Disco：在可扩展的多处理器上运行商品操作系统

由于商品操作系统不可扩展，因此随着硬件发展，应该修改哪些层以实现可扩展性？诸如可扩展共享内存多处理器之类的硬件创新需要对操作系统软件进行重大改变。这些变化有巨大的成本，它可能会带来不稳定性，不可靠性和不良数据共享、它需要商业公司的合作。因此，必须找到新的方法来开发用于可扩展多处理器机器的软件系统。本文试图通过引入一层虚拟机监视器来解决这个问题，该监视器能够通过多路复用硬件来运行多个商品操作系统以实现最大的利用率。

作者正试图解决的主要问题是缩小硬件升级和系统软件功能之间的差距。他们的观点是，系统软件需要大量的时间来整合硬件创新，而硬件供应商也面临着说服 OS 公司支持这些改变的额外挑战。考虑到 Windows 8 在 Windows 8 的推出中得到了很好的支持，让系统软件支持多代硬件显然是操作系统供应商面临的一个问题。此外，作者还需要解决通常影响虚拟机的问题，如开销。作者重新提出了虚拟机监视器的概念，并用它来从操作系统中抽象出底层硬件的细微差别。这使他们能够掩饰 FLASH 系统的不统一性质，同时也让他们的系统能够支持多种操作系统。

作者提出了当前操作系统可能面临的硬件演变方式的一些问题。我认为处理异构系统是一个值得关注的问题，分布式内核结构看起来像是一个有趣的潜在解决方案。然而，作者并没有真正解决如何编译这些分布式内核系统的应用程序。另外，为了克服通常困扰虚拟机系统的问题，他们还引入了一些数字，如果对虚拟机进行了增强。他们通过让虚拟机知道进程的优先级来减少资源管理的开销。在页面复制和迁移中，他们引入了优化 NUMA 系统内存布局的功能。他们为当今仍在使用的虚拟机引入了一些功能。例如，它们允许 VM 上运行的所有系统访问相同的地址空间，并将映射从系统物理地址映射到 TLB 中的实际机器地址。

作者将这些想法一步步实现，设计并实现了 Disco。Disco 是一款虚拟机监视器，它通过在操作系统和硬件之间增加一层额外的系统软件，在可扩展的共享内存多处理器系统上运行许多商品操作系统。由于代码库较小，与商品操作系统相比，它可以更容易地演变和修改。该监视器被推荐作为可扩展操作系统

的替代品，这些操作系统是复杂的，需要很多修改并且具有巨大的实施成本。Disco 通过在早期 FLASH 机的系统模拟器上运行实际工作负载进行评估，一些经验数据表明，虚拟化可以在很小的开销下完成，并获得类似于在大型 cc-NUMA 机器上运行的可扩展 OS 的好处。

Disco 的重要贡献在于它提供给在可扩展 cc-NUMA 机器上运行的商品操作系统的非 NUMA 界面。Disco 为所有正在运行的操作系统提供接口，并将所有这些资源（如 CPU，内存和 I / O 设备）虚拟化，作为硬件和操作系统之间的额外绝缘层。迪斯科使用标准网络接口支持虚拟机之间的通信，使共享存储器通信更加轻松高效。本机操作系统上的所有特权指令陷入 Disco，并为每个虚拟 CPU 模拟这些指令。它通过使用某些共享数据结构提供了从虚拟到机器映射的地址转换的高效机制。提供动态页面迁移和页面复制等功能，以支持机器的 NUMAness，因为商品操作系统是 NUMA 不知道的。所有 I / O 通信和硬件中断都被 Disco 拦截，从而有效地在虚拟机之间共享资源。

作者通过大量工作负载（包括工程，科学和数据库）提供了详细的实施评估结果。比较 IRIX 和迪斯科，与不使用 VMM（主要来自陷阱和 TLB 重新加载）相比，VMM 产生的开销在 3％和 16％之间。内存在虚拟机之间进行分区，NFS 提供的内存比可用内存更多。动态页面迁移和复制图表显示，与商品操作系统相比，性能提高了约 33％。