别来确保对临界中断数据结构的互斥访问。但是这样无法阻塞中断代码，导致可扩展性很差。Solaris 通过将中断作为线程来调度以提高并发度，并使用常规的内核锁来确保数据结构的互斥访问。
- 完全的抢占式内核——Solaris 内核是完全可抢占的，不需要通过硬件中断级的控制来保护临界数据——使用锁对内核数据的访问进行同步。也就是说，需要运行的线程可以中断另一个低优先级的线程；因此可以实现低延迟的调度和低延迟的中断分派。举个例子，一个进程在因等待磁盘 I/O 而睡眠之后被唤起，它可以被立即调度，而不需等待调度程序的运行。另外，由于没有产生优先级的等级和阻塞中断，系统就无需在中断处理期间定期将活动挂起，所以系统资源的利用率会更高些。
- 支持多调度程序——Solaris 提供有一个可配置的调度程序环境。多调度程序可并发运行，各有各的调度算法和优先级。调度算法作为内核模块，被动态加载到操作系统中。Solaris 提供了一个表驱动、使用衰减（usage-decayed）、分时的用户调度程序（TS）；还有一个视窗系统优化分时调度程序（IA）；一个实时的固定优先级调度程序（RT）。可以通过 Solaris 资源管理器包来加载一个可选的公平的共享调度程序类。
- 支持多文件系统——Solaris 提供了一个虚拟文件系统（VFS）结构，允许系统中配置多种文件系统。该结构实现了几个基于磁盘的文件系统（UNIX 文件系统、MS-DOS 文件系统、CD-ROM 文件系统等），以及网络文件系统（NFS V2 和 V3）。虚拟文件系统结构也实现了伪文件系统，包括进程文件系统 procfs，该系统将进程抽象为文件。虚拟文件系统和虚拟存储系统结合在一起，提供动态文件系统高速缓存，用可用的空闲内存作为一个文件系统高速缓存。
- 处理器划分和绑定——一些专门的软件允许细粒度的处理器控制，可以将处理器配置到调度群里以划分系统资源。
- 按需分页（Demand-paged）的虚拟存储系统——该特性允许系统按照需要加载应用程序，而不是将整个可执行应用程序或库映像都加载到内存中去。按需分页的方法加速了应用程序的启动，并相应地会减少内存的占用。
- 模块化虚拟存储系统——虚拟存储系统将虚存功能分散到不同的层次。地址空间层、分段驱动程序和相关硬件组件固定到一个硬件地址转换（HAT）层。分段驱动程序可以将内存抽象为文件，而文件可以将内存映射到一个地址空间。分段驱动程序能够使不同的抽象在一个地址空间中显示出来，包括物理内存和设备。
- 模块化设备 I/O 系统——动态加载设备和总线驱动程序允许安装和配置一整套分层的总线和设备。一个设备驱动接口（DDI）在特定平台底层结构和设备驱动程序之间进行屏蔽，因此使得设备驱动程序具有最大的可移植性。
- 整合网络——通过数据链路供应接口（DLPI），可以配置多个并发的网络接口，并在