系统管理 剖析Linux扩展文件系统ext4Linux认证考试 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/555/2021_2022__E7_B3_BB_E7_BB_9F_E7_AE_A1_E7_c103_555903.htm 第 4 个扩展文件系统，即 ext4，是下一代的日志文件系统，它与上一代文件系统 ext3 是向后兼容的。尽管 ext4 目前还不是标准文件系统，但它将成为大部分下一代 Linux? 发行版的默认文件系统。了解 ext4，以及它为什么将成为您最喜欢的新文件系统。 Linux 内核的每次发行都伴随一些惊喜，今年 12 月份发行的 2.6.28 也不例外。这个发行版是首个稳定的 ext4 文件系统（它还包含其他出色的特性，比如正在开发的 Btrfs）。这个下一代扩展文件系统提供更好的伸缩性、可靠性和许多新功能。ext4 的伸缩性如此之大，以致最大的文件系统所用的磁盘空间将达到 100 万 TB。 扩展文件系统的简史 第一个受 Linux 支持的文件系统是 Minix 文件系统。这个文件系统有严重的性能问题，因此出现了另一个针对 Linux 的文件系统，即扩展文件系统。第 1 个扩展文件系统（ext1）由 Remy Card 设计，并于 1992 年 4 月引入到 Linux 中。ext1 文件系统是第一个使用虚拟文件系统（VFS）交换的文件系统。虚拟文件系统交换是在 0.96c 内核中实现的，支持的最大文件系统为 2 GB。 第 2 个扩展文件系统（ext2）也是由 Remy Card 实现的，并于 1993 年 1 月引入到 Linux 中。它借鉴了当时文件系统（比如 Berkeley Fast File System [FFS]）的先进想法。ext2 支持的最大文件系统为 2TB，但是 2.6 内核将该文件系统支持的最大容量提升到 32TB。 第 3 个扩展文件系统（ext3）是 Linux 文件系统的重大改进，尽管它在性能方面逊色于某些竞争对手。ext3 文件系

统引入了日志 概念，以在系统突然停止时提高文件系统的可靠性。虽然某些文件系统的性能更好（比如 Silicon Graphics 的 XFS 和 IBM? Journaled File System [JFS]），但 ext3 支持从使用 ext2 的系统进行就地（in-place）升级。ext3 由 Stephen Tweedie 实现，并于 2001 年 11 月引入。 今天，我们已经拥有第 4 个扩展文件系统（ext4）。ext4 在性能、伸缩性和可靠性方面进行了大量改进。最值得一提的是，ext4 支持 1 EB 的文件系统。ext4 是由 Theodore Tso（ext3 的维护者）领导的开发团队实现的，并引入到 2.6.19 内核中。目前，它在 2.6.28 内核中已经很稳定（到 2008 年 12 月为止）。 ext4 从竞争对手那里借鉴了许多有用的概念。例如，在 JFS 中已经实现了使用区段（extent）来管理块。另一个与块管理相关的特性（延迟分配）已经在 XFS 和 Sun Microsystems 的 ZFS 中实现。 在 ext4 文件系统中，您可以发现各种改进和创新。这些改进包括新特性（新功能）、伸缩性（打破当前文件系统的限制）和可靠性（应对故障），当然也包括性能的改善。 功能 ext4 引入了大量新功能，但最重要的是与 ext3 的向后和向前兼容性，以及在时间戳上的改进。这些改进立足于提高未来的 Linux 系统的性能。 向后和向前兼容性 由于 ext3 是 Linux 上最受欢迎的文件系统之一，因此应该能够轻松迁移到 ext4。为此，ext4 被设计为在 extent 方面具有向后和向前兼容性。ext4 与 ext3 是向前兼容的，这样就可以将 ext3 文件系统挂载为 ext4 文件系统。为了充分利用 ext4 的优势，必须实现文件系统的迁移，以转换和利用新的 ext4 格式。您还可以将 ext4 挂载为 ext3（向后兼容），但前提是 ext4 文件系统不能使用区段（将在性能小节对其进行讨论）。 除了兼容性特性之外，您还

可以逐步地将 ext3 文件系统迁移到 ext4。这意味着没有移动的旧文件可以保留 ext3 格式，但新的文件（或已被复制的旧文件）将采用新的 ext4 数据结构。您可以通过这种方式在线将 ext3 文件系统迁移到 ext4 文件系统。 提高时间戳分辨率和扩展范围 令人惊讶的是，ext4 之前的扩展文件系统的时间戳都是以秒为单位的。这已经能够应付大多数设置，但随着处理器的速度和集成程度（多核处理器）不断提升，以及 Linux 开始向其他应用领域发展（比如高性能计算），基于秒的时间戳已经不够用。ext4 设计时间戳时考虑到未来的发展，它将时间戳的单位提升到纳秒。ext4 给时间范围添加了两个位，从而让时间寿命再延长 500 年。 伸缩性 文件系统未来发展的一个重要方面就是伸缩性，即根据需求进行伸缩的能力。ext4 以多种方式现实了强大的伸缩性，它的伸缩性超越了 ext3，并且在文件系统元数据管理方面开辟了新领域。 突破文件系统的限制 ext4 的一个明显差别就是它支持更大的文件系统、文件和子目录。ext4 支持的最大文件系统为 1 EB（1000 PB）。虽然根据今天的标准这个文件系统已经非常巨大，但存储空间的消费会不断增长，因此 ext4 必须考虑到未来的发展。ext4 支持最大 16 TB 的文件（假设由 4KB 的块组成），这个容量是 ext3 的 8 倍。 最后，ext4 也扩展了子目录的容量，将其从 32KB 扩展到无穷大。这是极端情况，我们还需要考虑文件系统的层次结构，因为它的最大存储容量为 1 EB。此外，目录索引也优化为类似于散列 B 树结构，因此尽管限制更加多，但 ext4 支持更快的查找。 区段 ext3 分配空间的方式是其主要缺点之一。ext3 使用空闲空间位映射来分配文件，这种方式不是很快，并且伸缩性不强。ext3 的格式对小

文件而言是很高效的，但对于大文件则恰恰相反。ext4 使用区段取代 ext3 的机制，从而改善了空间的分配，并且支持更加高效的存储结构。区段是一种表示一组相邻块的方式。使用区段减少了元数据，因为区段维护关于一组相邻块的存储位置的信息（从而减少了总体元数据存储），而不是一个块的存储位置的信息。 ext4 的区段采用分层的方法高效地表示小文件，并且使用区段树高效地表示大文件。例如，单个 ext4 inode 有足够的空间来引用 4 个区段（每个区段表示一组相邻的块）。对于大文件（包括片段文件），一个 inode 能够引用一个索引节点，而每个索引节点能够引用一个叶节点（引用多个区段）。这种持续的区段树为大文件（尤其是分散的文件）提供丰富的表示方式。这些节点还包含自主检查机制，以阻止文件系统损坏带来威胁。 性能 衡量一个新文件系统的最重要指标就是它的根本性能。这常常是最难实现的指标，因为当文件系统变得庞大并且要求实现高可靠性时，将会以损害性能为代价。但是，ext4 不仅解决了伸缩性和可靠性，它还提供各种改善性能的方法。 文件级预分配 某些应用程序，比如数据库或内容流，要求将文件存储在相邻的块上（利用相邻块的读优化和最大化读的命令-块比率）。尽管区段能够将相邻块划分为片段，但另一种更强大的方法是按照所需的大小预分配比较大的相邻块（XFS 以前就是采用这种方法）。ext4 通过一个新的系统调用来实现这个目的，这个调用将按照特定的大小预分配并初始化文件。然后，您就可以写入必要的数据，并为数据提供不错的读性能。 延迟块分配 另一个基于文件大小的优化是延迟分配。这种性能优化延迟磁盘上的物理块的分配，直到块被刷入到磁盘时才进行

分配。这种优化的关键是延迟物理块的分配，直到需要在磁盘上写这些物理块时才对其进行分配并写到相邻的块。这类似于持久化预分配，惟一的区别是文件系统会自动执行这项任务。不过如果预先知道文件的大小时，持久化预分配是更好的选择。 多个块分配 这是最后一个与相邻块相关的优化，即针对 ext4 的块分配器。在 ext3 中，块分配器的工作方式是每次分配一个块。当需要分配多个块时，非相邻块中可能存在相邻的数据。ext4 使用块分配器修复了这个问题，它能够在磁盘上一次分配多个块。与前面其他优化一样，这个优化在磁盘上收集相关的数据，以实现相邻读优化。 多个块分配的另一个方面是分配块时需要的处理量。记住，ext3 一次只分配一个块。在最简单的情况下，每个块的分配都要有一个调用。如果一次分配多个块，对块分配器的调用就会大大减少，从而加快分配并减少处理量。 可靠性 ext4 文件系统可能会扩展得比较大，这将导致可靠性问题。但 ext4 通过许多自主保护和自主修复机制来解决这个问题。 执行文件系统日志校验和 和 ext3 一样，ext4 也是一个日志文件系统。日志记录就是通过日记（磁盘上相邻区域的专门循环记录）记录文件系统的变更的过程。因此，根据日志对物理存储执行实际变更更加可靠，并且能够确保一致性，即使在操作期间出现系统崩溃或电源中断。这样做可以减少文件系统损坏的几率。但是即使进行日志记录，如果日志出现错误仍然会导致文件系统损坏。为了解决这个问题，ext4 对日志执行校验和，确保有效变更能够在底层文件系统上正确完成。在 参考资料 小节可以找到其他关于日志记录（ext4 的重要部分）的资料。ext4 支持根据用户需求采用多种模式的日志记录。例如，ext4

支持 Writeback 模式，它仅记录元数据；或 Ordered 模式，它记录元数据，但写为元数据的数据是从日志中写入的；或 Journal 模式（最可靠的模式），它同时记录元数据和数据。注意，虽然 Journal 模式是确保文件系统一致的最佳选择，但它也是最慢的，因为所有数据都要经过日志。 在线磁盘碎片整理 尽管 ext4 添加一些特性来减少文件系统的碎片（比如将相邻块分配为区段），但随着系统使用时间的增加，碎片是难以完全避免的。因此出现了在线碎片整理工具，它们可以对文件系统和单个文件执行碎片整理，从而改善性能。在线碎片整理程序是一个简单的工具，它将文件复制到引用相邻区段的新 ext4 inode。 在线碎片整理还可以减少检查文件系统所需的时间（fsck）。ext4 将未使用的块组标记到 inode 表中，并让 fsck 进程忽略它们以加快检查速度。当操作系统因内部损坏（随着文件系统变大，这是不可避免的）而检查文件系统时，ext4 的设计方式将能够提高总体可靠性。 结束语 针对 Linux 的扩展文件系统有着漫长而丰富的历史 从 1992 年首次引入 ext1 到 2008 年引入 ext4。ext4 是首个专门为 Linux 设计的文件系统，并且事实证明它是高效、稳定、强大的文件系统。ext4 随着文件系统研究的深入而不断发展，并且借鉴其他新文件系统的先进思想（比如 XFS、JFS、Reiser 和 IRON 容错文件系统技术）。尽管目前预测 ext5 将会是什么样子还为时过早，但有一点是很明确的，它将主导企业级 Liunx 系统。