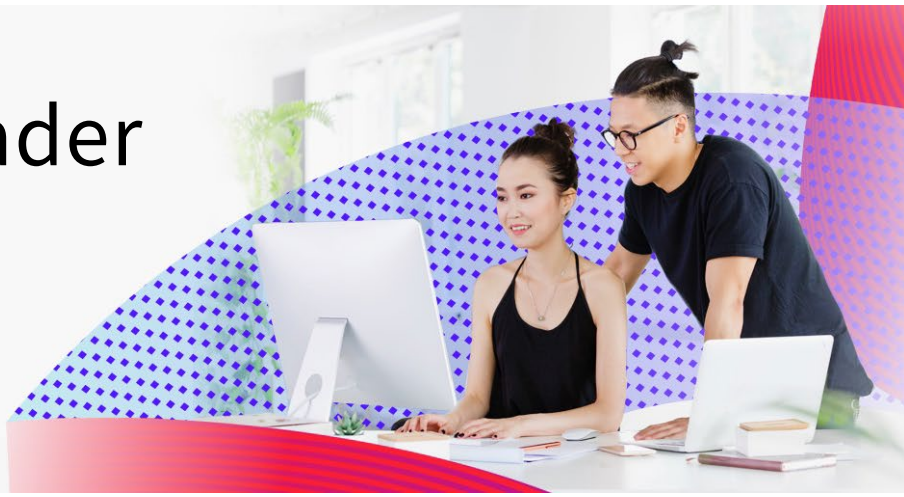
ABBYY

ABBYY FineReader Server 14

性能指南



概览

本指南说明了可帮助您实现 ABBYY FineReader Server 14 的最高性能的最佳实践。

ABBYY FineReader Server 14 是一款功能强大且可靠的基于服务器的解决方案，用于自动处理企业环境中的 OCR 和文档转换工作。 该解决方案可轻松扩展，满足多个作业的高容量并行处理需求。

主要特性：

- 高可扩展性，支持使用自有硬件或云资源来满足任何文档处理需求
- 日处理量高达 1,000,000 页彩色或 3,000,000 页黑白文档，可满足大型企业的需求
- 可靠性高，易于使用
- 支持多个并行会话

系统速度和质量的影响因素：

- 硬件特性
- 输入文件类型
- 选定处理参数

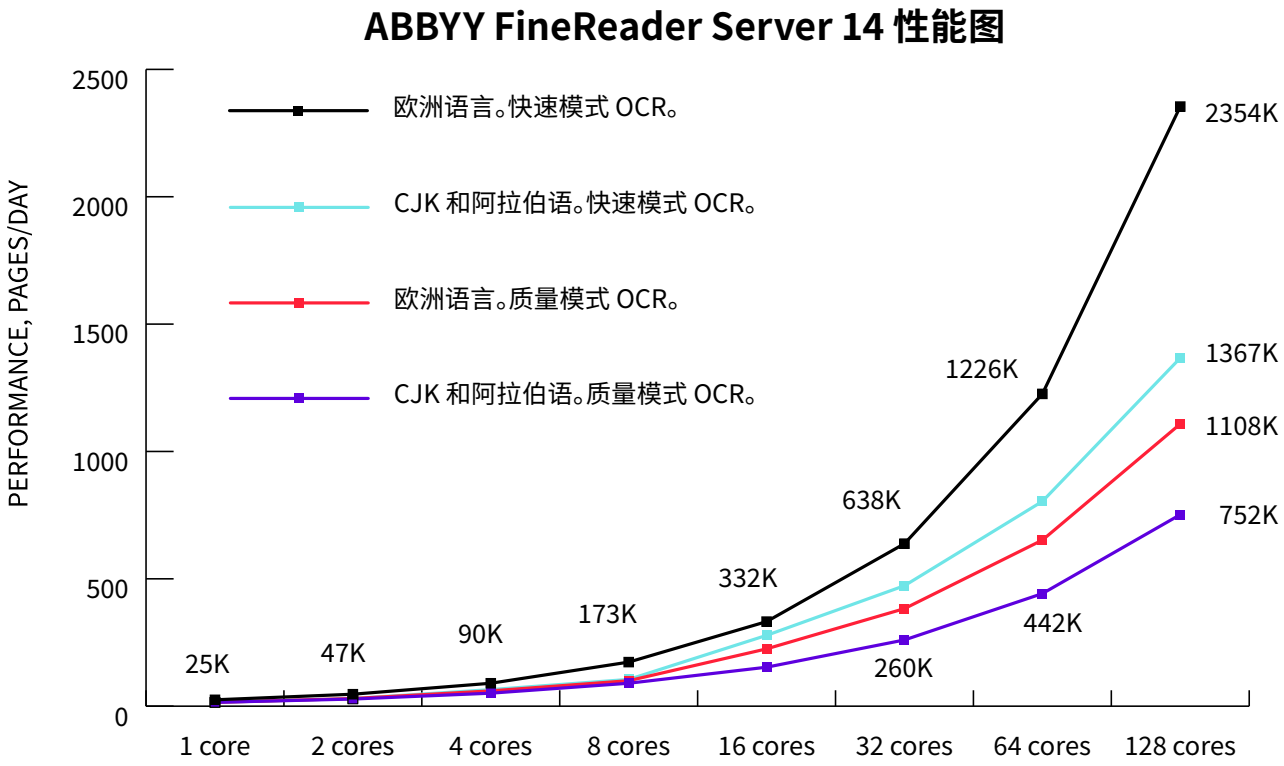
本指南包括性能测试数据，并提供可帮助您实现 ABBYY FineReader Server 14 系统最高性能的相关建议。

目录

ABBYY FineReader Server 14 性能

任何 ABBYY FineReader Server 系统的性能与 OCR 中所使用的 CPU 核数具有直接关系，因此您应该首先考虑项目所需的处理站和 CPU 数量。

下图说明了性能随 CPU 核数增加而提高的情况（使用默认工作流程处理同一类型的示例文件）。



计算项目所需的 CPU 数量

要计算所需的 CPU 数量：

1. 收集 50-100 个文档的代表性示例，并利用这些文档创建包含约 1,000 个文件的测试集
2. 在采用单核 CPU 的单处理站中处理测试集，并测量处理这些文件所需的时间 t
3. 计算平均处理速度（页数/分钟），方法是处理的页数除以时间 t：v=q/t
4. 计算单核 CPU 将在系统运行时间 T 内处理的页数：P=v*T
5. 计算项目所需的最少 CPU 核数，方法是待处理的总页数除以单核 CPU 可处理的页数：n=Q/P
6. 将第 5 步中得到的值增加 20%，以弥补共用系统资源（即缓存、网络等）导致的任何延迟：N=1.2*n

性能影响因素

在计算系统性能并调整以适应扩展系统时，应将以下因素考虑在内。

硬件性能

1 处理站 CPU 速度

CPU 速度越快，OCR 速度也越快。下表显示了处理性能与所使用的 CPU 以及核数之间的关系 。

表 1.使用 2、4 和 8 核 Intel® Xeon™ 时的黑白页面日处理量（单位为千页）

CPU 核数/处理器型号	Intel Xeon E5-2680 v4 2.4 GHz	Intel Xeon E5-2660 V2 2.6 GHz	Intel Xeon E5-2697A v4 2.6 GHz	Intel Xeon E5520 2.27 GHz	Intel Xeon E5-2640 v4 2.4 GHz
2	88	98	106	65	134
4	175	152	186	149	210
8	348	237	358	不适用	不适用

测试系统仅使用一个处理站，因此不能忽视网络、文件存储器、数据库和 ABBYY FineReader Server 负载的影响。增加多个处理站时会使性能出现非线性提升。

2 CPU 数量

系统可以将作业分配给多个处理站，让主机的多个 CPU 内核或网络中的任何其他计算机参与处理。请遵循以下建议：

- 如果计算结果显示需要 50 个以上 CPU 内核，请务必测试您的系统以找出瓶颈。
- 如果计算结果显示需要 50 至 100 个 CPU 内核，请将此估值增加 20%。
- 如果计算结果显示需要 100 个以上 CPU 内核，请务必进行系统负载测试。

3 磁盘速度

- 大规模处理作业需要在输入、输出和临时文件夹之间复制大量数据。磁盘读/写速度越快，文件复制速度就越快。
- 建议在托管服务器管理器并存储输入和输出文件的计算机中使用 SSD。
- 如果使用 HDD，则应考虑使用单独的硬盘执行导入和导出操作。

4 网络速度

- 如果服务器管理器连接有大量处理站，低带宽条件可能会延缓各站之间的作业分配。
- 与所有组件都安装在一台计算机中的系统相比，使用远程处理站可能会导致性能降低 10-20%。

5 输入文件类型

文件可能采用不同的格式（例如图像格式、PDF 或常用办公应用程序所采用的格式）。下表说明了输入文件格式对性能的影响。

表 2.使用 16 核 CPU Intel(R) Core(TM) I5-2400 CPU @3.10GHz 时的黑白页面日处理量（单位为千页）。

导出/导入	黑白 TIFF（100 页文档）	彩色 TIFF（单页文档）	彩色 PDF（5 页文档）
PDF	640	263	318
PDF/A	608	256	315

6 文档页数

ABBYY FineReader Server 能够轻松处理 25-250 页文档。处理包含 1,000 页以上的文档时可能会显著拖慢系统速度。

7 图像属性（质量、颜色、分辨率）

- 图像质量会影响 OCR 的速度和质量，质量较高的清晰文本能够更加快速和准确地予以识别。
- 彩色图像的处理时间比黑白或灰度图像更长。
- 高分辨率（即 600 dpi 以上）图像的加载时间长于标准 300-dpi 图像。

8 页面设计（字体和字号、文本和图片布局等）

复杂布局的处理时间更长，且 OCR 质量可能更低。

9 图像大小

处理 A3 以上尺寸的图像时，性能可能呈非线性降低。

处理参数

1 工作流程

- 如果文档存储在本地，使用热文件夹或 DocLibrary 工作流程时，将能够达到最大速度。
- 如果使用 DocLibrary 工作流程，ABBYY FineReader Server 将需要爬取文件存储以访问需要处理的文件。爬取为单线程流程，如果在此阶段（例如处理 SharePoint 库或远程文件存储时）发生任何延迟，将导致服务器在程序查找文件期间变为空闲状态。
- 如果使用电子邮件工作流程，单线程 POP3 和 Exchange 邮件(MAPI) 协议的处理速度将低于支持多线程的 IMAP。但请注意，IMAP 可能会限制服务器连接数量。
- 要加快电子邮件工作流程，请避免在邮箱中存储大量的邮件。首次运行电子邮件工作流程时，系统将加载邮箱中存储的所有消息。下次运行该工作流程时，ABBYY FineReader Server 将仅爬取邮箱以查找新消息。

2 将文档拆分为可并行处理的片段

- 默认情况下，多页文档将拆分为每个片段 25 页。您可以减小默认大小，将文档最少拆分为一个片段一页。这可能会使每项作业的处理速度提高 30-50%，但需要增加 CPU 核数才能维持同一负载。
- 处理大型文档时，您可以将每个片段的大小增加至 100-500 页甚至整个文档。这将会降低网络负载，适用于“常规”A4 文档，但在处理大型技术图纸或心电图等复杂文档时程序容易出错。

3 文档队列

默认队列大小为 50 项作业，但在某些情况下可能需要增加此值，例如，使用自定义验证/索引工作流程，或者系统中的 CPU 核数超过 25。一般而言，队列大小应为系统中 CPU 核数的两倍。

4 图像预处理参数

根据系统中的输入文件，额外的图像预处理阶段可能提高或降低性能。您将需要执行更多的测试来找出最佳图像预处理参数。

5 OCR 模式

选择质量模式时，系统性能将比快速模式降低约 45%，但会提高具有复杂布局和设计的文档的 OCR 准确性。

表 3.使用 16 核 CPU Intel(R) Core(TM) I5-2400 CPU @3.10GHz 时的日处理量（单位为千页）。

导出/导入	欧洲语言，快速模式	欧洲语言，质量模式	CJK 语言，快速模式	CJK 语言，质量模式
PDF	410	225	280	156

6 OCR 语言

- CJK 和阿拉伯语文档的处理时间比欧洲语言文档多出 30% 左右（详情请参见表 4）。

表 4.使用 16 核 CPU Intel(R) Core(TM) I5-2400 CPU @3.10GHz 和快速模式时的日处理量（单位为千页）。

导出/导入	欧洲语言，快速模式	欧洲语言，质量模式	CJK 语言，快速模式	CJK 语言，质量模式
PDF	236	272	332	286

- 如果选择的 OCR 语言不正确，处理时间可能会达到两倍。您将在作业状态中看到一条警告，表示语言选择错误。

7 输出格式

导出为 PDF/A 格式的速度比导出为 PDF 低 3% 左右。

8 打开办公格式文件时所使用的程序（内置或外部）

- 可使用外部处理器进行并行处理，这可能显著改进某些文档的性能。
- 处理数字原生文档时，应避免在单个处理站中运行 5 个以上进程，因为这可能导致队列溢出和某些作业超时。

9 开发人员日志

启用日志可能会将性能降低达 15%。

多处理站与采用多核 CPU 的单处理站比较

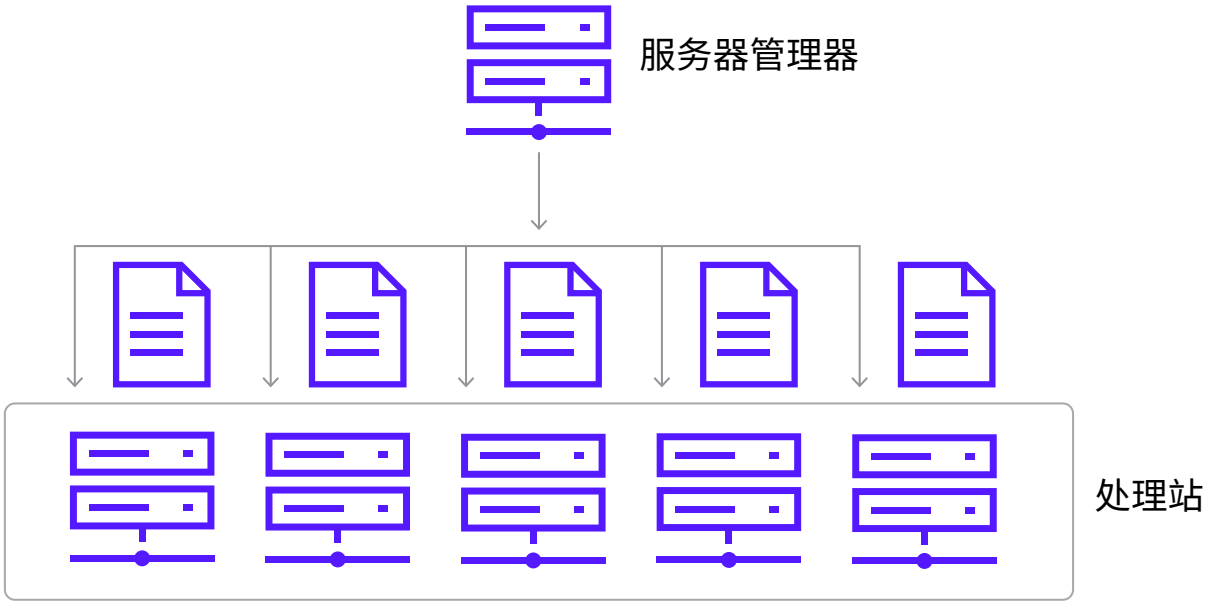
FineReader Server 由多个组件组成，其中最重要的是服务器管理器和处理站。

服务器管理器是中心组件，负责协调所有其他组件的运行，控制处理参数，创建和管理作业队列，在处理站之间分配作业，并将输出文档移至预期目标位置。

处理站是负责 OCR 和文档转换的组件。处理站从服务器管理器接收作业。

服务器管理器可以控制许多处理站。它可以在可用 CPU 内核之间平均分配工作负载，或为处理站分配指定的工作负载，以确保系统满负荷运行。该架构可提高 FineReader Server 的扩展性，使其能够处理大量的文档。FineReader Server 完全支持多核、多 CPU 和超线程技术。可同时使用的 CPU 核数和处理站数量仅受您的许可证参数的限制。

单个服务器管理器可连接的处理站数量没有任何技术限制。FineReader Server 已成功通过单个服务器管理器与超过 100 个处理站的连接测试。处理速度与所有处理站中的可用 CPU 内核总数具有直接关系。例如，在相同时间内，10 个单核 CPU 处理站的文件处理速度是一个类似处理站的 10 倍。



FineReader Server 采用同一算法在 CPU 内核（无论是用于支持单个处理站还是多个不同的处理站）之间分配作业。从这个角度讲，单个多核处理站与采用单核 CPU 的多处理站没有任何区别。但是，如果需要选择一个多核 CPU 或多个两核或四核 CPU，请注意以下几点：

1. 通常情况下，一个多核 CPU 的处理速度要比相应数量的单核 CPU 更低。例如，一个八核 CPU 的性能略低于八个单核 CPU。
2. 不建议使用 32 个以上的逻辑 CPU 内核，因为多个运行进程将会争夺访问硬盘和 CPU 缓存。
3. 处理站不能使用 32 个以上的 CPU 内核。
4. 如果发生硬件故障，处理站将无法运行，同时 FineReader Server 将无法使用所有 CPU 内核，直至处理站恢复正常运行。另一方面，所有多处理站同时发生故障的几率非常低。

建议

1 开发和集成架构建议。

要使用 FineReader Server COM/Web API 开发应用程序，您将需要一台至少具有 4 核 CPU、6 GB RAM 和 7,200 RPM 硬盘的计算机。

如果计算机满足以上要求，您就能够在其中安装服务器管理器、处理站、COM API、Web API 以及任何其他所需组件。

2 负载评估建议。

- 如需提高负载评估的准确性，请在合理范围内尽可能多次试运行工作流程。
- 使用工作流程中将要使用的 CPU、RAM 模块和 HDD 来测试一台或多台计算机。测试单个分布式系统时需要使用两台计算机，其中文档将在远程处理站进行处理。
- 如果预计负载较大，请在包含数千项作业的大型队列以及所有处理站满负荷连续运行时测试您的工作流程。
- 请检查局域网的吞吐量，确保可以处理工作流程中生成的流量。建议带宽为 1 Gbit/s。
- 对于实时处理，请根据峰值负载（而非平均负载）计算 CPU 核数。

3 如何加快不同大小文档的处理速度？

如果需要处理许多大型多页文件，单个此类文件的处理可能会极大延迟所有其他文件的处理。要避免这种情况，应根据以下公式设置 PagesSlice 参数：PagesSlice = 2 * X * T / N，其中

- X 是每分钟允许的多页文件最大数量，
- T 是采用单线程时单个此类文件的平均识别时间，
- N 是总进程数。

4 如何加快办公文件（DOCX、XLSX、PPTX 等）的处理速度？

- 使用外部组件（例如 Microsoft Office 或 LibreOffice）来充分利用多线程代码的优势。
- 创建专用的办公文档工作流程，并为该工作流程设置 PagesSlice=0。

5 如何改进容错能力？

- 请考虑使用 Microsoft Cluster Server。
- 或者考虑：
  - 将共享数据存储在外部 RAID 磁盘或云端。如果服务器出现硬件故障，您就能够将故障服务器更换为其他服务器（为最大限度减少延迟，可提前准备好备用服务器）
  - 为处理站创建定期备份

6 FineReader Server 所支持的 CPU 核数是否存在限制？

- FineReader Server 已在采用 100 核 CPU 的处理站中成功通过测试，而不会对服务器管理器的性能造成任何降级。
- 如果需要，您可以创建多服务器系统。您将能够从同一远程管理控制台控制多个服务器管理器。

7 如果发生系统故障，是否需要重新扫描或重新加载任何未完成的作业？

- 发生服务器管理器故障时：
  - 如果未丢失任何操作数据（参见“使用集群”），新的服务器管理器将接管故障服务器剩下的作业。只需要重新扫描或重新加载发生故障时正在进行的作业。
- 发生处理站故障时：
  - 故障处理站中正在处理的作业将自动重定向至第一个可用的处理站（FineReader Server 会每秒轮询一次处理站的可用性）。

8 输入文件是否存在任何特定要求（例如最大大小、每个文件的最大页数等）？

- 最大图像大小为 32,000 * 32,000 像素。
- 每个文件中的页数没有任何限制。
- 每个文件的大小也没有任何限制。