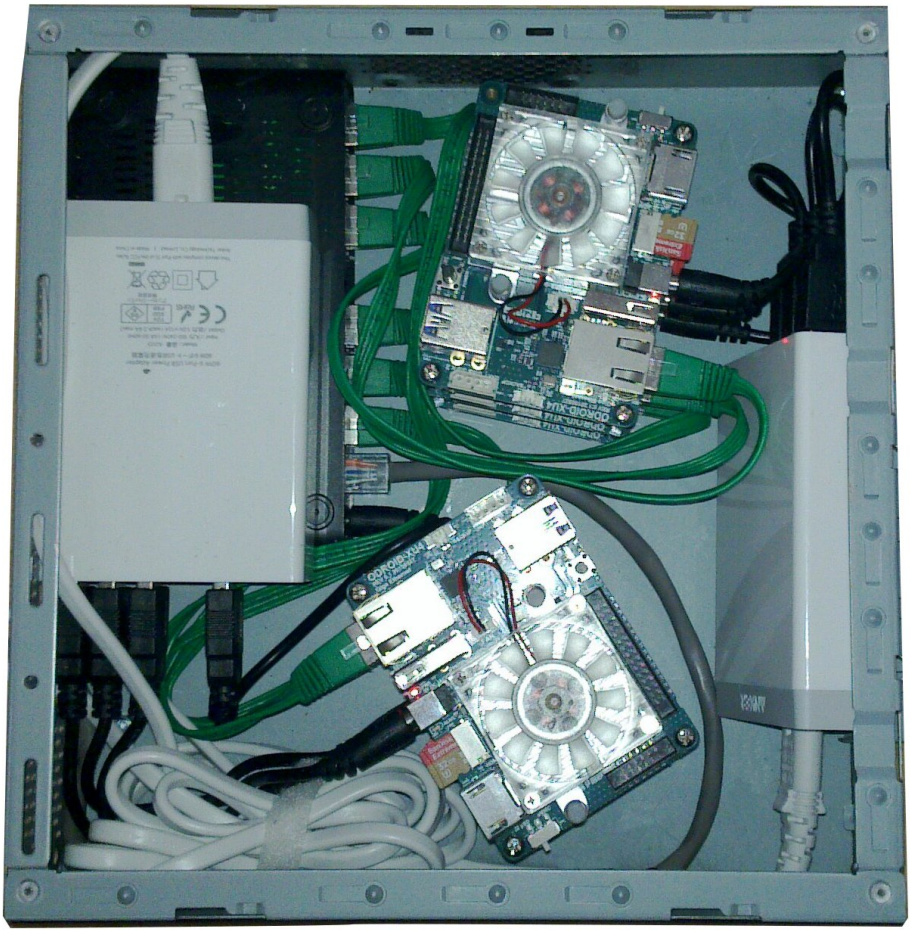
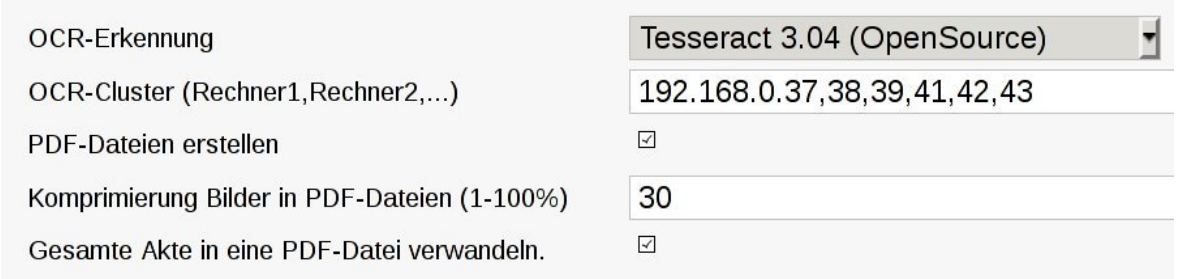
Bis zu 10 Millionen/Seiten Texterkennung pro Tag mit dem ArchivistaBox OCR-Cluster

Egg, 20. November 2015: *Mit dem ArchivistaBox OCR-Cluster (Rechnerverbund) können Bilddaten vollautomatisiert mit Texterkennung (OCR) in durchsuchbare PDF- bzw. Text-Dateien umgewandelt werden. Dank skalierbarer Cluster-Technologie von 24 bis 1920 Prozessoren (CPU-Kerne) ist der ArchivistaBox OCR-Cluster in der Lage, zwischen 120'000 und 10 Millionen Bild-Dateien pro Tag in durchsuchbare Textdaten (OCR) umzuwandeln.*



Der OCR-Cluster wird durch stromsparende ARM-Prozessoren (CPUs) angetrieben. So findet ein **48-CPU-Cluster Platz in einem 3-Liter mITX-Gehäuse und benötigt unter Last in etwa 75 Watt** an Energie. Dabei werden pro Minute 180 Seiten verarbeitet. Dies ergibt eine **Tagesleistung von 250'000 Seiten.** Die Verwaltung des OCR-Clusters erfolgt webbasiert. Bei der Auslieferung sind die notwendigen IP-Adressen der Knoten bereits eingetragen, die weitere Konfiguration wie gewünschte Sprachen, Textlayout, Scan-Profile und Netzlaufwerke werden ebenfalls per Web-Interface vorgenommen.



Um die Erkennung zu steuern, steht optional ein API (Application Programming Interface) mit HTTP-Aufrufen zur Verfügung. Ferner kann die Texterkennung direkt auf der Konsole gestartet und überwacht werden. Die zu verarbeitenden Dokumente können per FTP (Datei-Upload), SMB (Netzlaufwerk), HTTP bzw. HTTPS (Web) oder mittels angeschlossener Dokumenten-Scanner zur Verarbeitung herangezogen werden.
Bei der Texterkennung, die auf Tesseract 3.0x basiert, stehen mehr als **50 Sprachen zur Verfügung, darunter alte Zeichensätze wie Fraktur und/oder Gothik.** Zusätzliche Sprachen und/oder spezielle Zeichensätze lassen sich jederzeit integrieren. Die Auslieferung der erkannten Texte erfolgt über das integrierte Dokumenten-Management-System ArchivistaDMS. Optional können durchsuchbare PDF-Dateien direkt auf externe Laufwerke exportiert werden.



Ausgeliefert werden die OCR-Cluster in Form von Mini-Rechnern (je ca. 100 Gramm schwer) oder (optional) montiert in klassichen Gehäusen bis hin zur Rack-Bauweise. Die Preisstruktur des OCR-Clusters richtet sich nach der Anzahl CPU-Kerne. Ein einzelner Knoten enthält acht CPU (Prozessoren) und entspricht einer ArchivistaBox mit dem gewünschten Leistungsumfang. So kostet z.B. ein **OCR-Cluster mit 24 CPU-Kernen und einer Tagesleistung von 120'000 Seiten 981,18 EURO (3 x ArchivistaBox Dolder).** Die für den OCR-Cluster notwendigen Knoten (ArchivistaBoxen) können unter **shop.archivista.ch** bestellt werden.
Hinweis: *Der ArchivistaBox OCR-Cluster wurde am 21.11.2015 anlässlich des linuxday.at-Vortrages '**ARM-Plattform reif für den Alltag?**' der Öffentlichkeit vorgestellt.*