ABBYY Recognition Server 4.0 OCR on Amazon AWS

Amazon EC2 testing environment is provided by WiseTREND (www.wisetrend.com)
Created on 07/29/15

Table of Contents

OVERVIEW 2
ENVIRONEMENT 2
COST 3
TECHNICAL CHARACTERISTICS 4
PassMark Performance Test Score 4
Hardware Details as reported by PassMark Performance Test 5
PassMark Performance Test Hardware Comparison 7
Internet Speed Test 9
Main Drive Storage 10
SOFTWARE 10
TEST IMAGE SET 11
PROCESSING SETTINGS 11
PROCESSING STATISTICS 12
CONCLUSION 12
CONTACT & ADDITIONAL TESTING DETAILS 12

OVERVIEW

The purpose of this Amazon AWS environment is to benchmark the Cloud-based implementation of ABBYY Recognition Server 4.0 build 4.0.4.1425.

ENVIRONMENT

This test environment can be easily scaled on servers of various size from small (4 cores) to very large (36 cores) implementation.

From Amazon documentation:

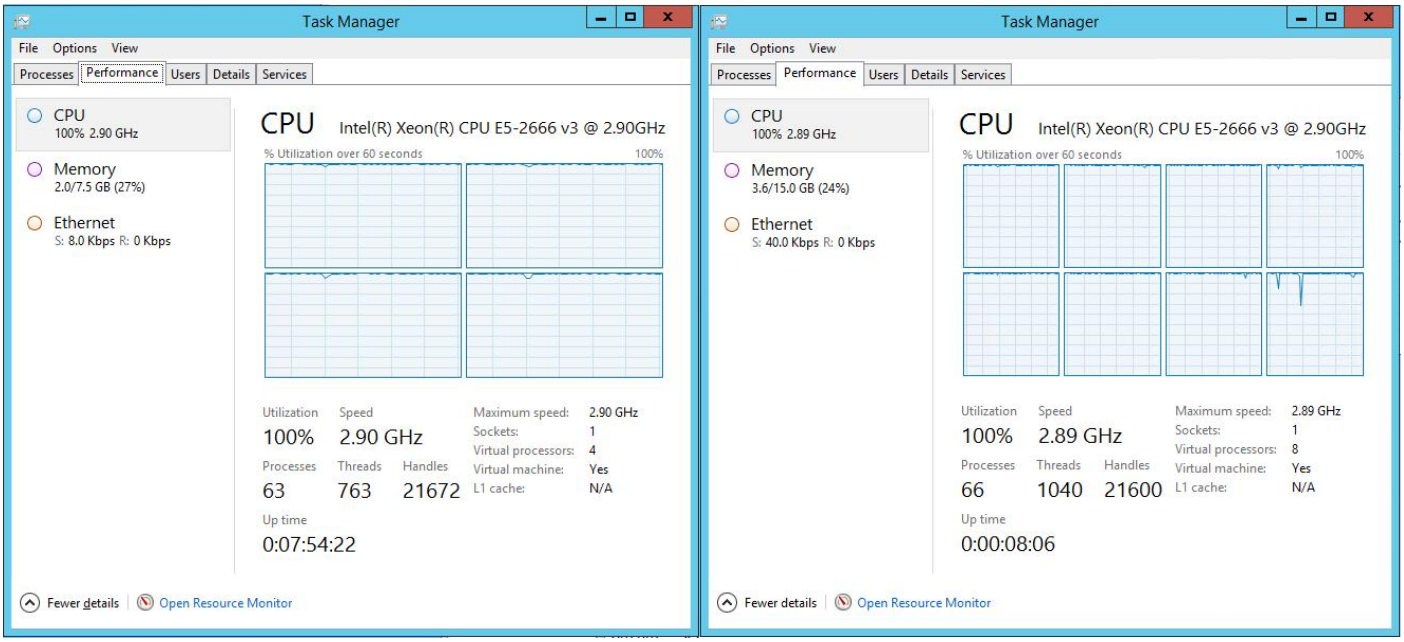
C4 instances are the latest generation of Compute-optimized instances, featuring the highest performing processors and the lowest price/compute performance in EC2.

Features:

- High frequency Intel Xeon E5-2666 v3 (Haswell) processors optimized specifically for EC2
- EBS-optimized by default and at no additional cost
- Ability to control processor C-state and P-state configuration on the c4.8xlarge instance type
- Support for Enhanced Networking and Clustering

Model	vCPU	Mem (GiB)	Storage	Dedicated EBS Throughput (Mbps)
c4.large	2	3.75	EBS-Only	500
c4.xlarge	4	7.5	EBS-Only	750
c4.2xlarge	8	15	EBS-Only	1,000
c4.4xlarge	16	30	EBS-Only	2,000
c4.8xlarge	36	60	EBS-Only	4,000

This particular test has been run on c4.xlarge (4 core, 7.5 GB RAM) and c4.2xlarge (8 core, 15 GB RAM) instance types in N.Virginia zone. 40 GB magnetic virtual hard drive was used for testing.

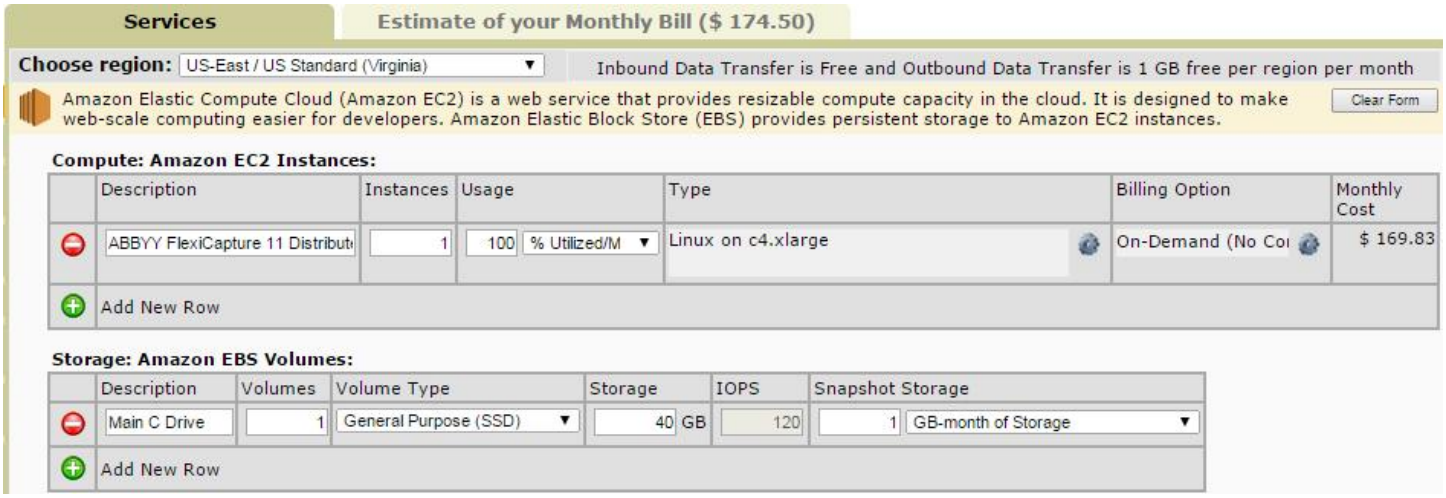


(shown at full CPU load during processing)

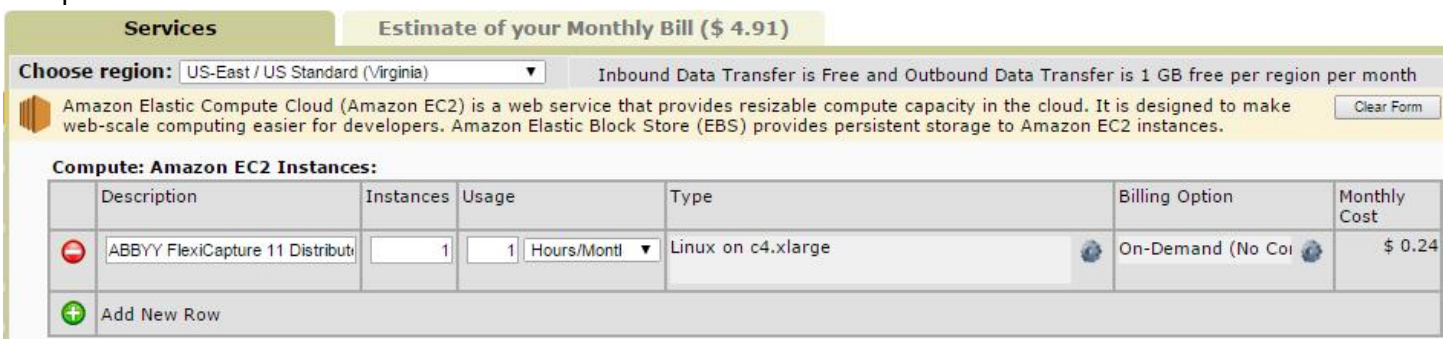
More information about various Amazon EC2 instance types is available here:
http://aws.amazon.com/ec2/instance-types/

COST

The estimated monthly cost for running c4.xlarge instance type in us-east-a1 availability zone as of 03/25/2015 is:



Cost per hour:



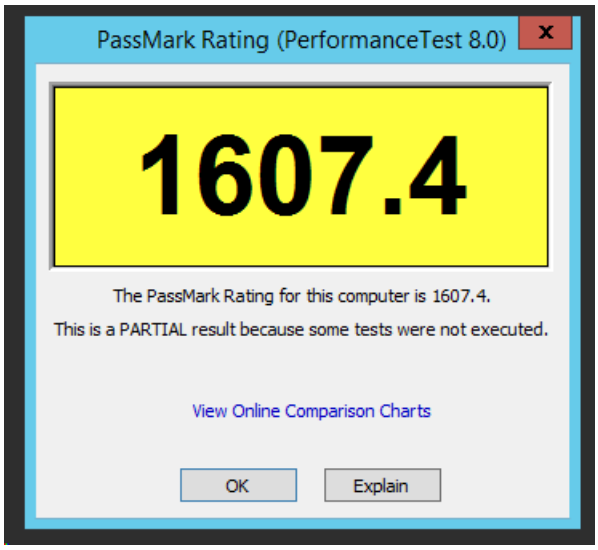
NOTE: Cost may change at any time per Amazon AWS price list. Other minor associated costs may apply, such as for additional data transfer in/out, Elastic IP, monitoring, and other services.

NOTE: This calculation does NOT include the cost of ABBYY Recognition Server 4.0 licensing and setup.

TECHNICAL CHARACTERISTICS

This c4.xlarge virtual machine instance type exhibited the following performance characteristics:

PassMark Performance Test Score



c4.xlarge

c4.2xlarge

Hardware Details as reported by PassMark Performance Test

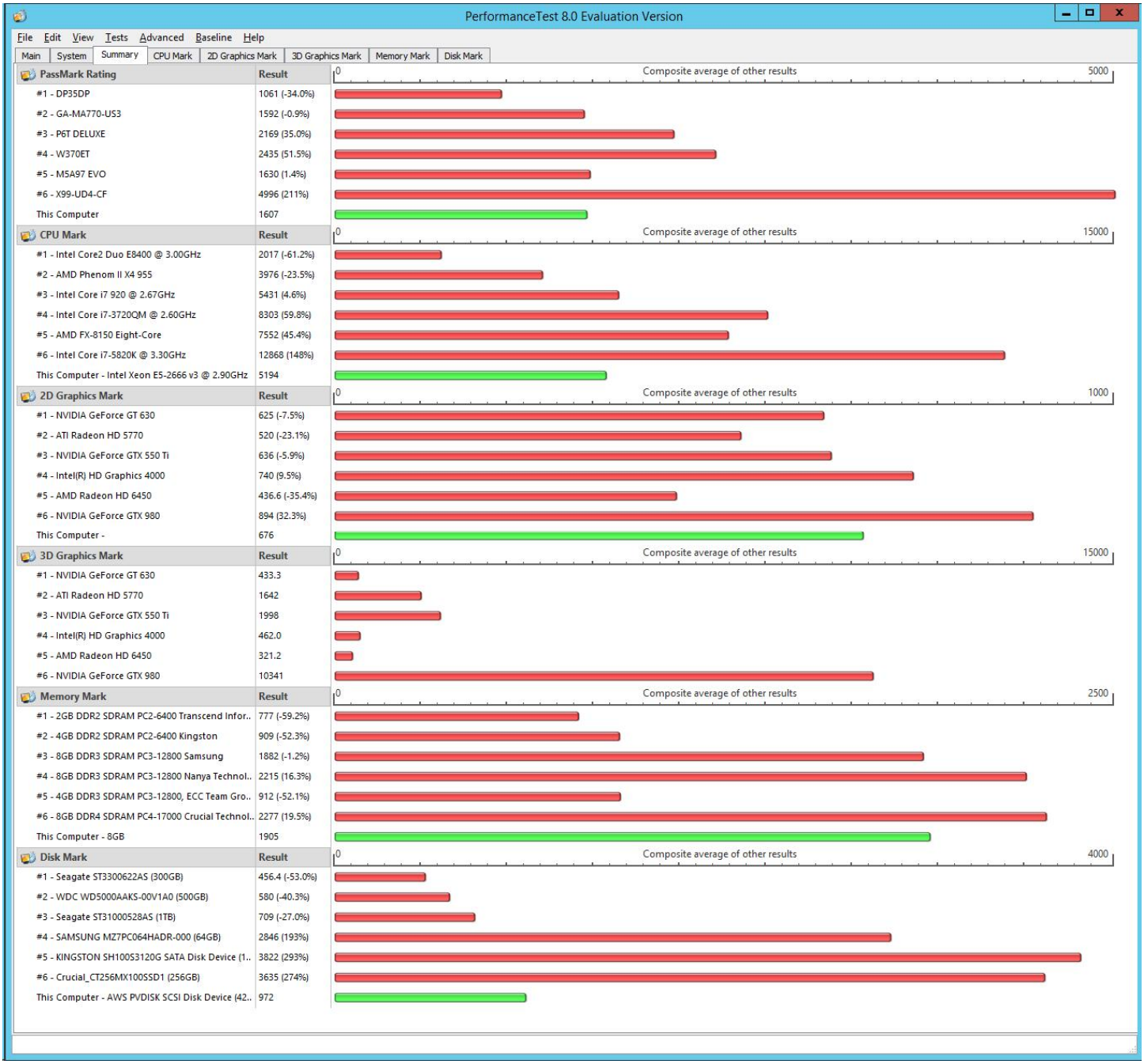
Item	This Computer
PerformanceTest Information	
PerformanceTest Version	8.0 (1046) WIN64
PassMark Rating	1607
System Information	
System Name	WIN-MCPGC64ISOF
Model	
Operating System	Windows Server 2012 R2 Server Standard Edition (full installation) build 9600 (64-bit)
Motherboard Manufacturer	
Motherboard Model	
Motherboard Version	
BIOS Manufacturer	Xen
BIOS Version	Xen - 0
BIOS Release Date	2014/12/04
CPU Information	
Manufacturer	GenuineIntel
Type	Intel Xeon E5-2666 v3 @ 2.90GHz
Codename	Haswell-E
CPUID	Family 6, Model 3F, Stepping 2
Socket	LGA2011-v3
Lithography	22nm
Number of CPU's	1
Cores per CPU	2
Logicals per Core	2
Clock Frequencies	
Measured Speed	2893.5 MHz
Multiplier	(N/A)
Bus Speed	99.8 MHz
Front Side Bus Speed	(N/A)
Timing Error Ratio	1.000
Cache per CPU package	
L1 Instruction Cache	2 x 32 KB
L1 Data Cache	2 x 32 KB
L2 Cache Size	2 x 256 KB
L3 Cache	25 MB
Memory Information	
Total Physical Memory	7680MB
Available Physical Memory	6386MB
Memory Devices	
Slot 1	RAM, 7680MB,

c4.xlarge

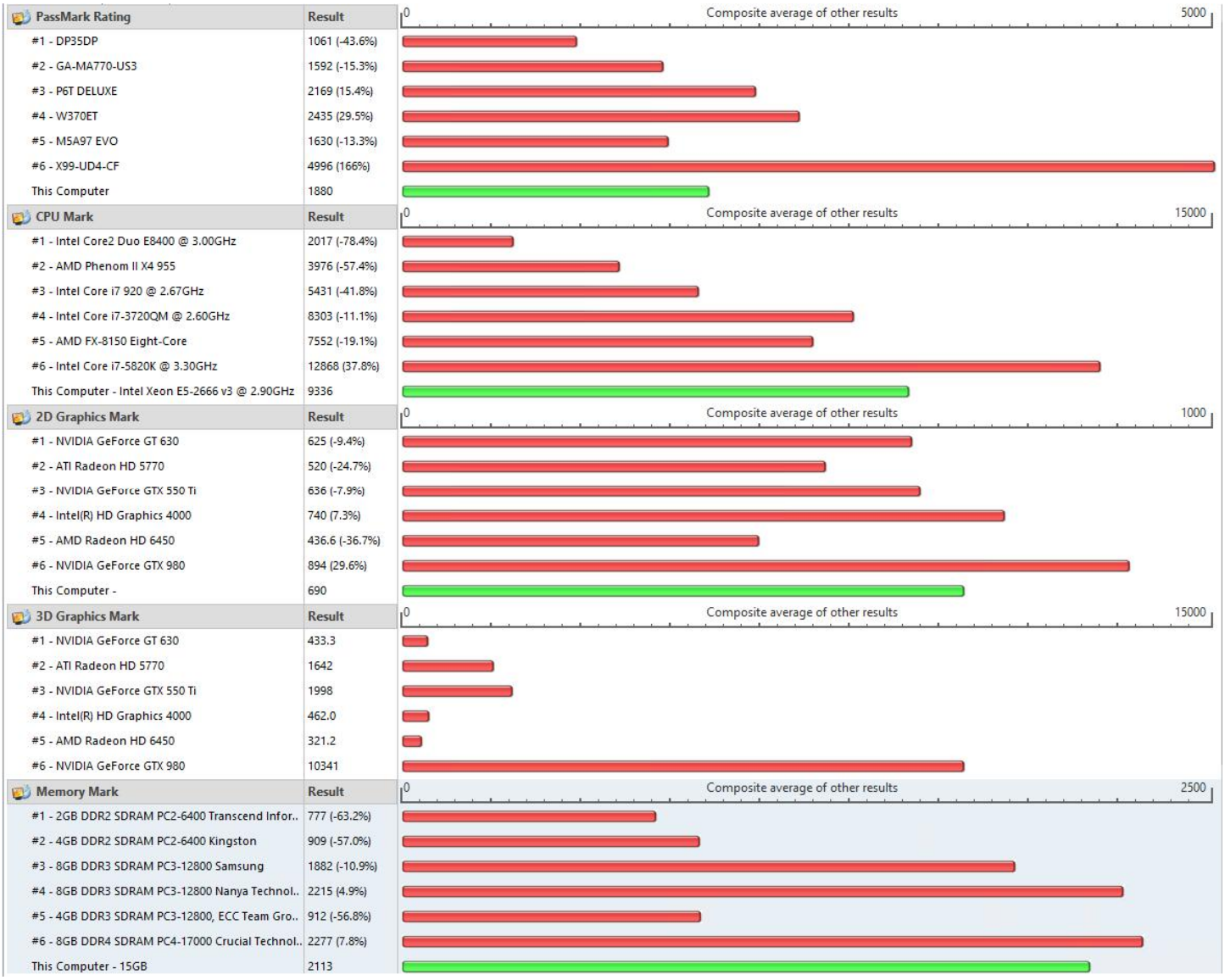
Item	This Computer
PerformanceTest Information	
PerformanceTest Version	8.0 (1046) WIN64
PassMark Rating	1880
System Information	
System Name	WIN-FC11DISTR
Model	
Operating System	Windows Server 2012 R2 Server Standard Edition (full installation) build 9600 (64-bit)
Motherboard Manufacturer	
Motherboard Model	
Motherboard Version	
BIOS Manufacturer	Xen
BIOS Version	Xen - 0
BIOS Release Date	2015/05/06
CPU Information	
Manufacturer	GenuineIntel
Type	Intel Xeon E5-2666 v3 @ 2.90GHz
Codename	Haswell-E
CPUID	Family 6, Model 3F, Stepping 2
Socket	LGA2011-v3
Lithography	22nm
Number of CPU's	1
Cores per CPU	4
Logicals per Core	2
Clock Frequencies	
Measured Speed	2893.7 MHz
Multiplier	(N/A)
Bus Speed	99.8 MHz
Front Side Bus Speed	(N/A)
Timing Error Ratio	1.000
Cache per CPU package	
L1 Instruction Cache	4 x 32 KB
L1 Data Cache	4 x 32 KB
L2 Cache Size	4 x 256 KB
L3 Cache	25 MB
Memory Information	
Total Physical Memory	15360MB
Available Physical Memory	14025MB
Memory Devices	
Slot 1	RAM, 15360MB,

c4.2xlarge

PassMark Performance Test Hardware Comparison



c4.xlarge



PassMark Rating	Result
#1 - DP35DP	1061 (-43.6%)
#2 - GA-MA770-US3	1592 (-15.3%)
#3 - P6T DELUXE	2169 (15.4%)
#4 - W370ET	2435 (29.5%)
#5 - M5A97 EVO	1630 (-13.3%)
#6 - X99-UD4-CF	4996 (166%)
This Computer	1880

CPU Mark	Result
#1 - Intel Core2 Duo E8400 @ 3.00GHz	2017 (-78.4%)
#2 - AMD Phenom II X4 955	3976 (-57.4%)
#3 - Intel Core i7 920 @ 2.67GHz	5431 (-41.8%)
#4 - Intel Core i7-3720QM @ 2.60GHz	8303 (-11.1%)
#5 - AMD FX-8150 Eight-Core	7552 (-19.1%)
#6 - Intel Core i7-5820K @ 3.30GHz	12868 (37.8%)
This Computer - Intel Xeon E5-2666 v3 @ 2.90GHz	9336

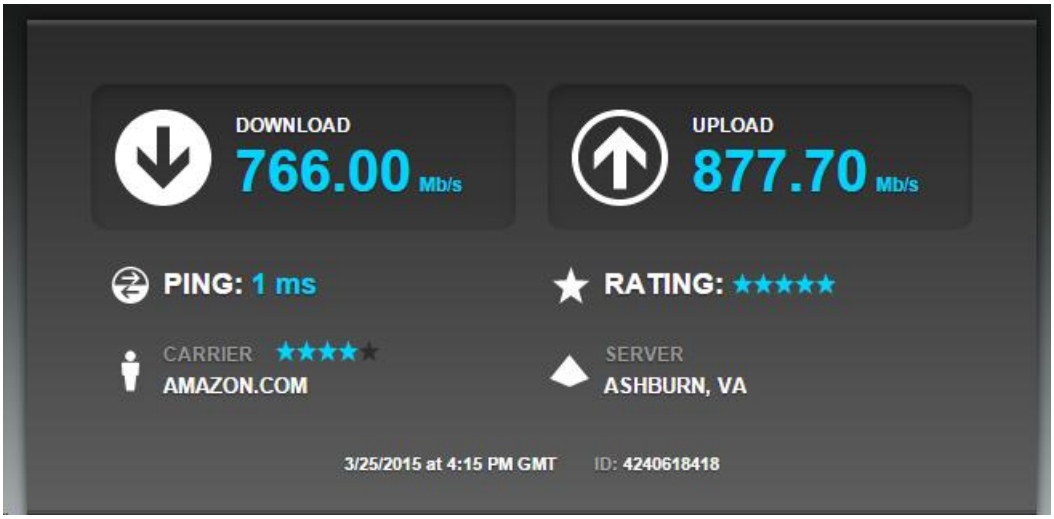
2D Graphics Mark	Result
#1 - NVIDIA GeForce GT 630	625 (-9.4%)
#2 - ATI Radeon HD 5770	520 (-24.7%)
#3 - NVIDIA GeForce GTX 550 Ti	636 (-7.9%)
#4 - Intel(R) HD Graphics 4000	740 (7.3%)
#5 - AMD Radeon HD 6450	436.6 (-36.7%)
#6 - NVIDIA GeForce GTX 980	894 (29.6%)
This Computer -	690

3D Graphics Mark	Result
#1 - NVIDIA GeForce GT 630	433.3
#2 - ATI Radeon HD 5770	1642
#3 - NVIDIA GeForce GTX 550 Ti	1998
#4 - Intel(R) HD Graphics 4000	462.0
#5 - AMD Radeon HD 6450	321.2
#6 - NVIDIA GeForce GTX 980	10341

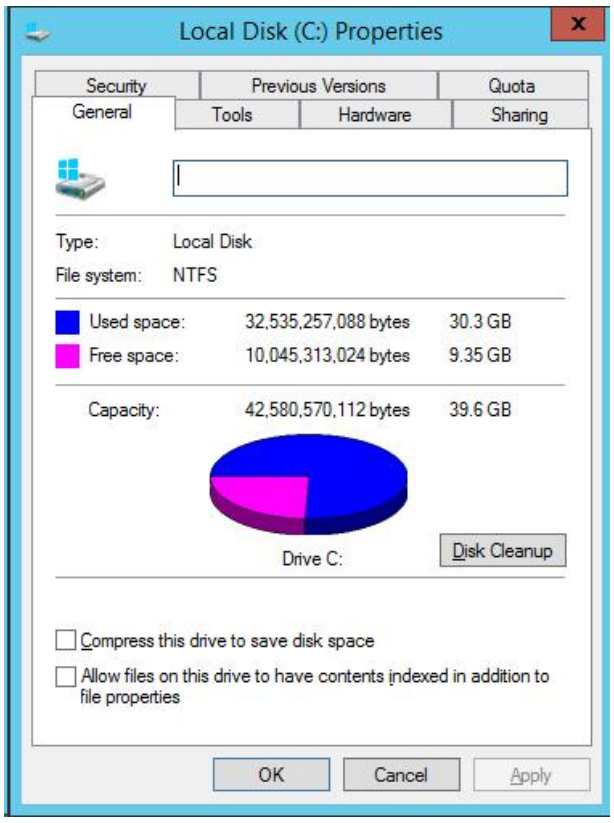
Memory Mark	Result
#1 - 2GB DDR2 SDRAM PC2-6400 Transcend Infor..	777 (-63.2%)
#2 - 4GB DDR2 SDRAM PC2-6400 Kingston	909 (-57.0%)
#3 - 8GB DDR3 SDRAM PC3-12800 Samsung	1882 (-10.9%)
#4 - 8GB DDR3 SDRAM PC3-12800 Nanya Technol..	2215 (4.9%)
#5 - 4GB DDR3 SDRAM PC3-12800, ECC Team Gro..	912 (-56.8%)
#6 - 8GB DDR4 SDRAM PC4-17000 Crucial Technol..	2277 (7.8%)
This Computer - 15GB	2113

c4.2xlarge

Internet Speed Test



Main Drive Storage



The available free storage on the bundled 40 GB hard drive is about 9 GB. This free storage will be sufficient for testing and minimal processing. For large volume processing, the bundled drive should be expanded, or a separate Amazon Elastic Storage drive should be connected for the additional data store. Expanded storage will increase the cost per month.

SOFTWARE

This tested Amazon EC2 server contains these software packages:

- Windows Server 2012 R2
- ABBYY Recognition Server 4.0 build 4.0.4.1425
- Chrome Browser
- ClipX clipboard management utility
- Performance Text 8.0 Trial for benchmarking
- Adobe Acrobat Reader XI (11)
- Microsoft SQL Server 2008 Express
- Microsoft SQL Server 2008 Express Management Studio
- Crystal Reports Basic Runtime

TEST IMAGE SET

This test is based on a specific image set, which we selected from a large pool of documents representing typical processing needs of our clients. In order to run similar benchmarking on your specific image set, please contact WiseTREND (sales@wisetrend.com).

Image set contained 10 different documents:

Image Name	Pages Qty	Description
Sample 001.PDF	2	Product return UPS label + packing slip. Digitally generated image-based produced by Amazon.com.
Sample 002.PDF	58	Mortgage document scanned on Kodak ScanStaion 100 at 200 dpi. Mixed content.
Sample 003.PDF	1	Invoice digitally generated image-based from QuickBooks.
Sample 004.PDF	4	City of Detroit utility bill. Scanned at 300 dpi, with substantial background noise.
Sample 005.PDF	2	NDA. Scanned on Kodak i1210 and previously processed with FineReader 8.0 into searchable PDF.
Sample 006.PDF	19	Financial report. Scanned on Kodak Scan Station 100 at 200 dpi.
Sample 007.PDF	5	Legal document. Color. Produced as a print screen from some other application.
Sample 008.TIF	99	Agreement with Amendments and Exhibits. Includes handwriting on every 5th or so page.
Sample 009.JPG	1	Mobiel camera picture of a US Letter size document. Average lighting conditions.
Sample 010.PDF	5	NDA. Scanned on Kodak i1210 and previously processed with FineReader 8.0 into searchable PDF.

This set was duplicated 30 times to produce a total of 300 documents for load testing.

Total pages in this set: 5880

All pages in this set are US Letter size.

PROCESSING SETTINGS

OCR Language: English
High-Quality OCR Mode
Auto-rotation correction: Enabled
Extract barcodes
Extract text from pictures
Special processing for technical drawings
Correct resolution of photos
Clear background noise
Deskew
Despeckle
Output: TXT, Searchable PDF (text under image, 70% image layer compression)

PROCESSING STATISTICS

Test No	Server	Cores Qty Used	Run Time	Avg Pages Per Min	Avg Pages Per Min Per Core
1	c4.xlarge	3 (max-1)	95 min	62	21
2	c4.xlarge	4 (max)	82 min	72	18
3	c4.2xlarge	7 (max-1)	39.5 min	149	21
4	c4.2xlarge	8 (max)	37 min	159	20

Highlighted fields indicate a more degraded performance per page, even though the entire files set has been processed faster. In general, it is recommended not to max out the server (computer executing management of the workflow) by allocating all cores to OCR, which may starve the server-related processes. This effect will likely be more noticeable as more cores are controlled by the same server. A rule of thumb is to use 'n-1' formula on the server performing both workflow management and OCR processes. This rule allows at least one available core for other non-OCR processing tasks.

CONCLUSION

The c4.xlarge instance is capable of processing 1 Million pages in 9.7 days of 24-hour operation.

The c4.2xlarge instance is capable of processing 1 Million pages in 4.4 days of 24-hour operation.

CONTACT & ADDITIONAL TESTING DETAILS

We will be glad to run benchmarking on your specific image set to determine processing throughputs for your specific documents.

Please contact WiseTREND for any further assistance.

www.WiseTREND.com

support@wisetrend.com

(510)754-9866 USA