RESULTS FOR THE STUDY "AN EXPERIMENTAL AND COMPARATIVE BENCHMARK STUDY EXAMINING RESOURCE UTILIZATION IN MANAGED HADOOP CONTEXT"

SELECTED CONFIGURATIONS:

	GCP	**Azure**	**Alibaba Cloud**
Service	Dataproc	HDInsight	e-MapReduce
Region	europe-west3-a	Germany West Central	eu-central-1
Location	Frankfurt	Frankfurt	Frankfurt
Image	1.4-ubuntu18	HDI 3.6	EMR-3.32.0
OS	ubuntu18.04	ubuntu 16.04	Aliyun Linux 2
Hadoop	2.9	2.7.3	2.8.5
Java	1.8.0_275	1.8.0_275	1.8.0_252
Infrastructure	GCE	HDP	ECS
MASTER NODE			
Machine Type	e2-highmem-8	A8m v2	ecs.se1.2xlarge
Processors	8 vCPU	8 cores	8 vCPU
Memory	64 GB RAM	64 GB RAM	64 GB RAM
Extras	–	1 masternode for HA 3 nodes for Zookeeper	–
WORKER NODES			
# of Nodes	3	3	3
Machine Type	e2-highmem-4	A4m v2	ecs.se1.xlarge
Processors	4 vCPU	4 cores	4 vCPU
Memory	32 GB RAM	32 GB RAM	32 GB RAM
Storage Replication Block size	HDFS 1000 GB 2 128 MB	WASB Azure blob storage	HDFS 1000 GB 2 128 MB

USE CASE 1 BENCHMARK RESULTS

DATA SCALA: HUGE							
		Dataproc		**HDInsight**		**e-MapReduce**	
Benchmark	**IDS**	$D_{(s)}$	$T_{(MB/s)}$	$D_{(s)}$	$T_{(MB/s)}$	$D_{(s)}$	$T_{(MB/s)}$
Sort	3.28	70	47.11	131	25.08	111	29.42
Terasort	32.00	667	47.99	858	37.28	1054	30.37
Wordcount	32.85	978	33.60	1470	22.34	889	36.95
Dfsioe-r	26.99	294	91.77	662	40.79	245	110.21
Dfsioe-w	27.16	379	71.73	658	41.30	281	96.49
Scan	2.01	73	27.63	157	12.83	74	27.19
Join	1.92	181	10.61	356	5.39	175	10.95
Aggregation	0.37	97	3.86	215	1.73	97	3.85
Bayes	1.88	2604	0.72	6120	0.31	3017	0.62
Kmeans	20.08	2321	8.65	2313	8.68	2070	9.70
Pagerank	2.99	1544	1.94	3334	0.90	2458	1.22
DATA SCALE: GIGANTIC							
		Dataproc		**HDInsight**		**e-MapReduce**	
Benchmark	**IDS**	$D_{(s)}$	$T_{(MB/s)}$	$D_{(s)}$	$T_{(MB/s)}$	$D_{(s)}$	$T_{(MB/s)}$
Sort	32.85	715	45.94	787	41.72	896	36.68
Terasort	320.00	9821	32.58	(*)	(*)	9660	33.13
Wordcount	328.49	10131	32.42	13596	24.16	8671	37.88
Dfsioe-r	216.03	915	236.11	1886	114.54	660	327.29
Dfsioe-w	217.33	1347	161.39	1914	113.57	1060	205.12
Scan	20.10	457	43.96	514	39.09	407	49.38
Join	19.19	595	32.27	761	25.24	594	32.32
Aggregation	3.69	523	7.05	594	6.20	565	6.52
Bayes	3.77	5350	0.70	12589	0.30	6363	0.60
Kmeans	40.16	4541	8.84	4042	9.94	4034	9.96
Pagerank	19.93	8371	2.38	11779	1.70	13893	1.43
IDS: Input Data Size (GB); $D_{(s)}$: Duration (sec); $T_{(MB/s)}$; Throughput (MB/sec)							
(*) Workload failed to run within 3 attempts							

USE CASE 2 BENCHMARK RESULTS

		Dataproc		**HDInsight**		**e-MapReduce**	
Benchmark	**IDS**	$D_{(s)}$	$T_{(MB/s)}$	$D_{(s)}$	$T_{(MB/s)}$	$D_{(s)}$	$T_{(MB/s)}$
Sort (t)	39.30 KB	36	0.0012	69	0.0006	32	0.0012
Sort (s)	3.28 MB	36	0.09	70	0.0471	31	0.105
Sort (l)	328.50 MB	42	7.86	81	4.07	42	7.74
Sort (h)	3.28 GB	70	47.08	141	23.36	107	30.69
Sort (g)	32.85 GB	694	47.30	699	47.00	883	37.20
Wordcount (t)	38.65 KB	38	0.001	68	0.0006	31	0.0012
Wordcount (s)	348.29 MB	50	6.51	98	3.34	47	7.06
Wordcount (l)	3.28 GB	129	25.45	269	12.20	120	27.27
Wordcount (h)	32.85 GB	952	34.51	1487	22.10	888	37.00
Wordcount (g)	328.49 GB	9749	33.70	13286	24.73	8622	38.10
IDS: Input Data Size (GB); $D_{(s)}$: Duration (sec); $T_{(MB/s)}$; Throughput (MB/sec) (t): tiny, (s): small, (l): large, (h): huge, (g): gigantic							

USE CASE 1 SLOWDOWN ESTIMATES

DATA SCALA: HUGE					
Benchmark	**First**	**Second**	**SE**	**Third**	**SE**
Sort	GCP	Alibaba	1.59	Azure	1.87
Terasort	GCP	Azure	1.29	Alibaba	1.58
Wordcount	Alibaba	GCP	1.10	Azure	1.65
Dfsioe-r	Alibaba	GCP	1.20	Azure	2.70
Dfsioe-w	Alibaba	GCP	1.35	Azure	2.34
Scan	GCP	Alibaba	1.01	Azure	2.15
Join	Alibaba	GCP	1.03	Azure	2.03
Aggregation	GCP-Alibaba	Azure	2.22	---	---
Bayes	GCP	Alibaba	1.16	Azure	2.35
Kmeans	Alibaba	Azure-GCP	1.12	---	---
Pagerank	GCP	Alibaba	1.59	Azure	2.16
DATA SCALE: GIGANTIC					
Benchmark	**First**	**Second**	**SE**	**Third**	**SE**
Sort	GCP	Azure	1.10	Alibaba	1.12
Terasort	Alibaba	GCP	1.02	Azure	(*)
Wordcount	Alibaba	GCP	1.17	Azure	1.57
Dfsioe-r	Alibaba	GCP	1.39	Azure	2.86
Dfsioe-w	Alibaba	GCP	1.27	Azure	1.81
Scan	Alibaba	GCP	1.12	Azure	1.26
Join	Alibaba-GCP	Azure	1.28	---	---
Aggregation	GCP	Alibaba	1.08	Azure	1.14
Bayes	GCP	Alibaba	1.19	Azure	2.35
Kmeans	Alibaba-Azure	GCP	1.13	---	---
Pagerank	GCP	Azure	1.41	Alibaba	1.66
(*) Workload failed to run within 3 attempts					

USE CASE 2 SLOWDOWN ESTIMATES

DATA SCALA: HUGE					
Benchmark	**First**	**Second**	**SE**	**Third**	**SE**
Sort (t)	Alibaba	GCP	1.13	Azure	2.16
Sort (s)	Alibaba	GCP	1.16	Azure	2.26
Sort (l)	GCP-Alibaba	Azure	1.93	---	---
Sort (h)	GCP	Alibaba	1.53	Azure	2.01
Sort (g)	GCP	Azure	1.01	Alibaba	1.27
Wordcount (t)	Alibaba	GCP	1.23	Azure	2.19
Wordcount (s)	Alibaba	GCP	1.06	Azure	2.09
Wordcount (l)	Alibaba	GCP	1.08	Azure	2.24
Wordcount (h)	Alibaba	GCP	1.07	Azure	1.67
Wordcount (g)	Alibaba	GCP	1.13	Azure	1.54
(t): tiny, (s): small, (l): large, (h): huge, (g): gigantic					

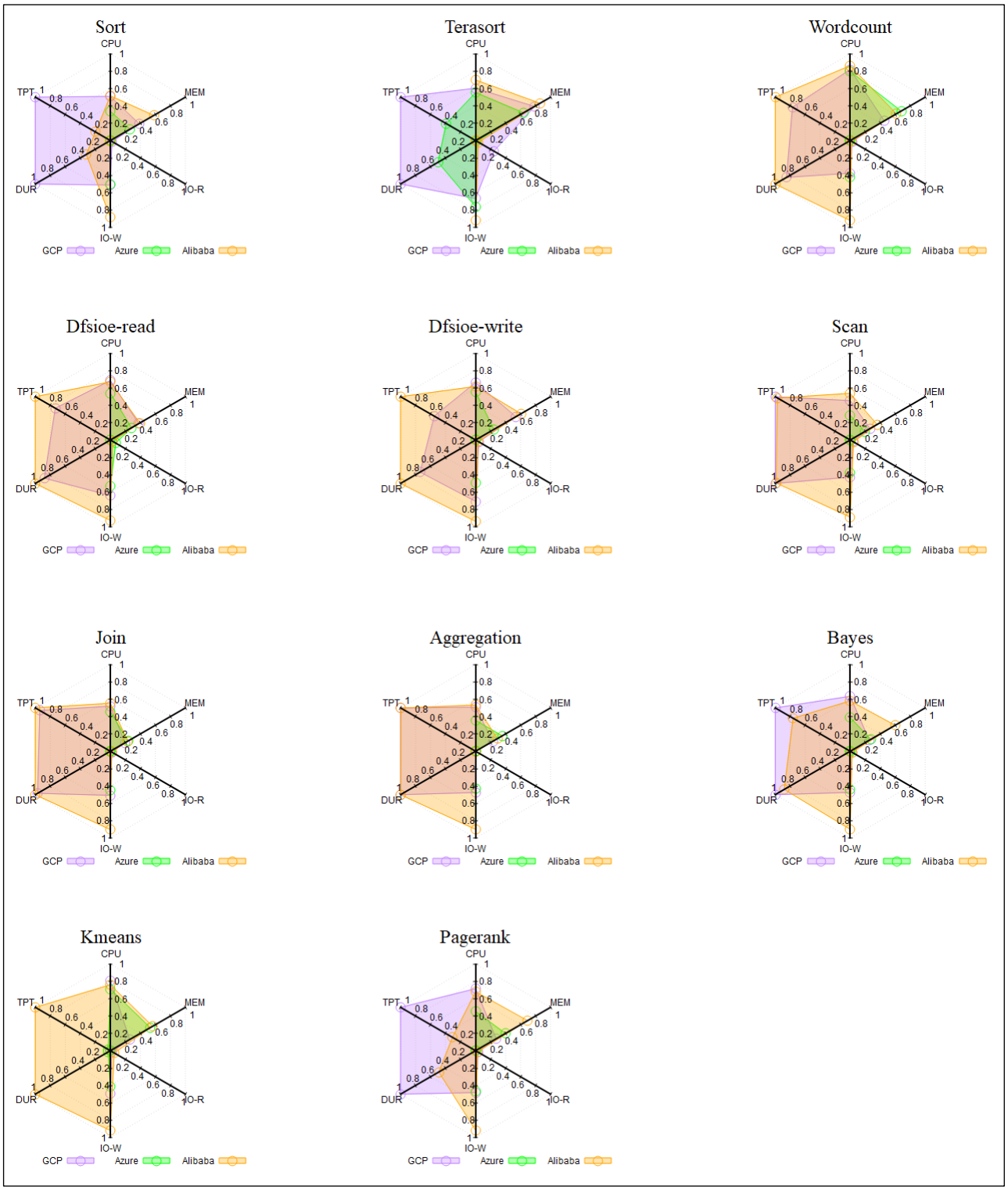
ALLOCATED MAP AND REDUCE SLOTS IN SQL BENCHMARKS OF USE CASE 1

	GCP		Azure		Alibaba	
Benchmark	**Maps**	**Reduces**	**Maps**	**Reduces**	**Maps**	**Reduces**
Scan* (h)	24	---	12	---	24	---
Scan* (g)	144	---	36	---	144	---
Join (h)	60	25	48	25	60	25
Join (g)	180	25	72	25	180	25
Aggregation (h)	24	12	12	12	24	12
Aggregation (g)	144	12	36	12	144	12
IDS: Input Data Size (GB); $D_{(s)}$: Duration (sec); $T_{(MB/s)}$; Throughput (MB/sec) (h): huge, (g): gigantic						

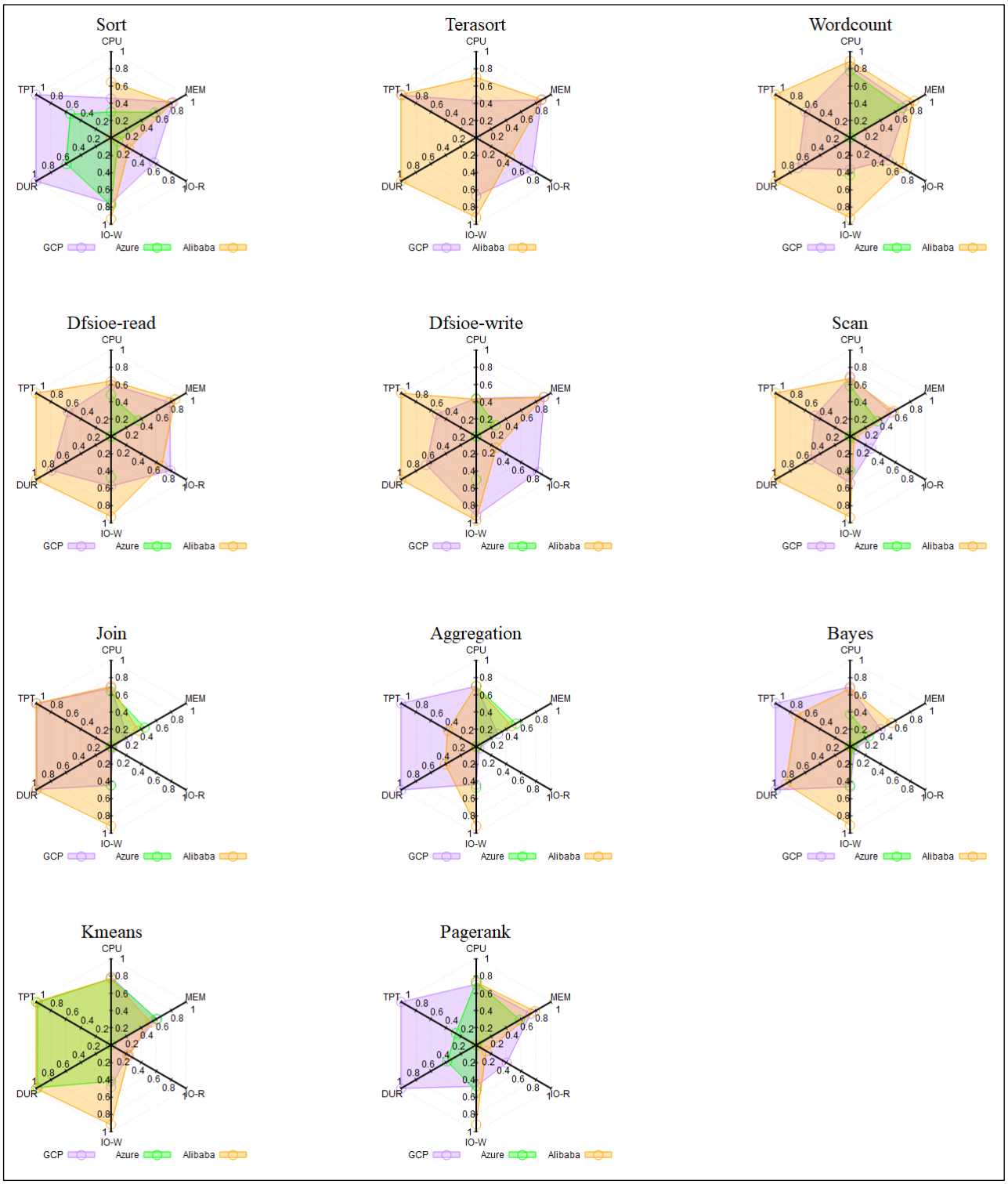
ALLOCATED MAP AND REDUCE SLOTS IN USE CASE 2

	GCP		Azure		Alibaba	
Benchmark	**Maps**	**Reduces**	**Maps**	**Reduces**	**Maps**	**Reduces**
Sort (t)	11	12	12	12	11	12
Sort (s)	11	12	12	12	11	12
Sort (l)	12	11	12	12	11	11
Sort (h)	24	12	12	12	23	12
Sort (g)	251	12	60	12	252	12
Wordcount (t)	11	12	12	12	11	12
Wordcount (s)	12	11	12	12	12	11
Wordcount (l)	24	11	12	12	23	12
Wordcount (h)	250	11	60	12	252	12
Wordcount (g)	2447	12	612	12	2447	11
(h): huge, (g): gigantic						

USE CASE 1 – AVERAGE SYSTEM UTILIZATION IN DATA SCALE HUGE

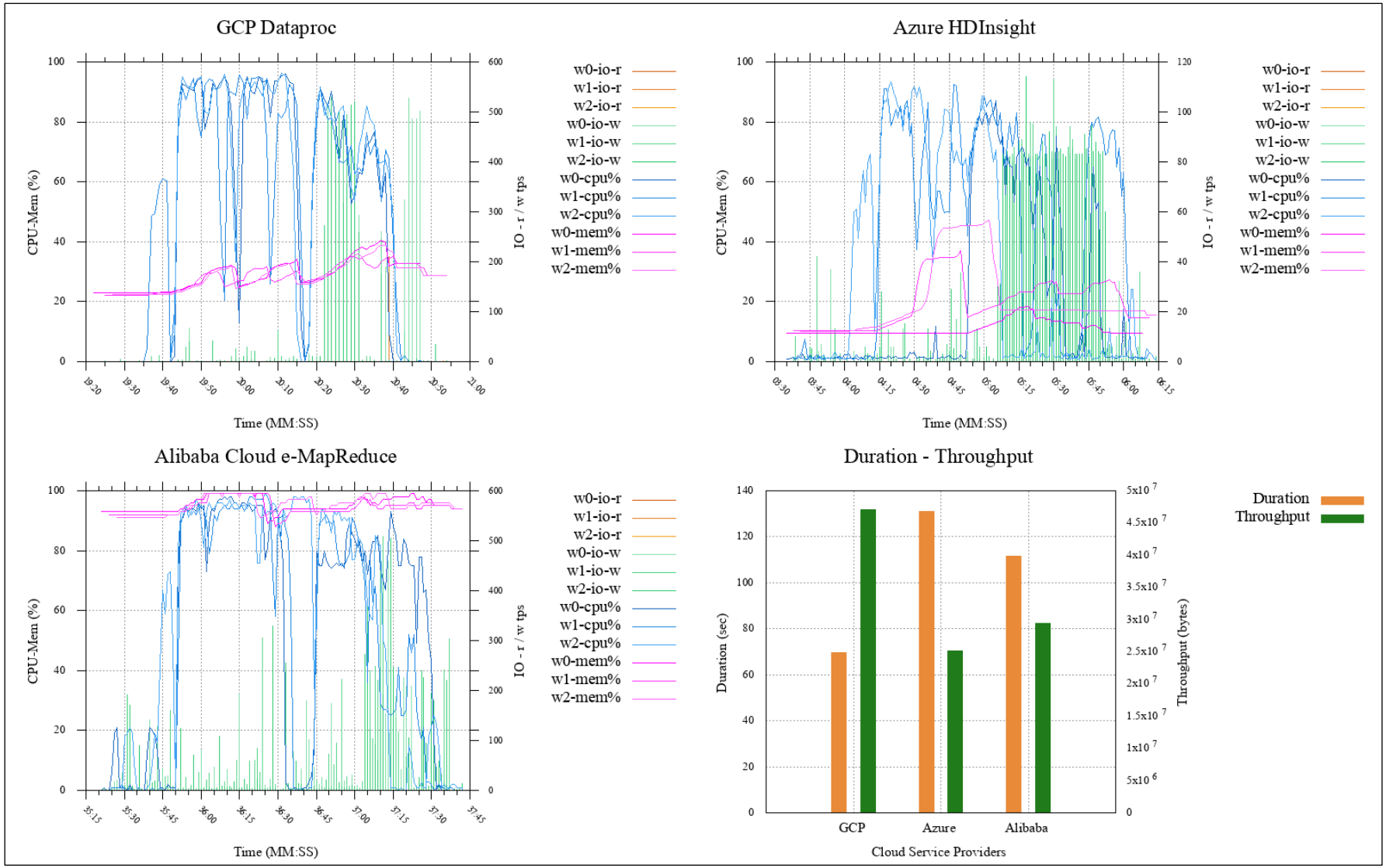


USE CASE 1 – AVERAGE SYSTEM UTILIZATION IN DATA SCALE GIGANTIC

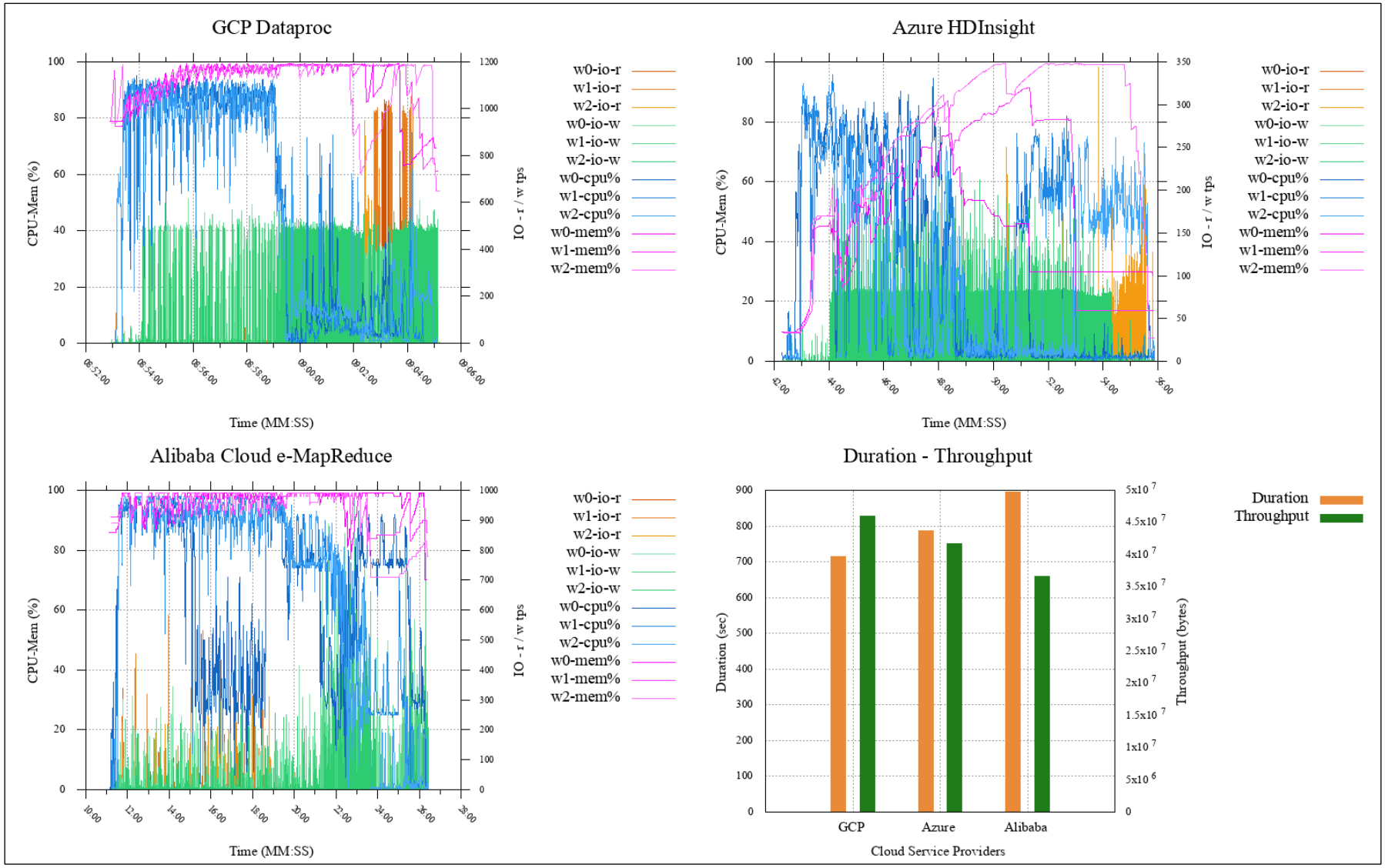


In Terasort plot, Azure results could not be visualized due to incomplete workload/benchmark.

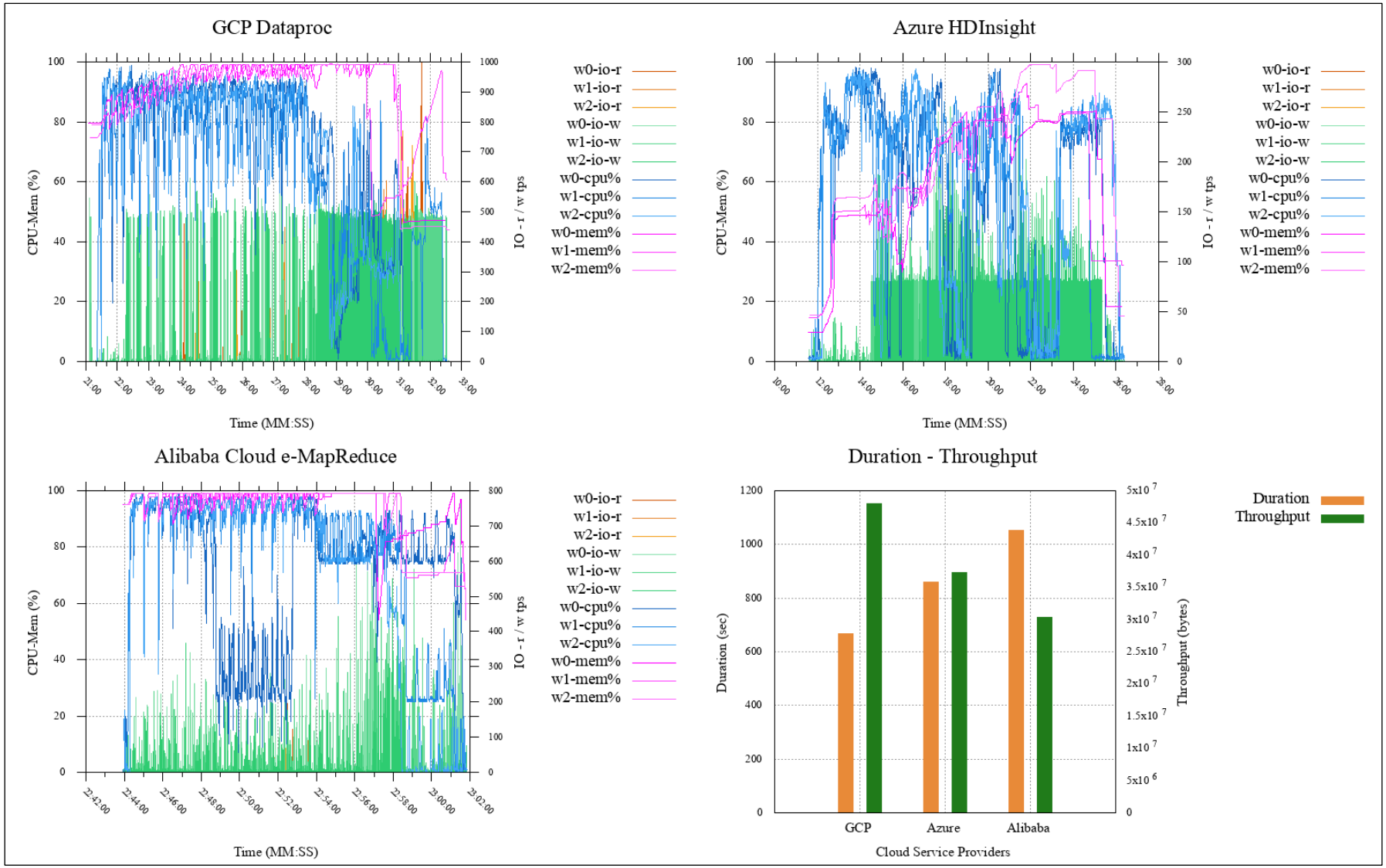
USE CASE 1 - Sort (Huge; 3.2 GB)



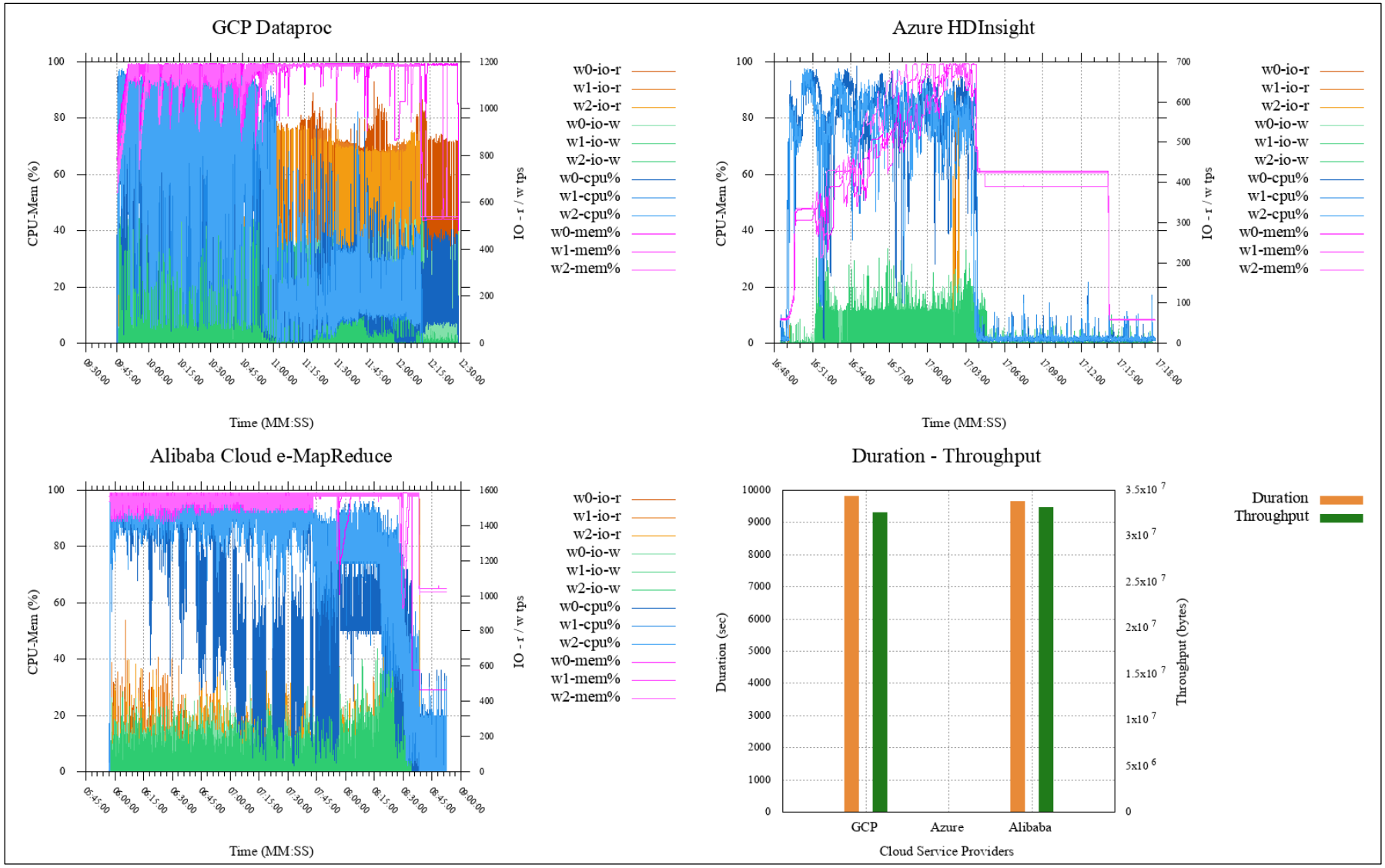
USE CASE 1 - Sort (Gigantic; 32 GB)



USE CASE 1 - Terasort (Huge; 320 MB)

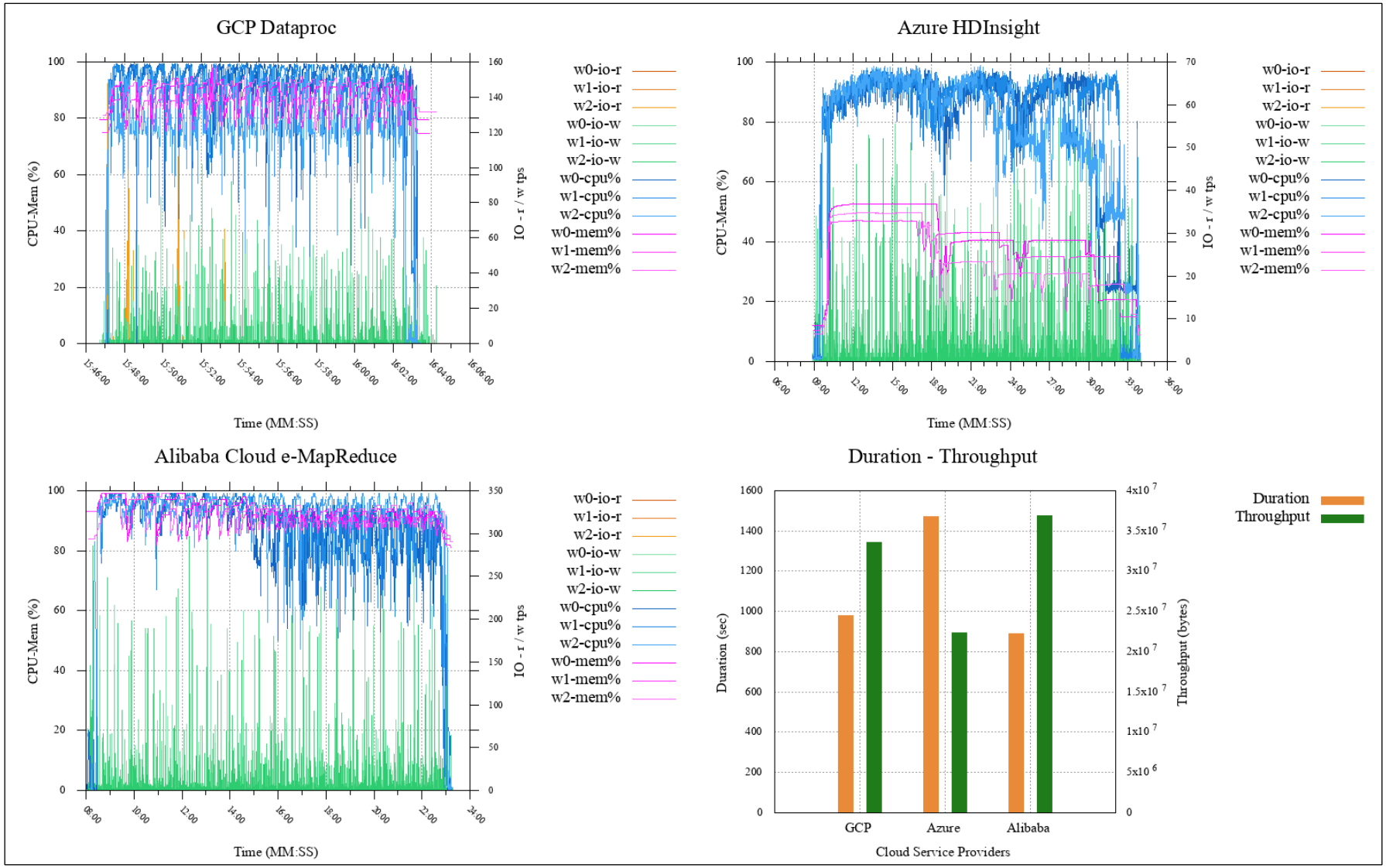


USE CASE 1 - Terasort (Gigantic; 3.2 GB)

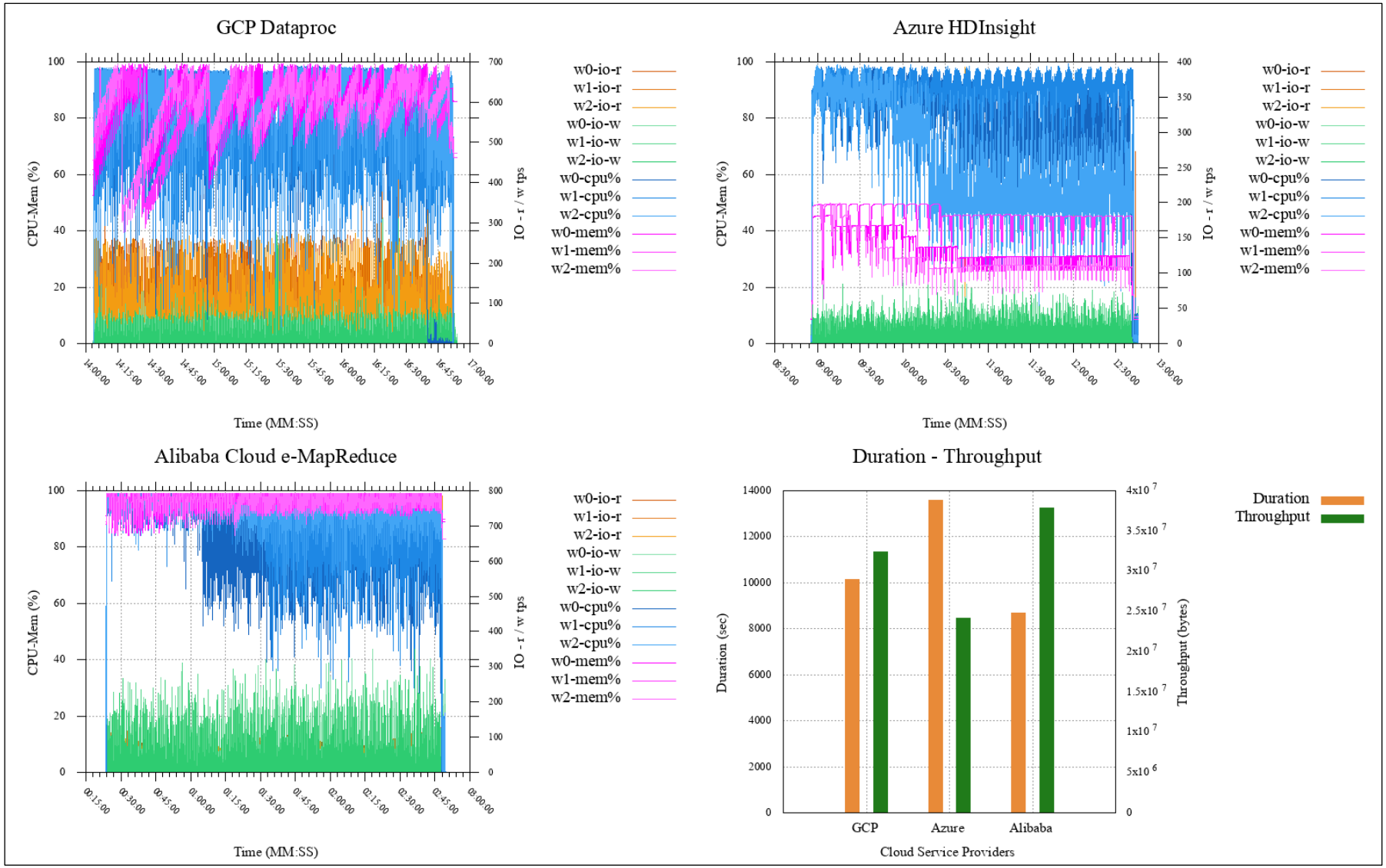


Azure's Duration-Throughput results could not be visualized due to incomplete workload/benchmark.

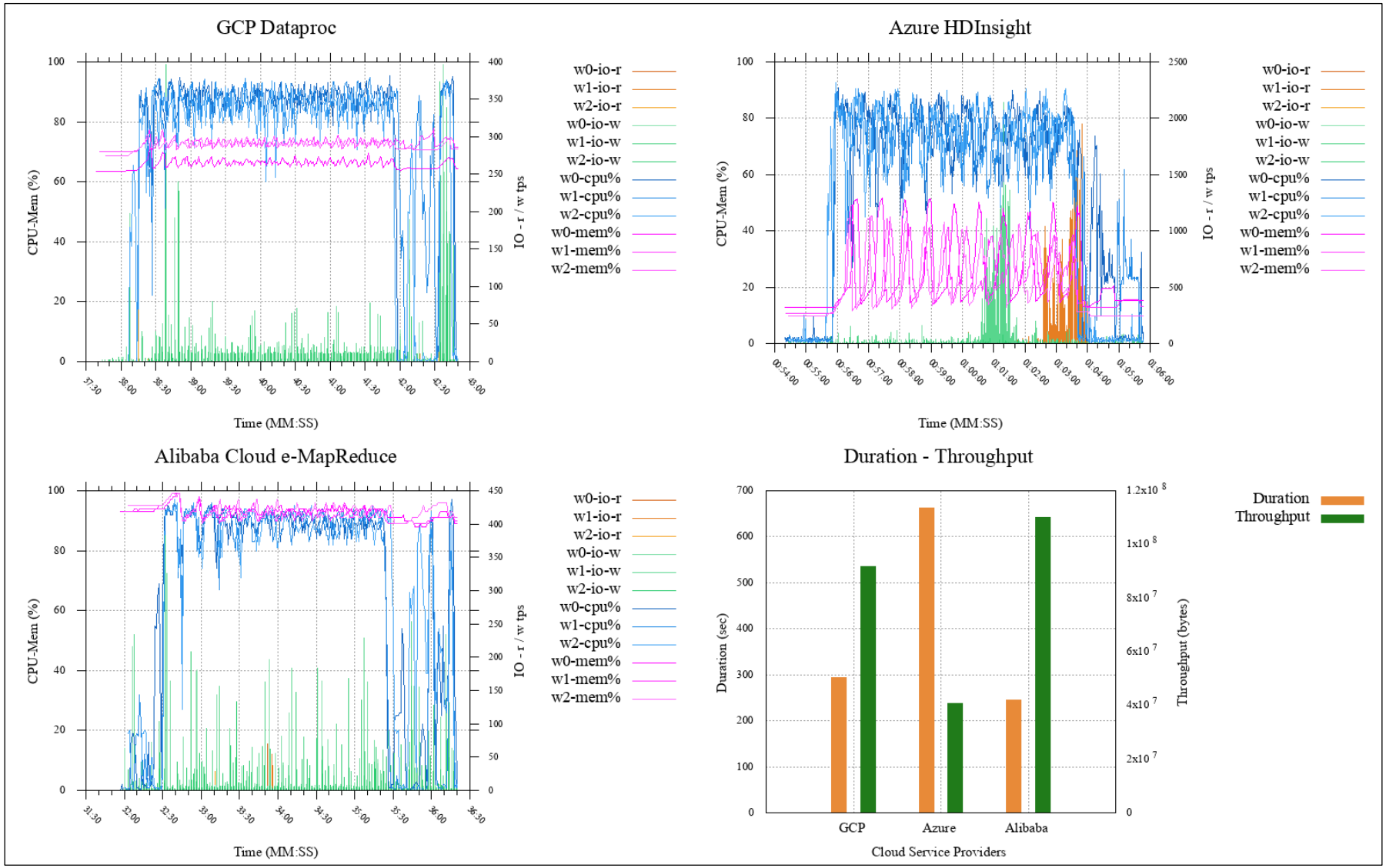
USE CASE 1 - Wordcount (Huge; 32 GB)



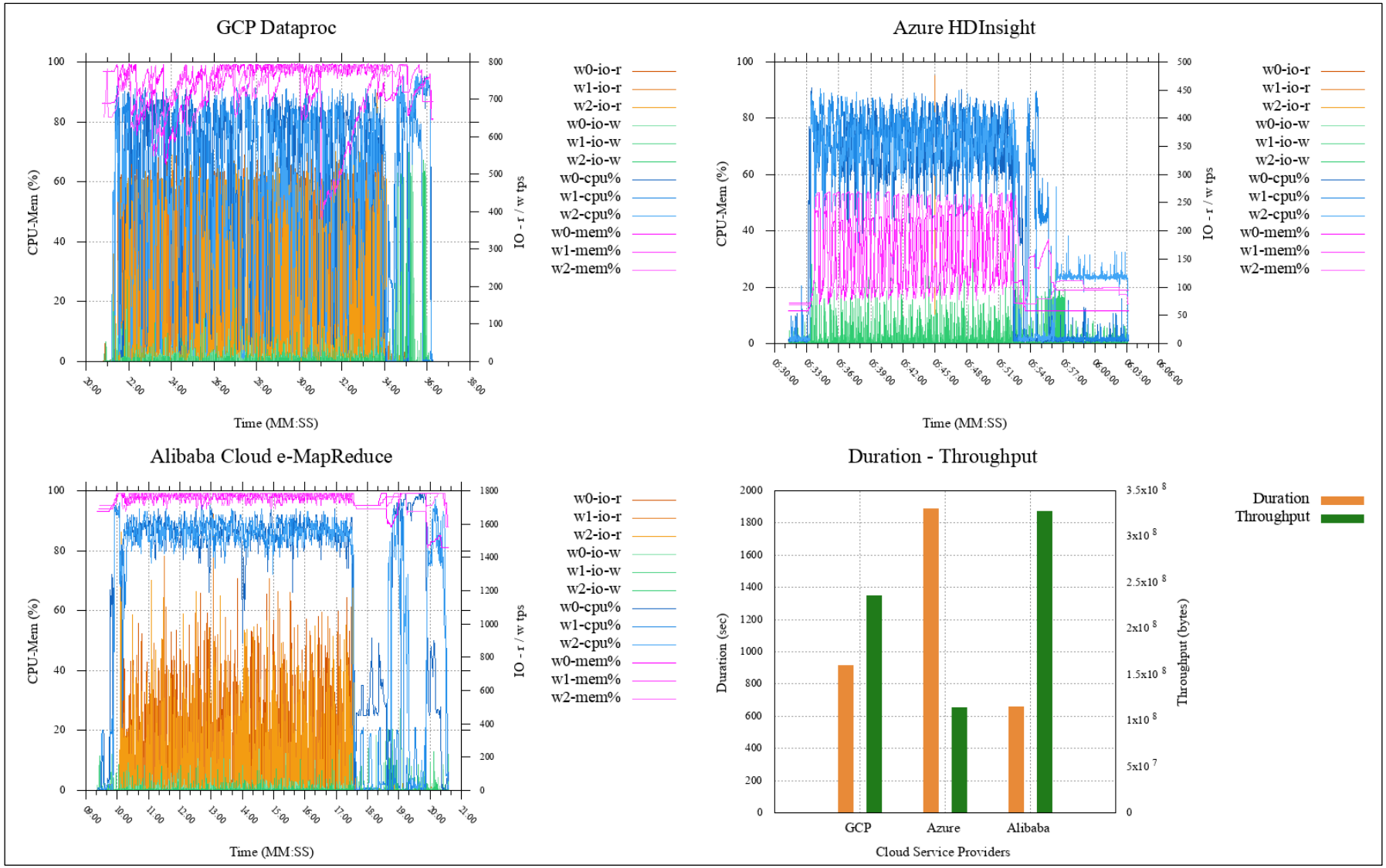
USE CASE 1 - Wordcount (Gigantic; 320 GB)



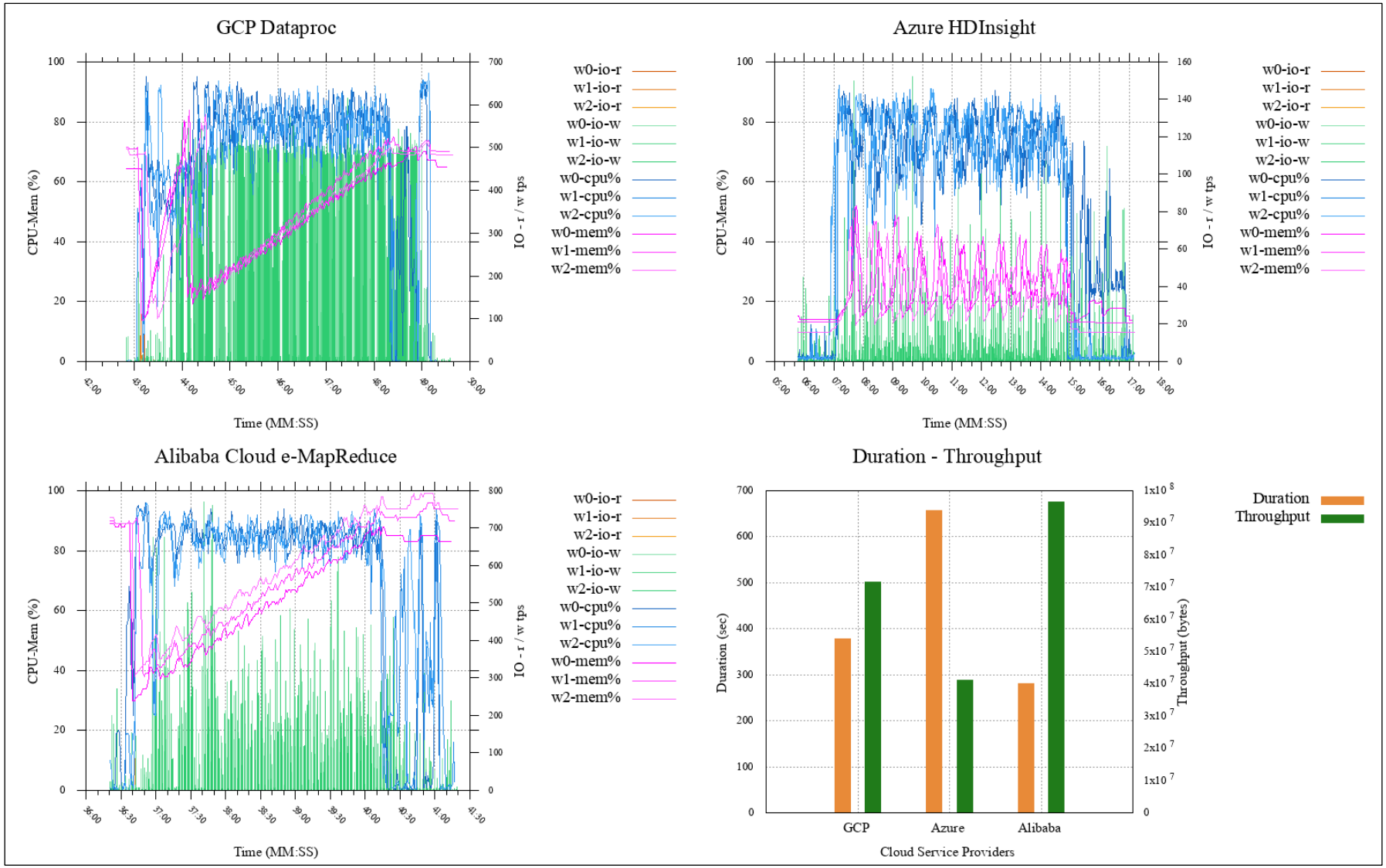
USE CASE 1 - Dfsioe-read (Huge; No of Files: 256, File size: 100 MB)



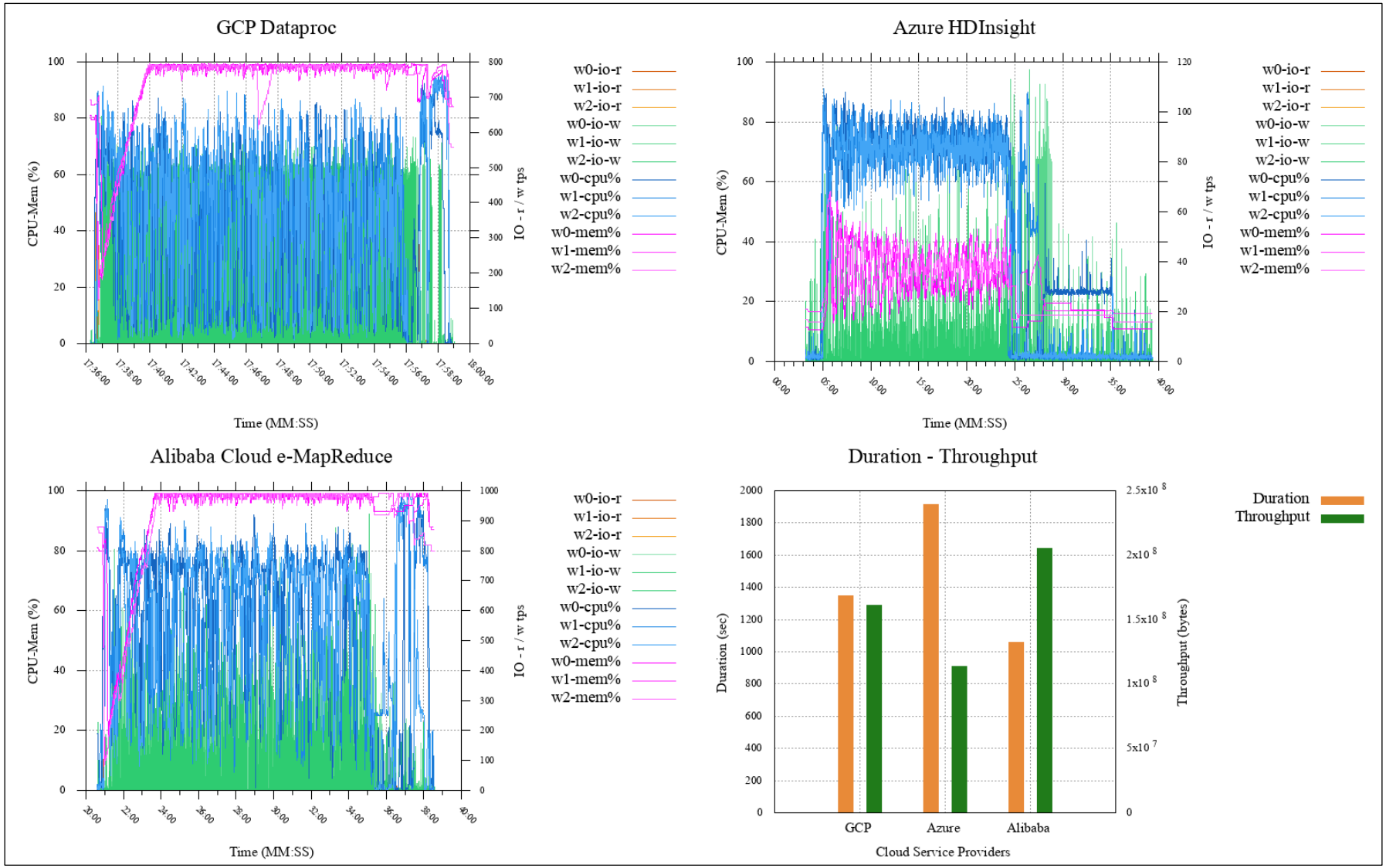
USE CASE 1 - Dfsioe-read (Gigantic; No of Files: 512, File size: 400 MB)



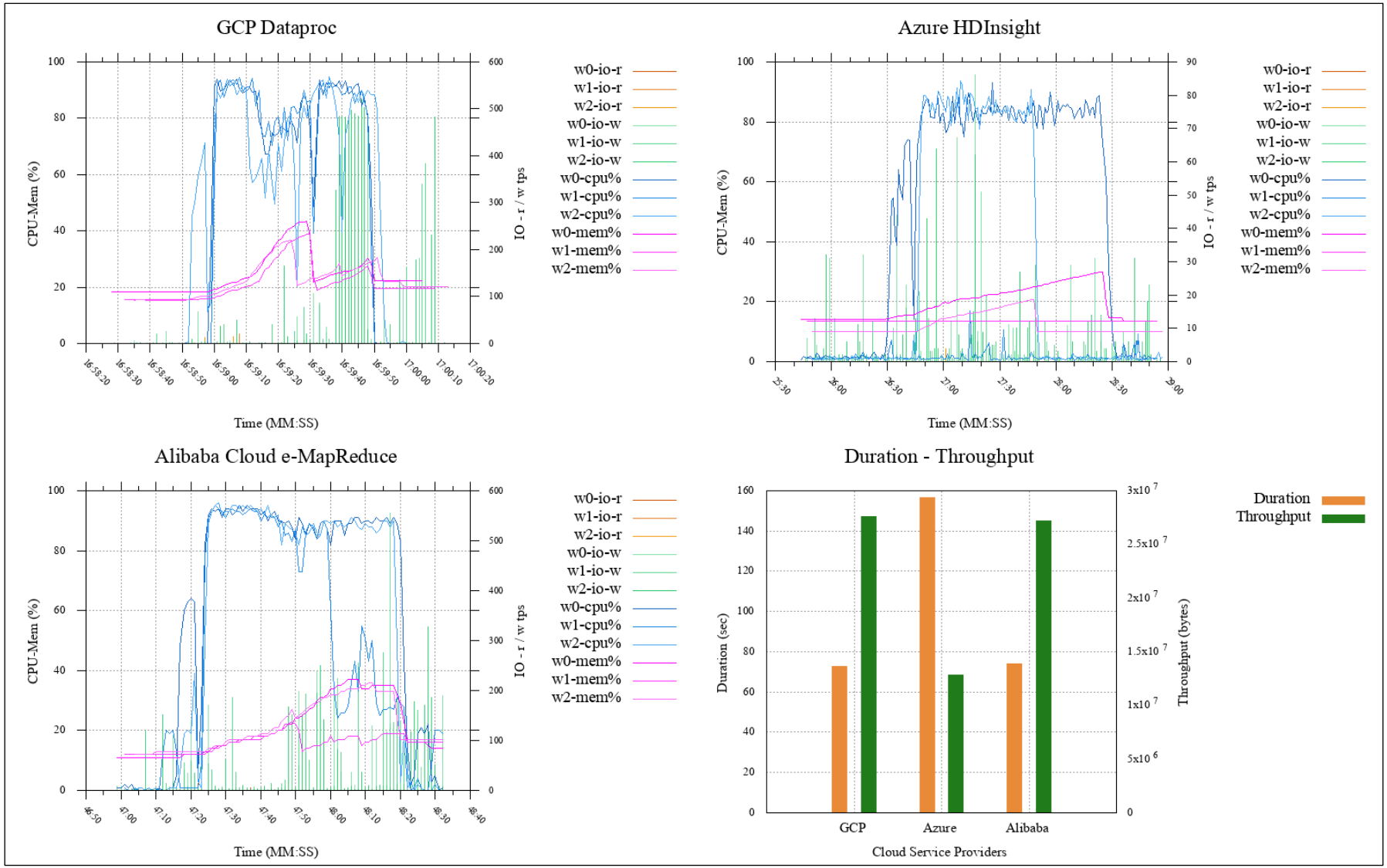
USE CASE 1 - Dfsioe-write (Huge; No of Files: 256, File size: 100 MB)



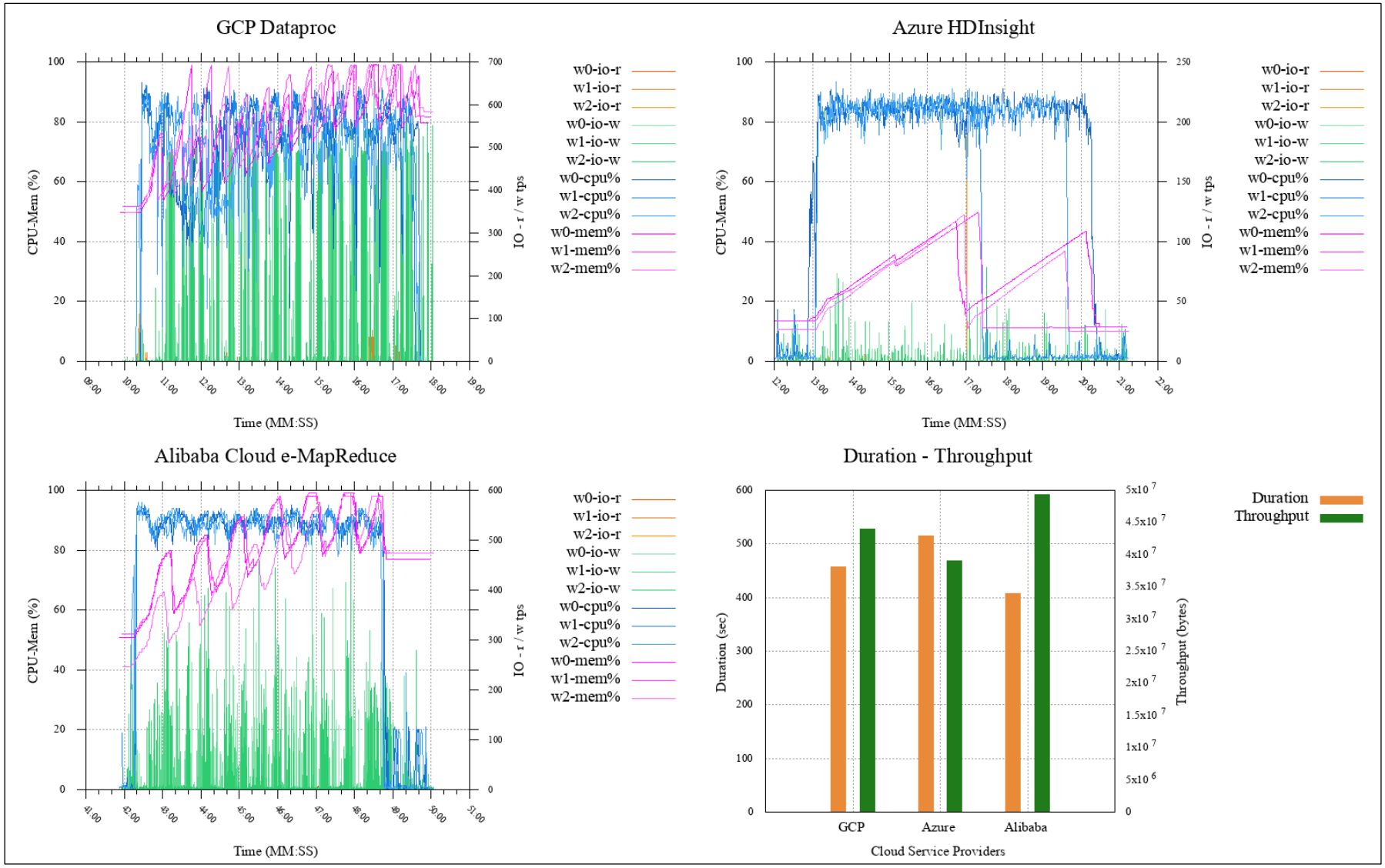
USE CASE 1 - Dfsioe-write (Gigantic; No of Files: 512, File size: 400 MB)



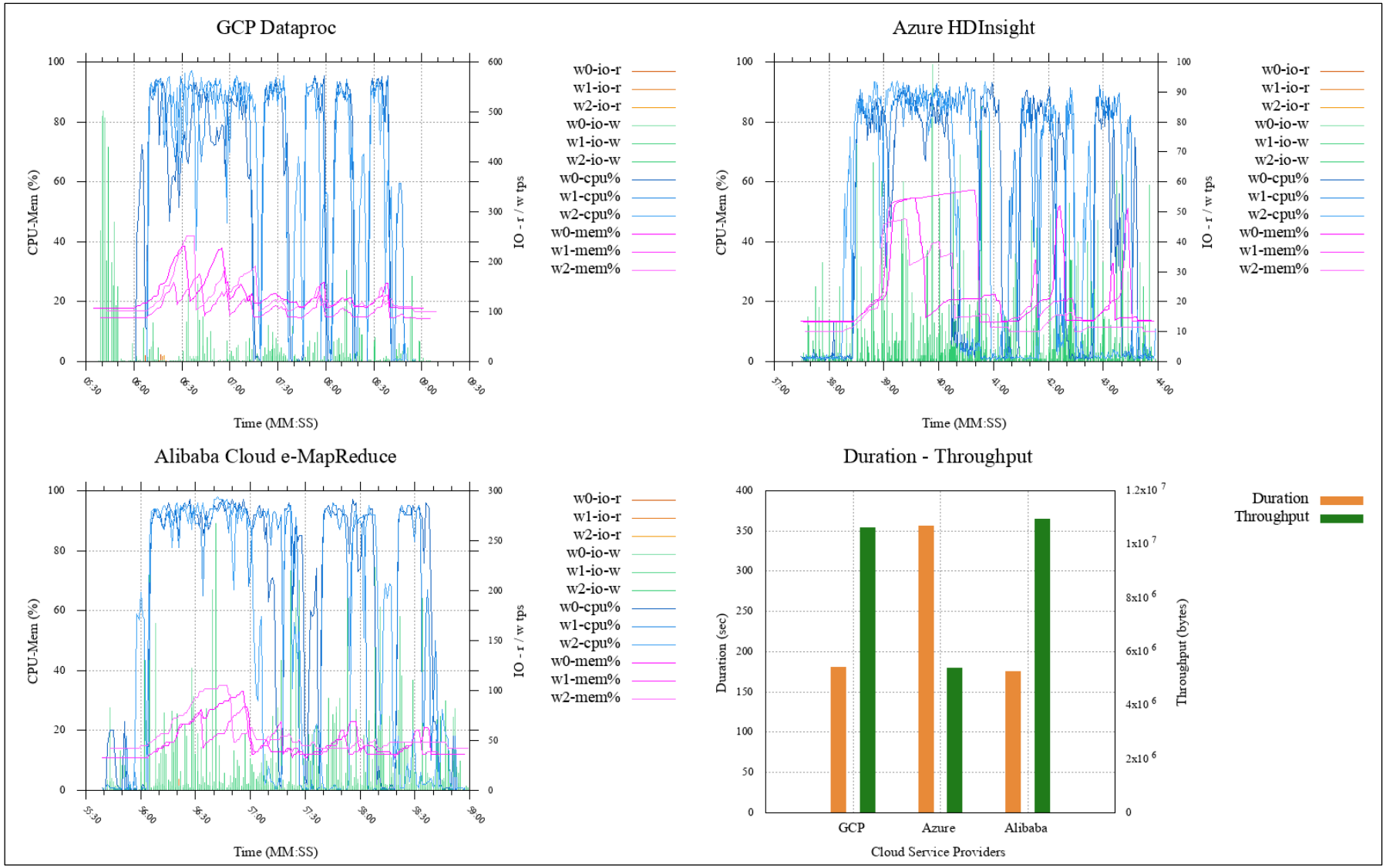
USE CASE 1 - Scan (Huge; Uservisits: 10,000,000 Pages: 1,200,000)



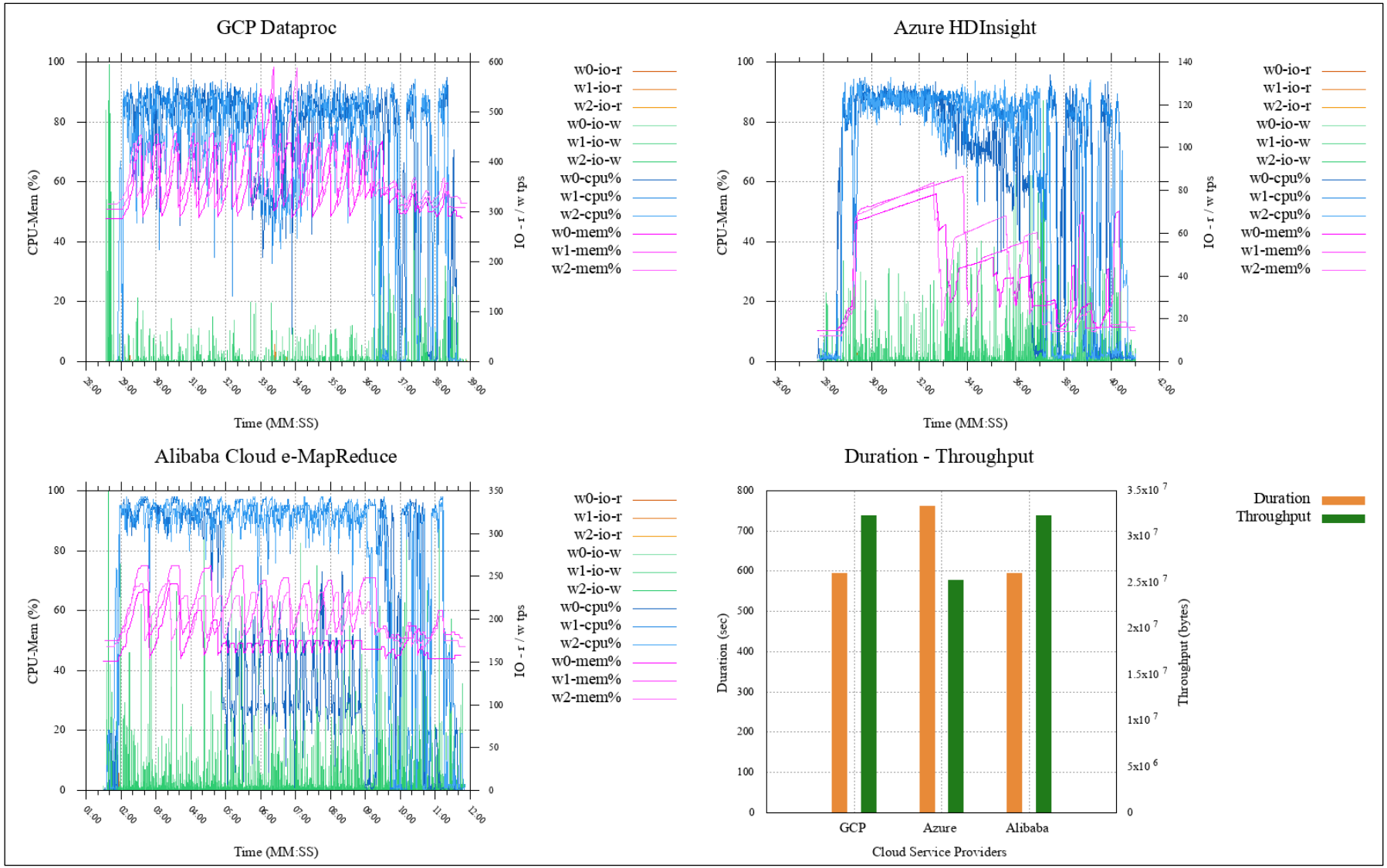
USE CASE 1 - Scan (Gigantic; Uservisits: 100,000,000 Pages: 12,000,000)



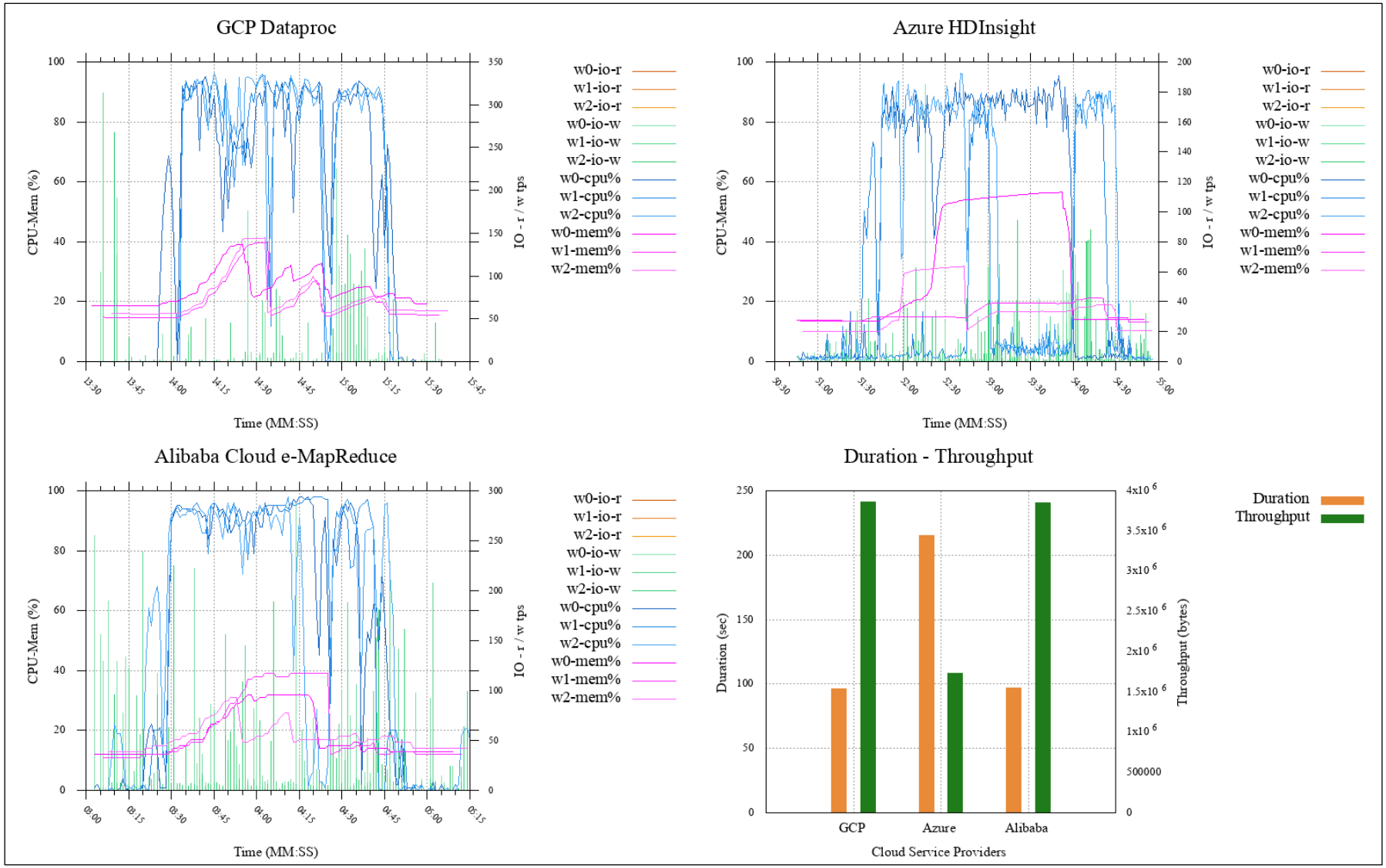
USE CASE 1 - Join (Huge; Uservisits: 10,000,000 Pages: 1,200,000)



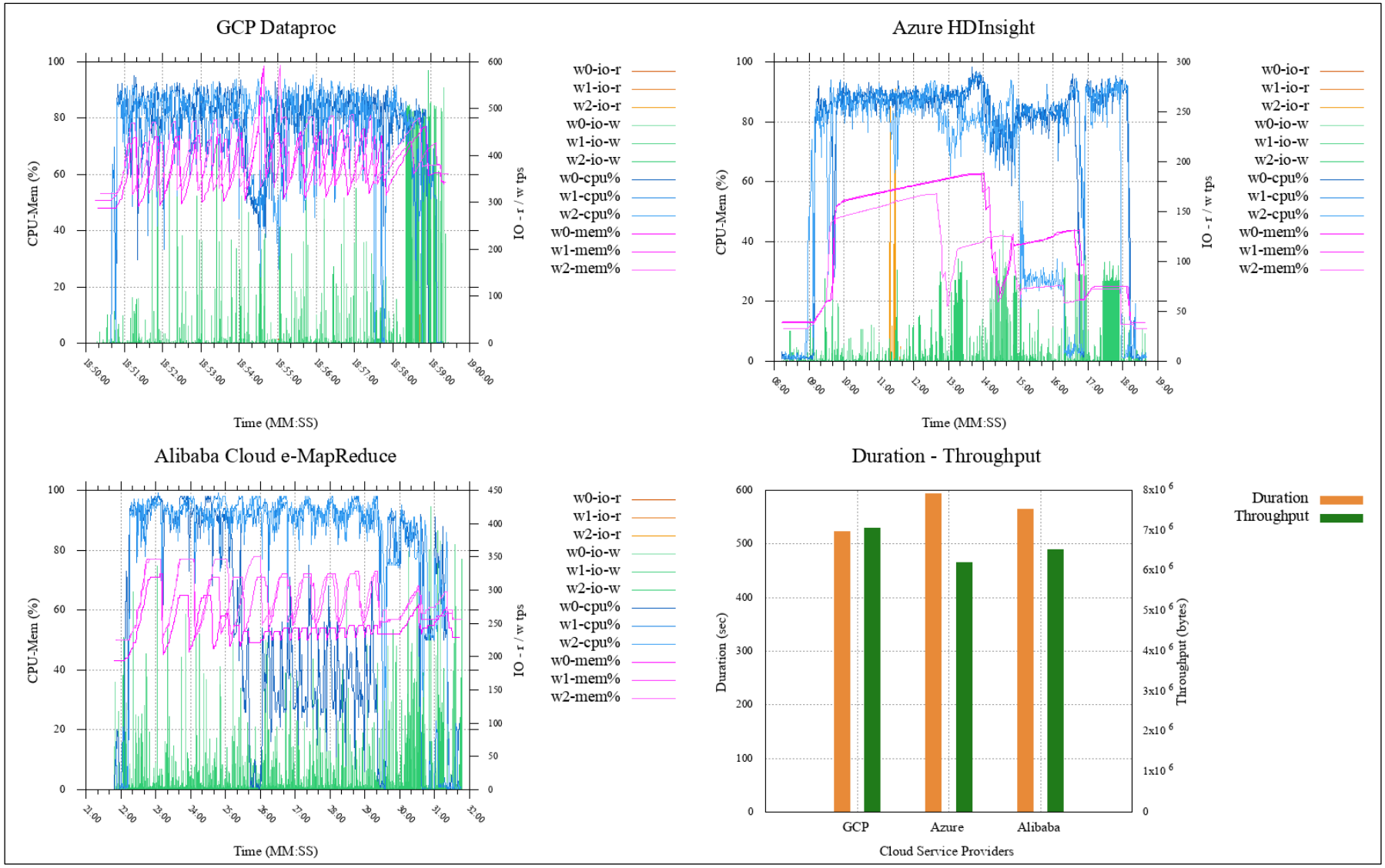
USE CASE 1 - Join (Gigantic; Uservisits: 100,000,000 Pages: 12,000,000)



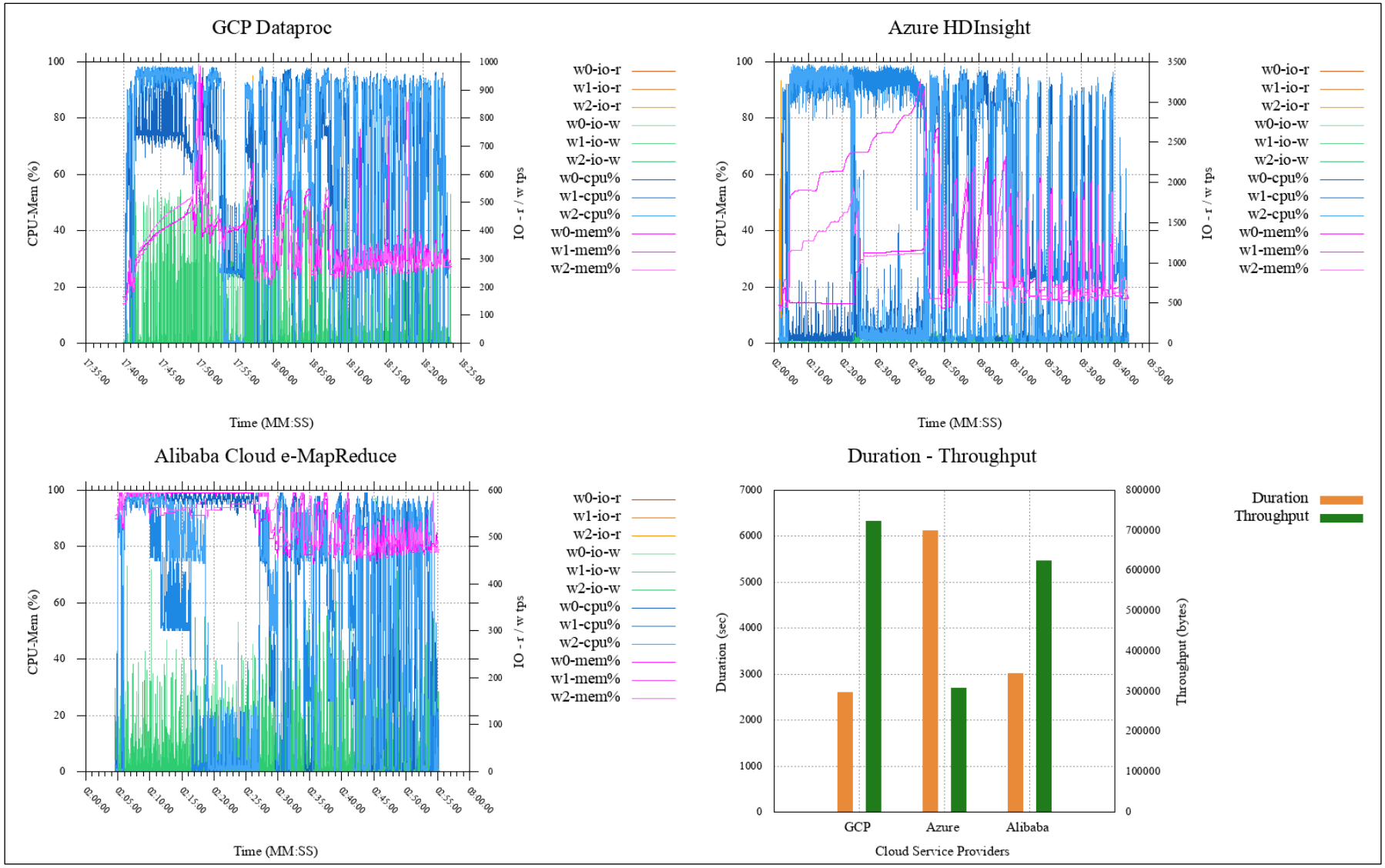
USE CASE 1 - Aggregation (Huge; Uservisits: 10,000,000 Pages: 1,200,000)



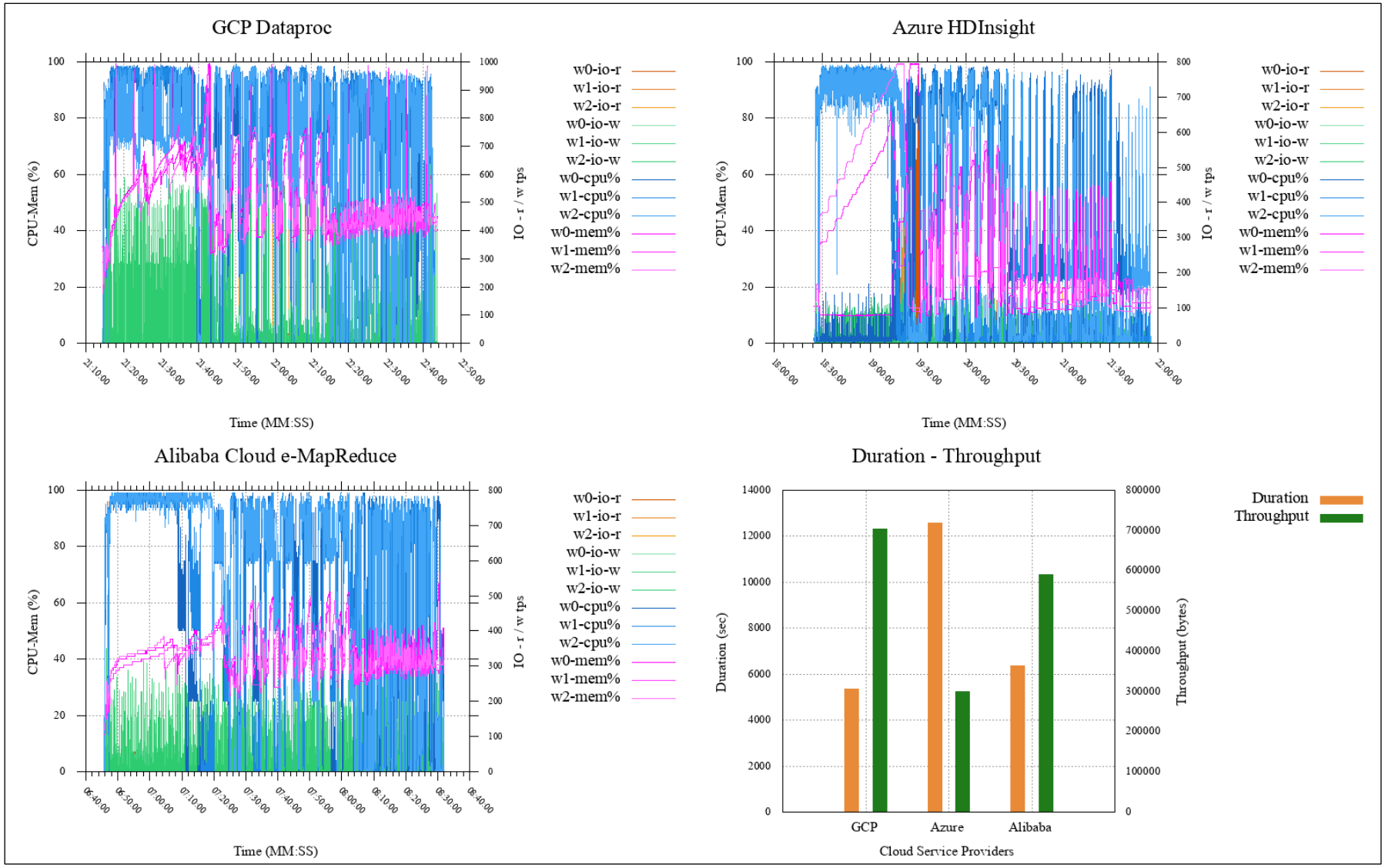
USE CASE 1 - Aggregation (Gigantic; Uservisits: 100,000,000 Pages: 12,000,000)



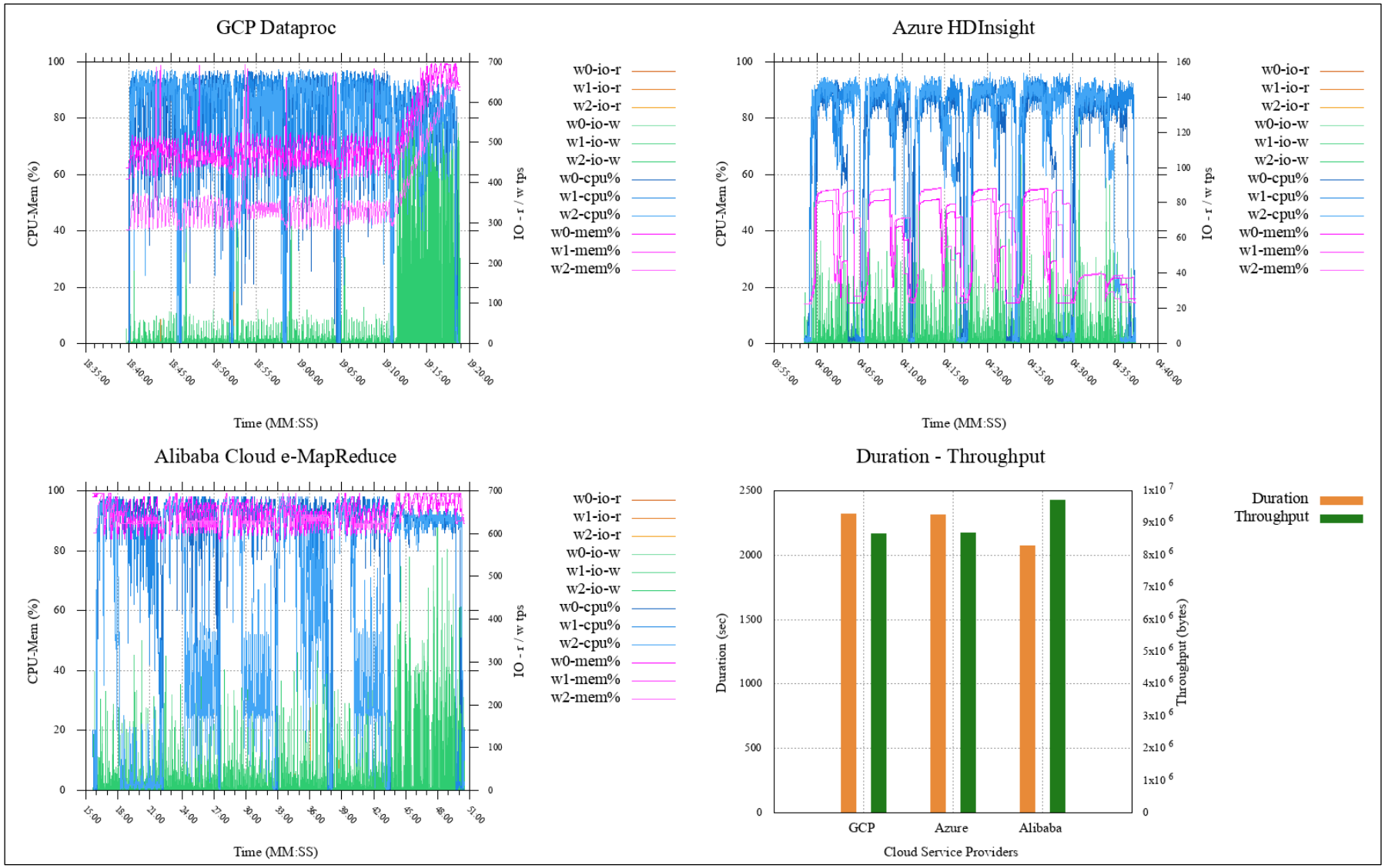
USE CASE 1 - Bayes (Huge; Pages: 500,000 Classes: 100 Ngrams: 2)



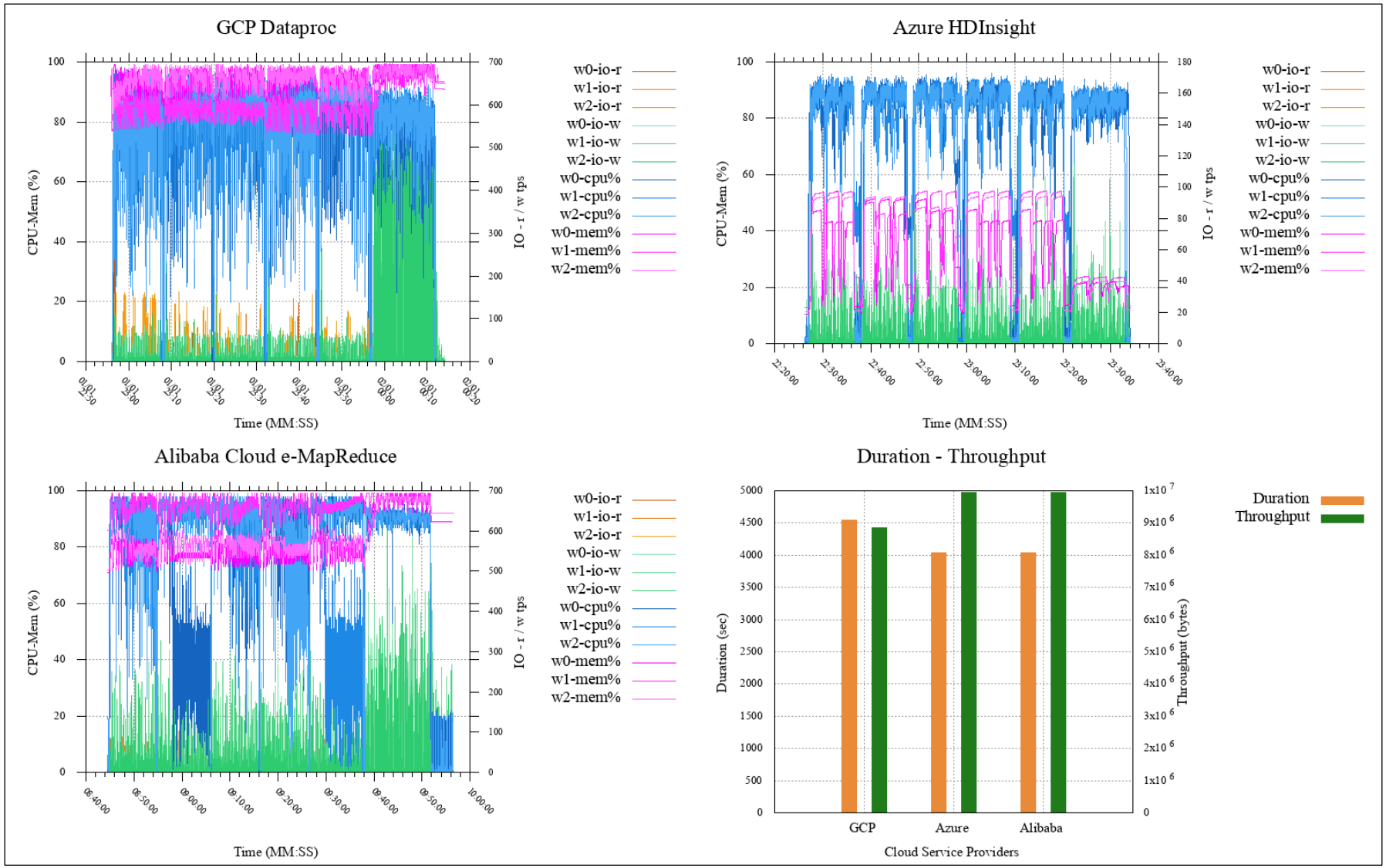
USE CASE 1 - Bayes (Gigantic; Pages: 1,000,000 Classes: 100 Ngrams: 2)



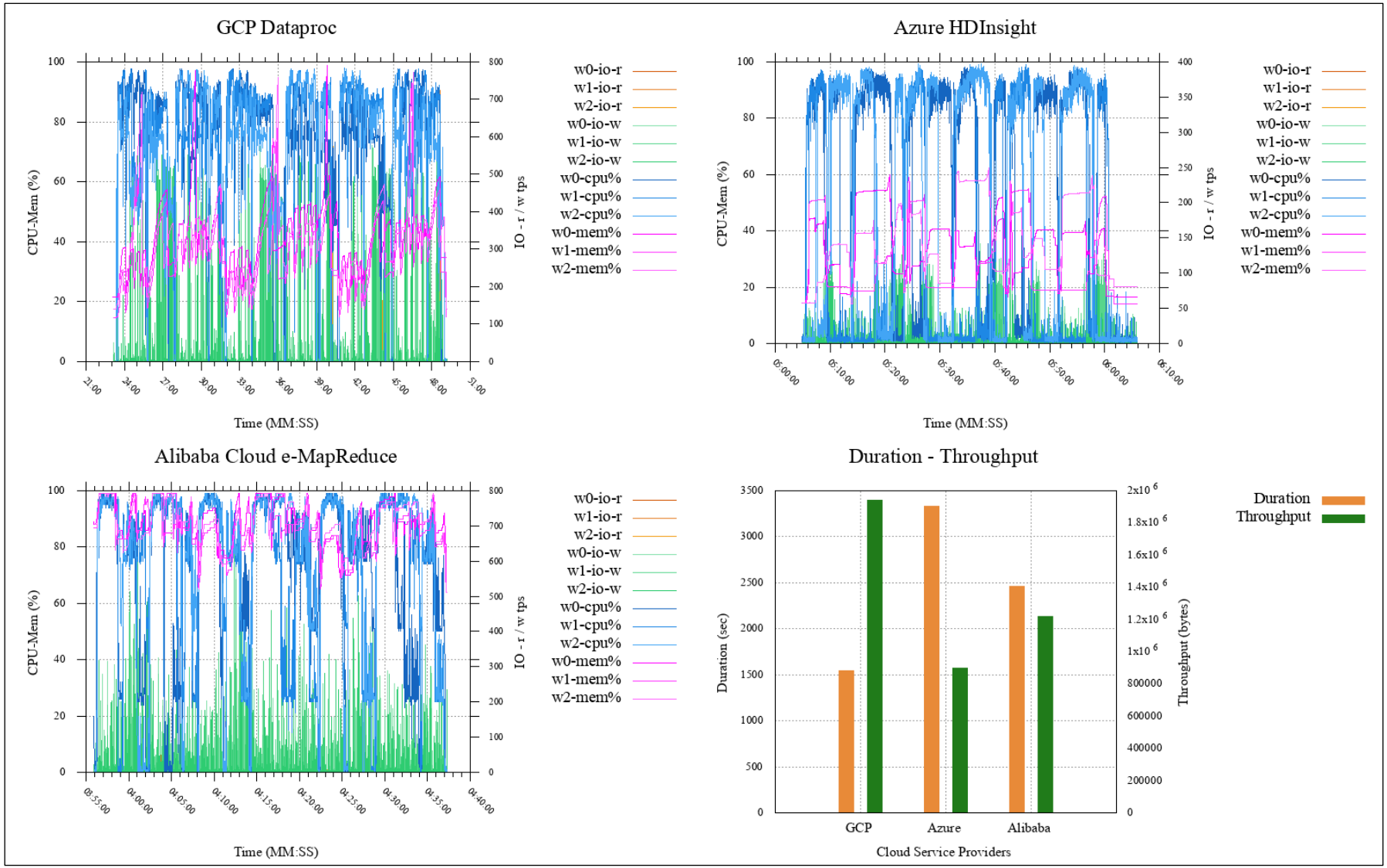
USE CASE 1 - Kmeans (Huge; Clusters: 5 Dimensions: 20 Samples: 100,000,000 Samp Per Input: 20,000,000 Max It: 5 K: 10 Convergedist: 0.5)



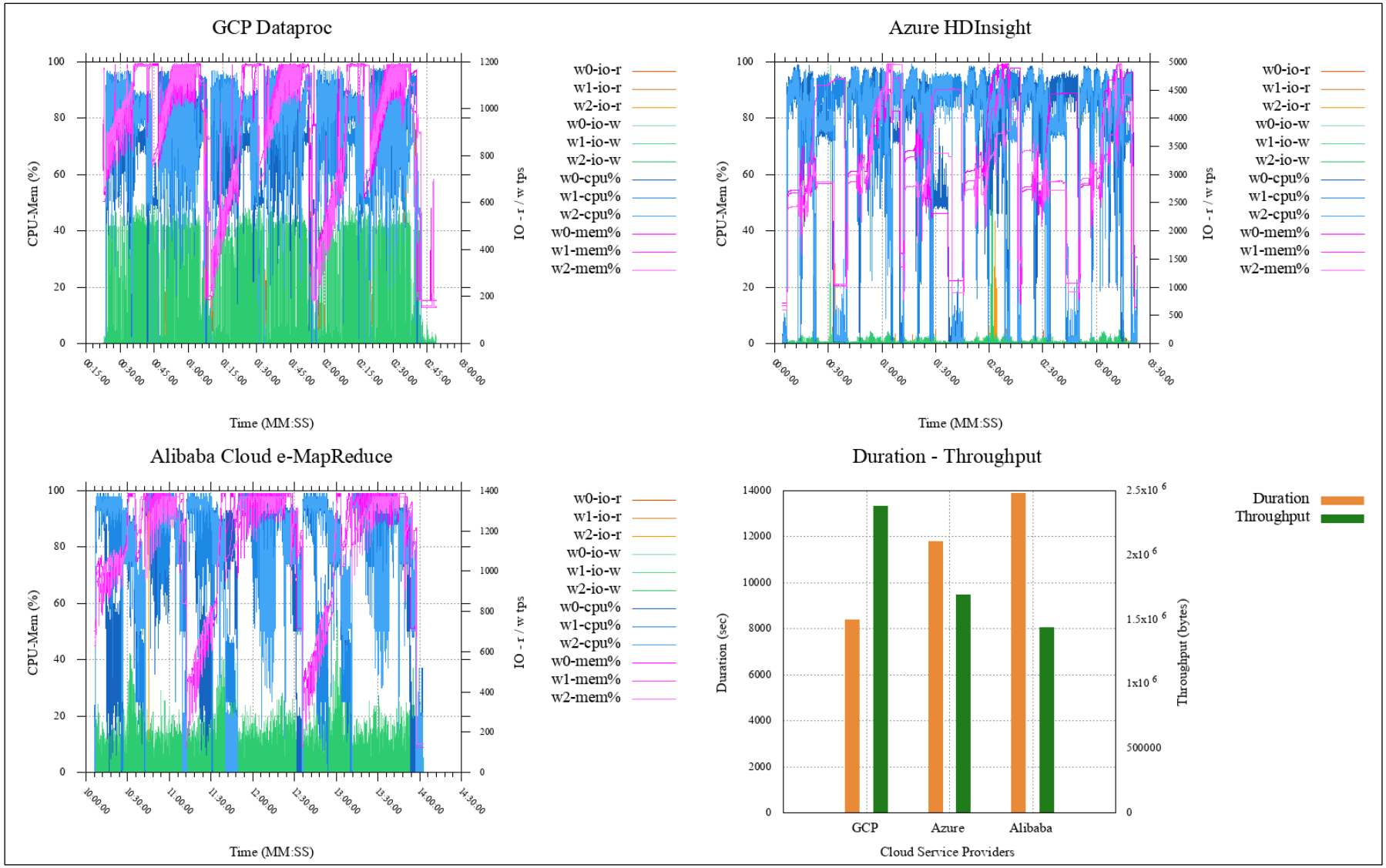
USE CASE 1 - Kmeans (Gigantic; Clusters: 5 Dimensions: 20 Samples: 200,000,000 Samp Per Input: 40,000,000 Max It: 5 K: 10 Convergedist: 0.5)



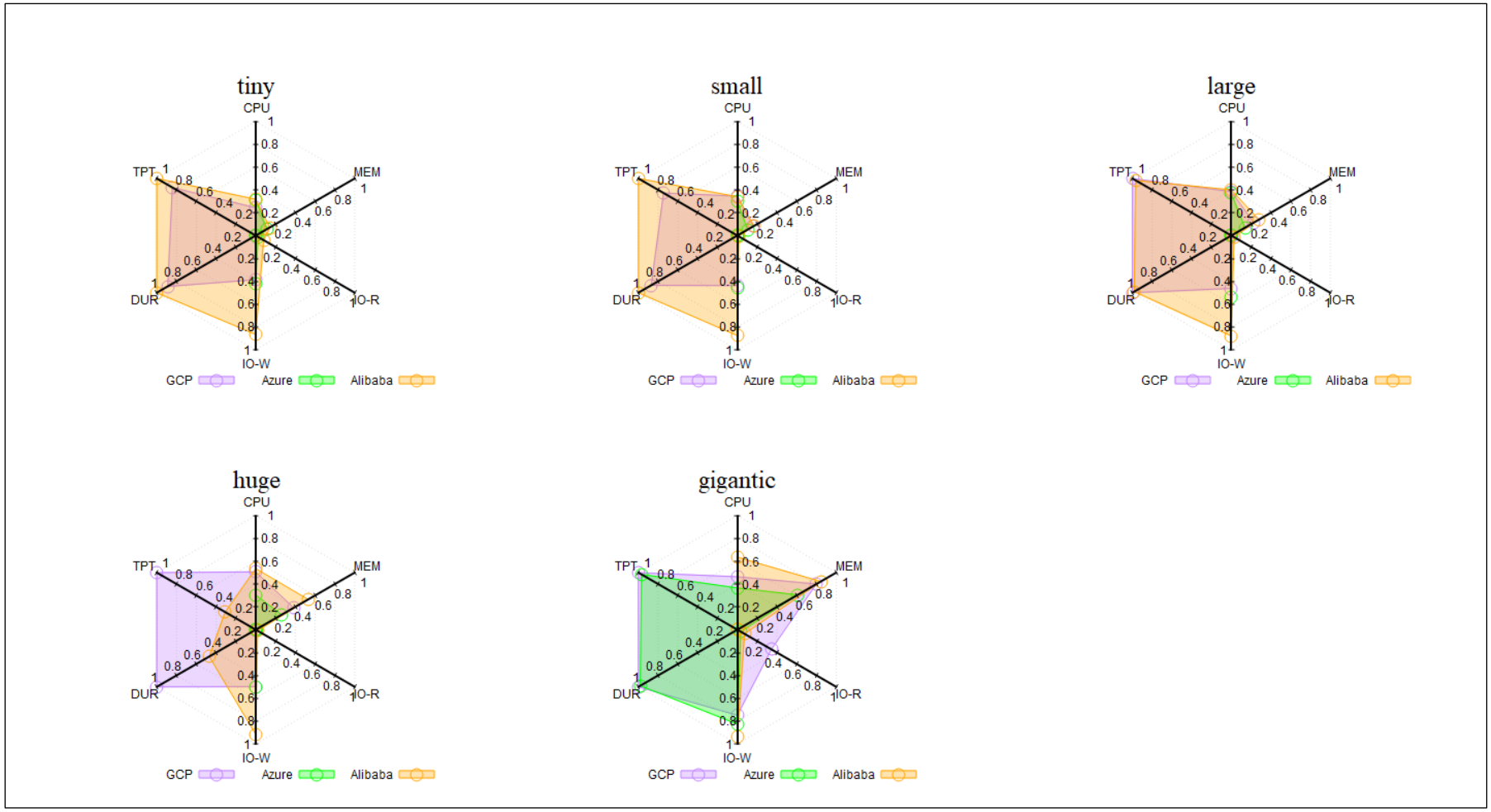
USE CASE 1 - Pagerank (Huge; Pages: 5,000,000 Num Iterations: 3 Block: 0 Block Width: 16)



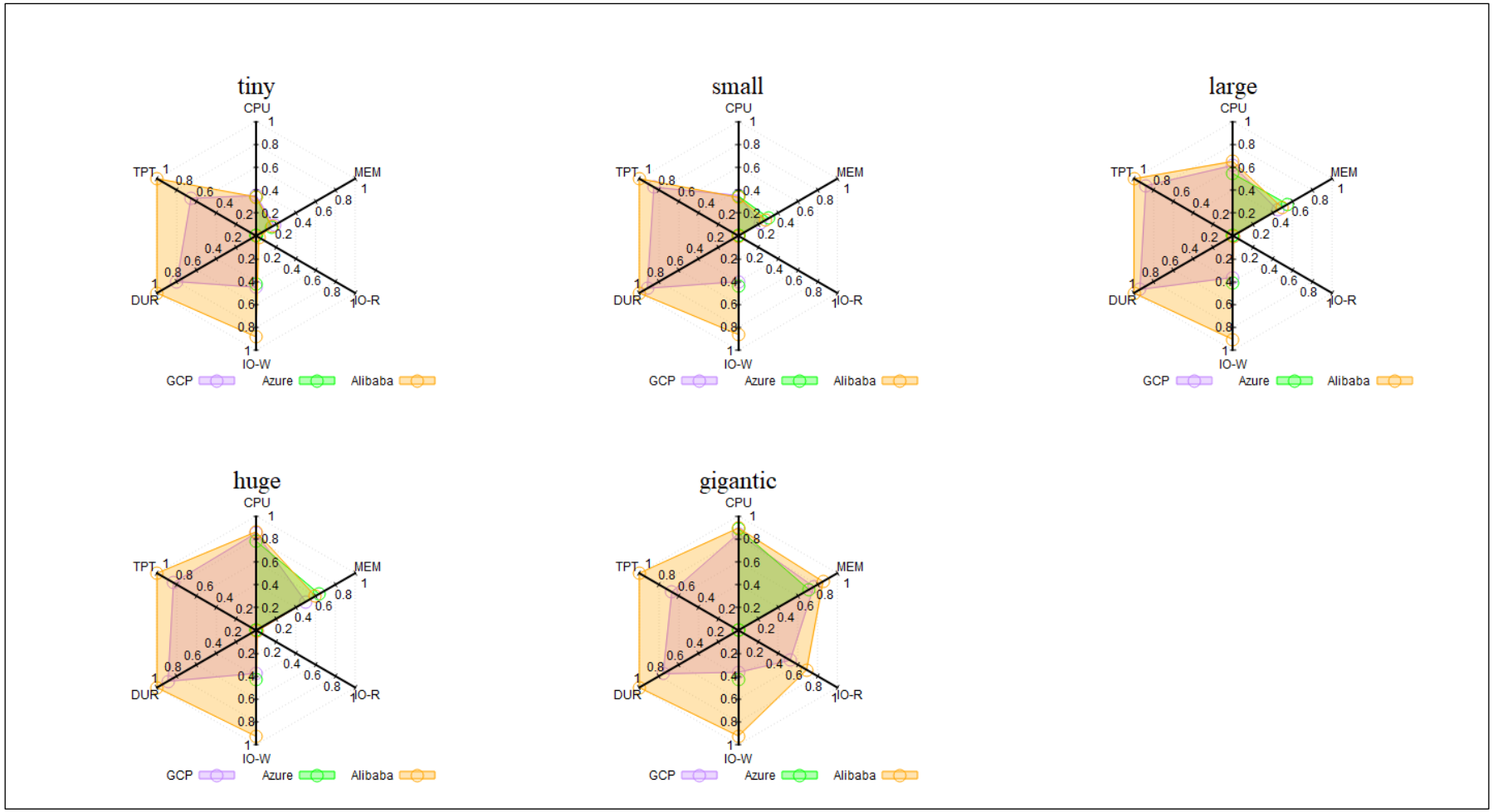
USE CASE 1 - Pagerank (Gigantic; Pages: 30,000,000 Num Iterations: 3 Block: 0 Block Width: 16)



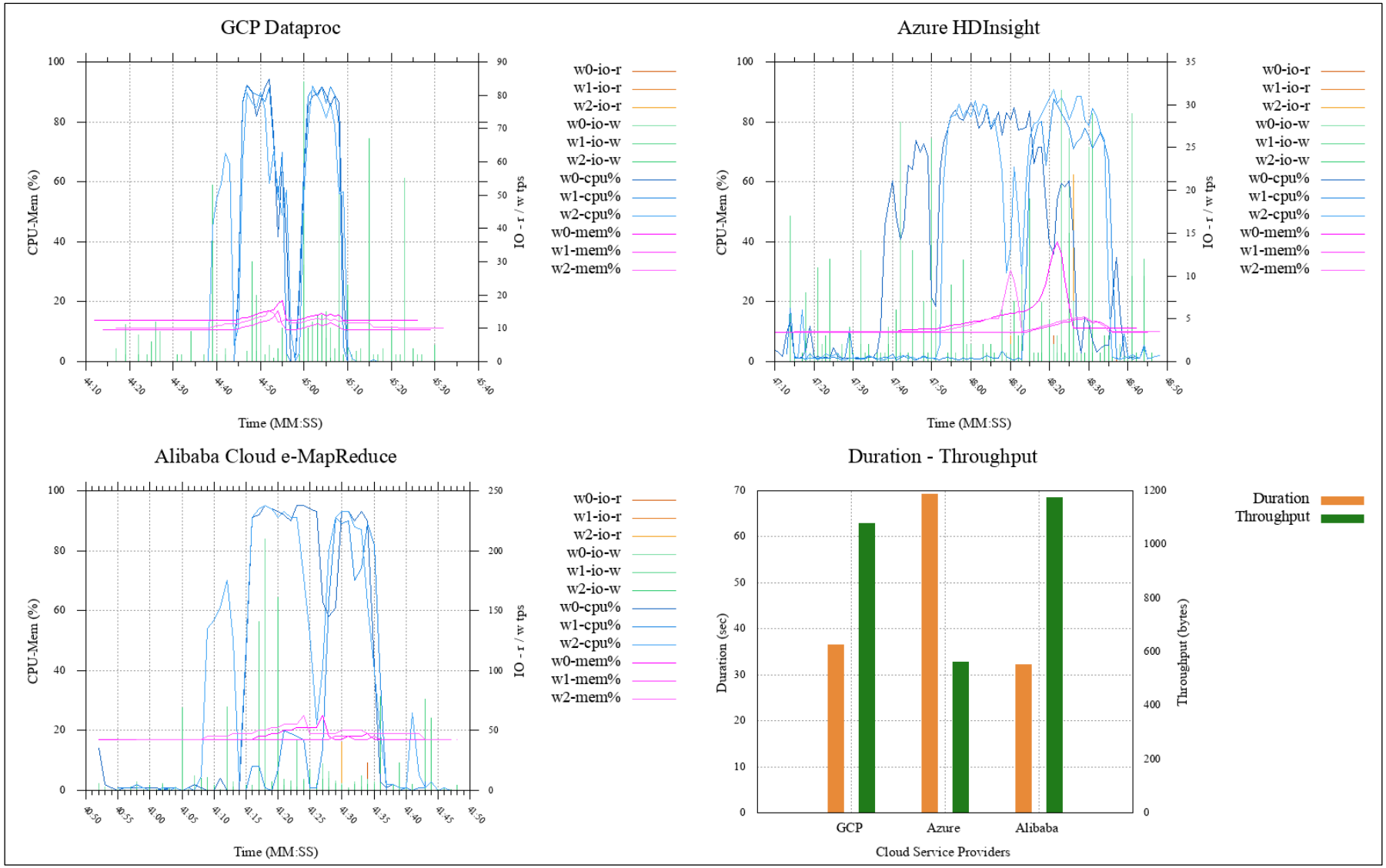
Use Case 2 – Sort: Average System Utilization Along with Data Scales



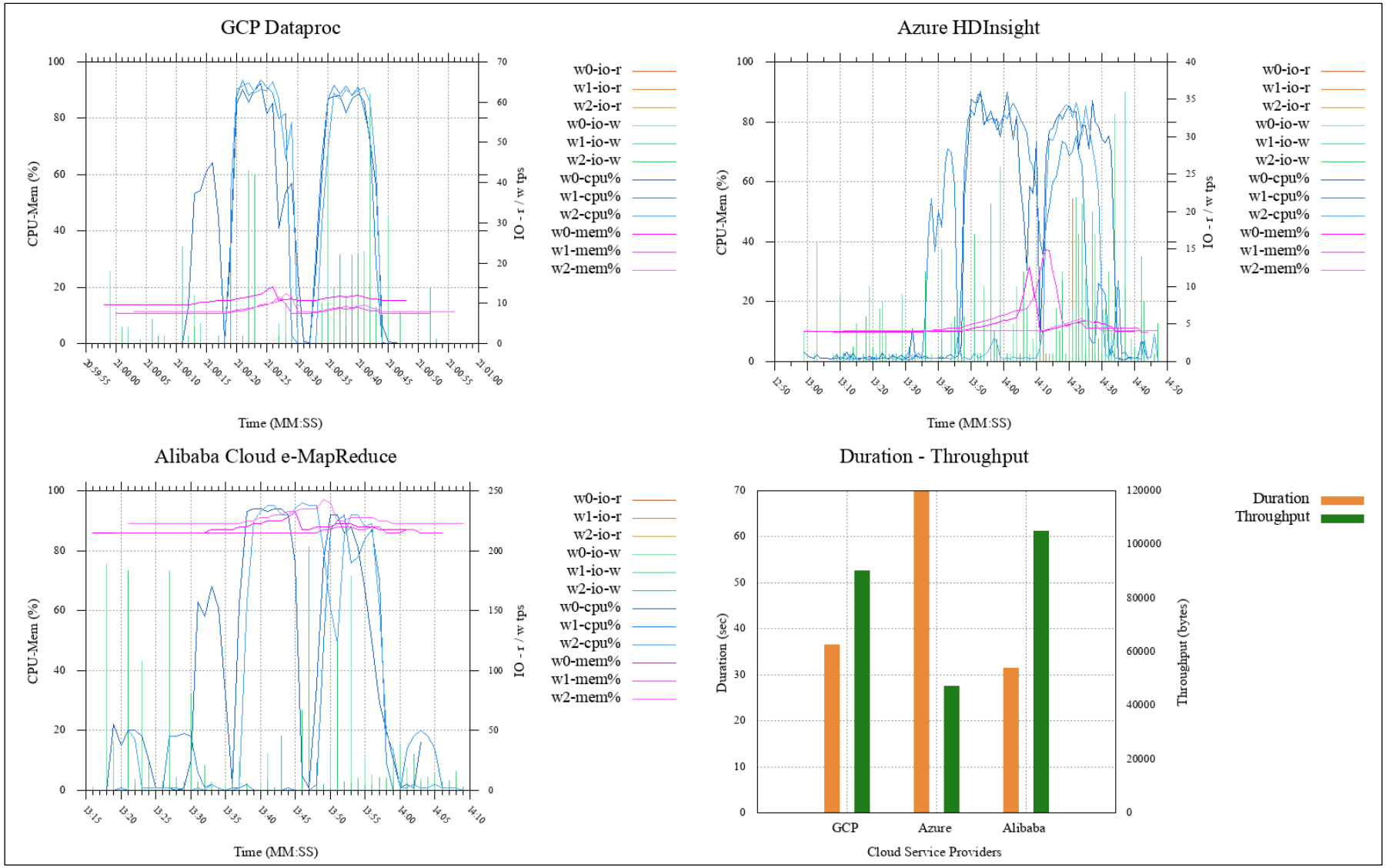
Use Case 2 – Wordcount: Average System Utilization Along with Data Scales



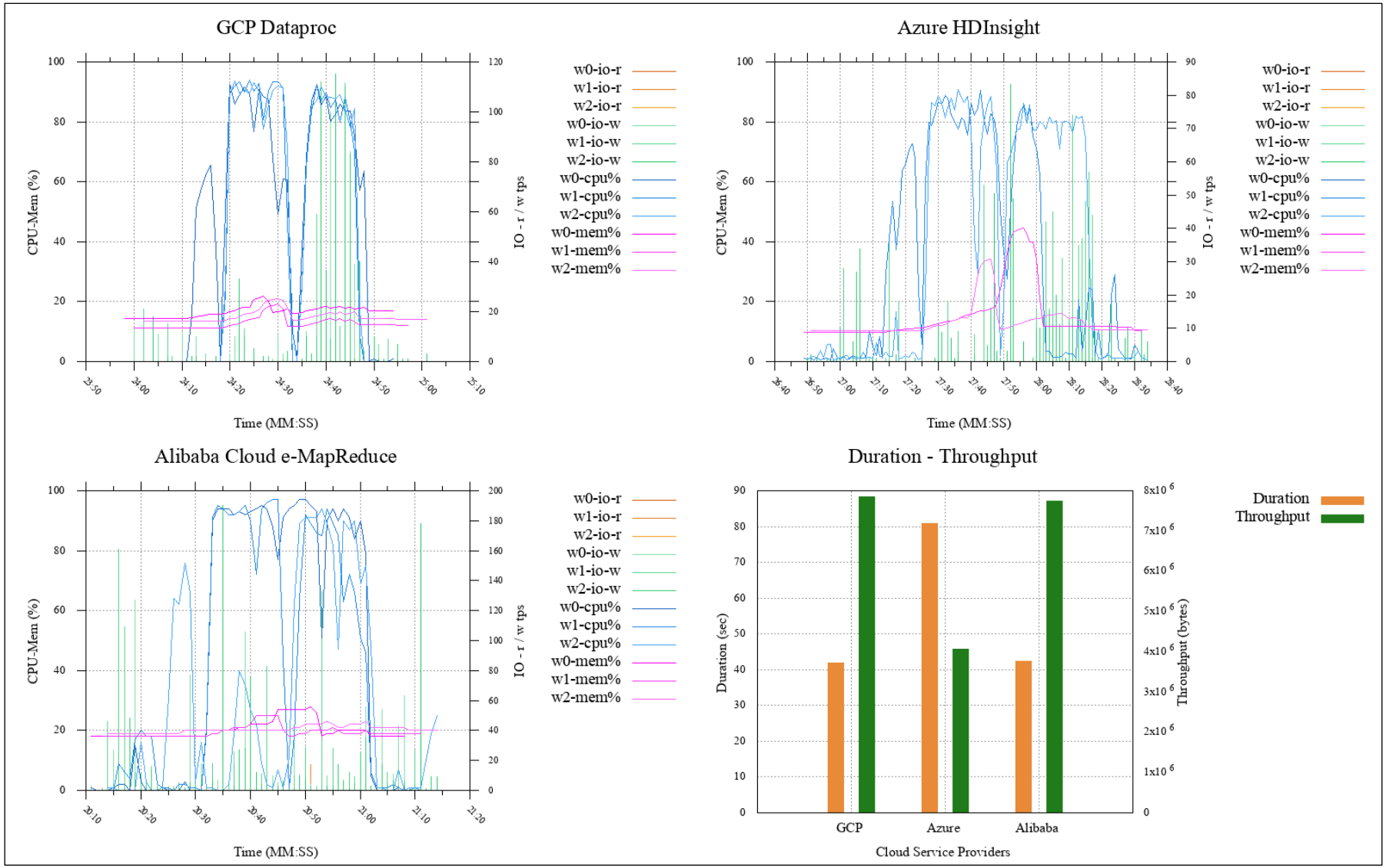
USE CASE 2 - Sort (Tiny; 32 KB)



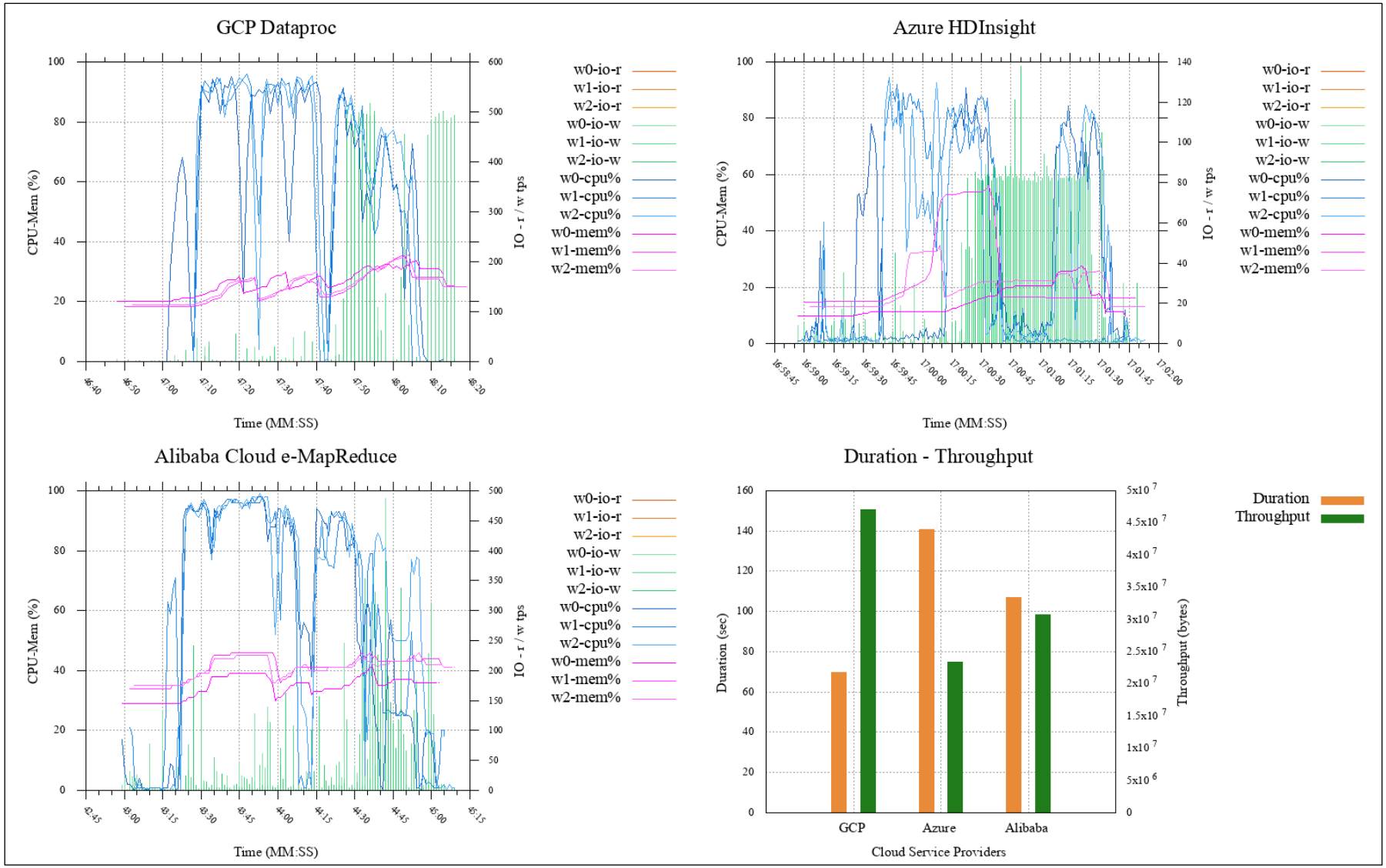
USE CASE 2 - Sort (Small; 3.2 MB)



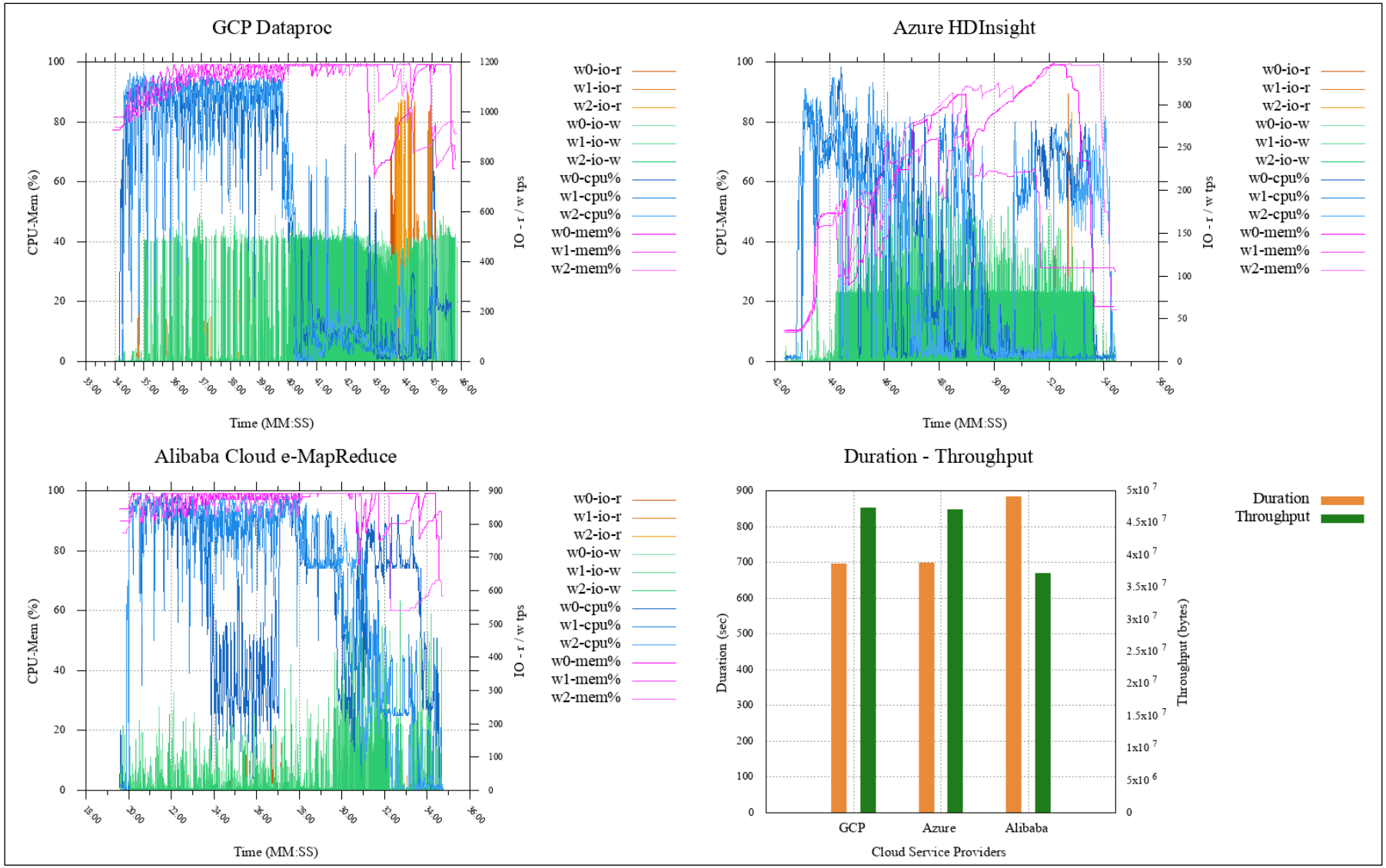
USE CASE 2 - Sort (Large; 320 MB)



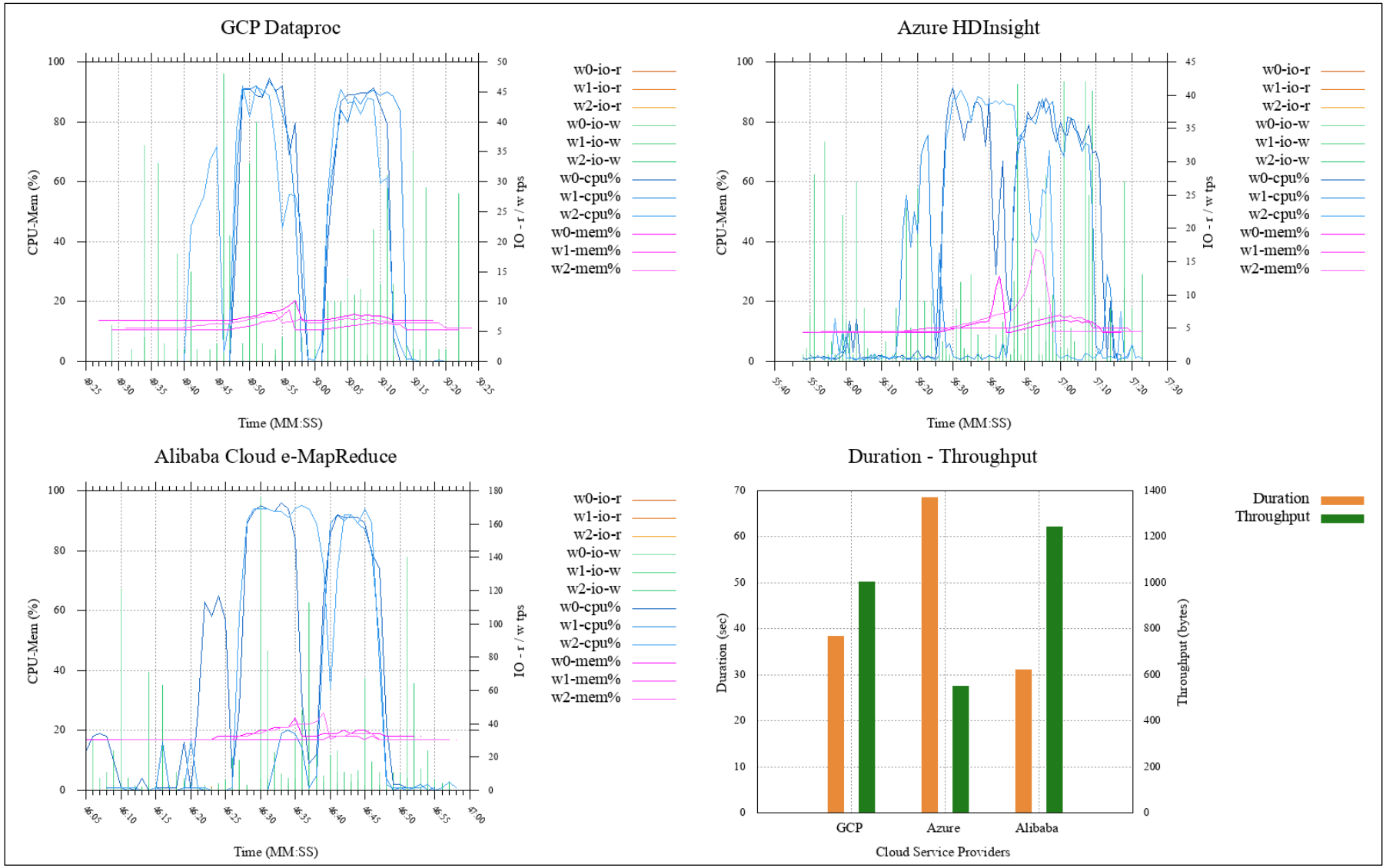
USE CASE 2 - Sort (Huge; 3.2 GB)



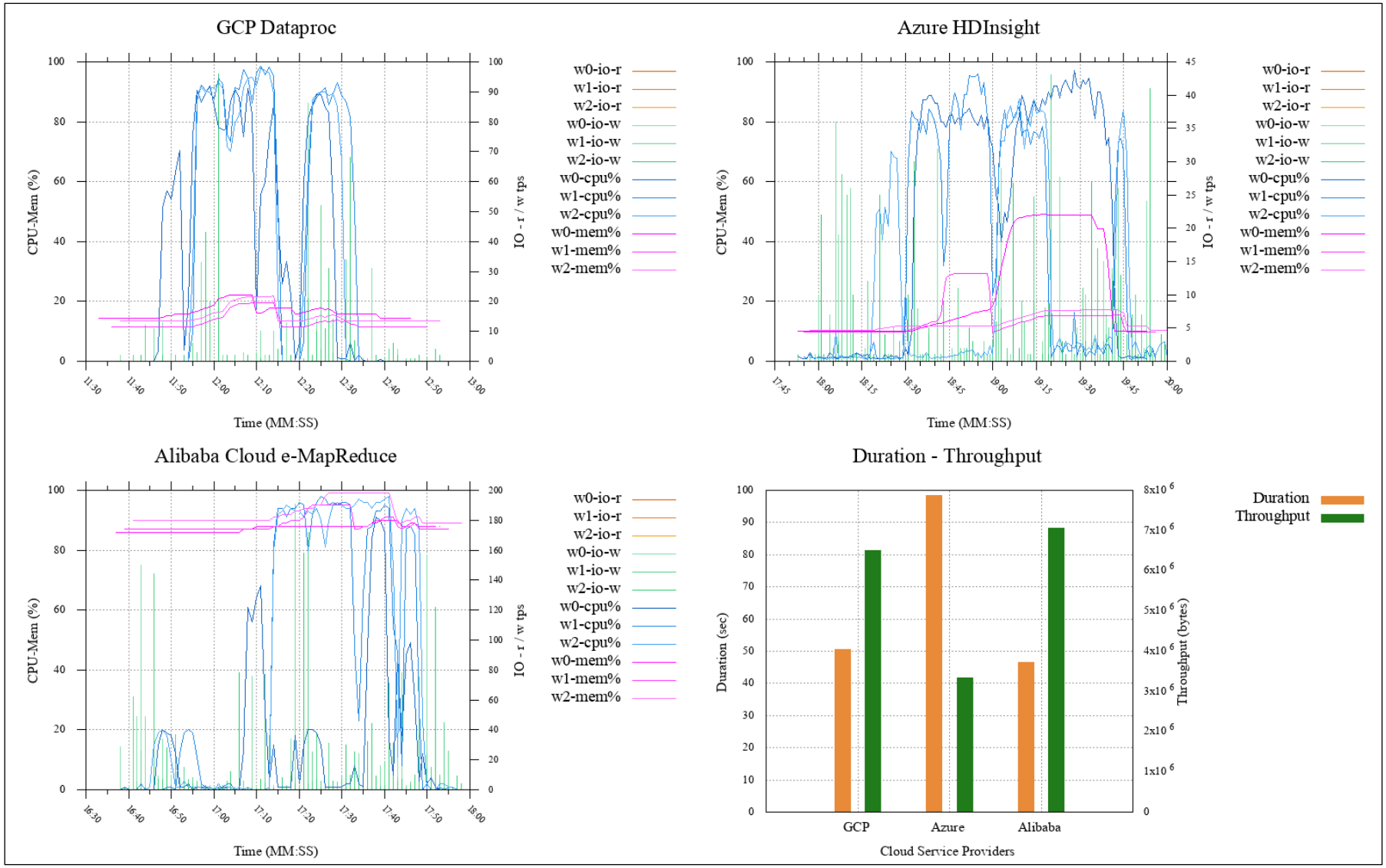
USE CASE 2 - Sort (Gigantic; 32 GB)



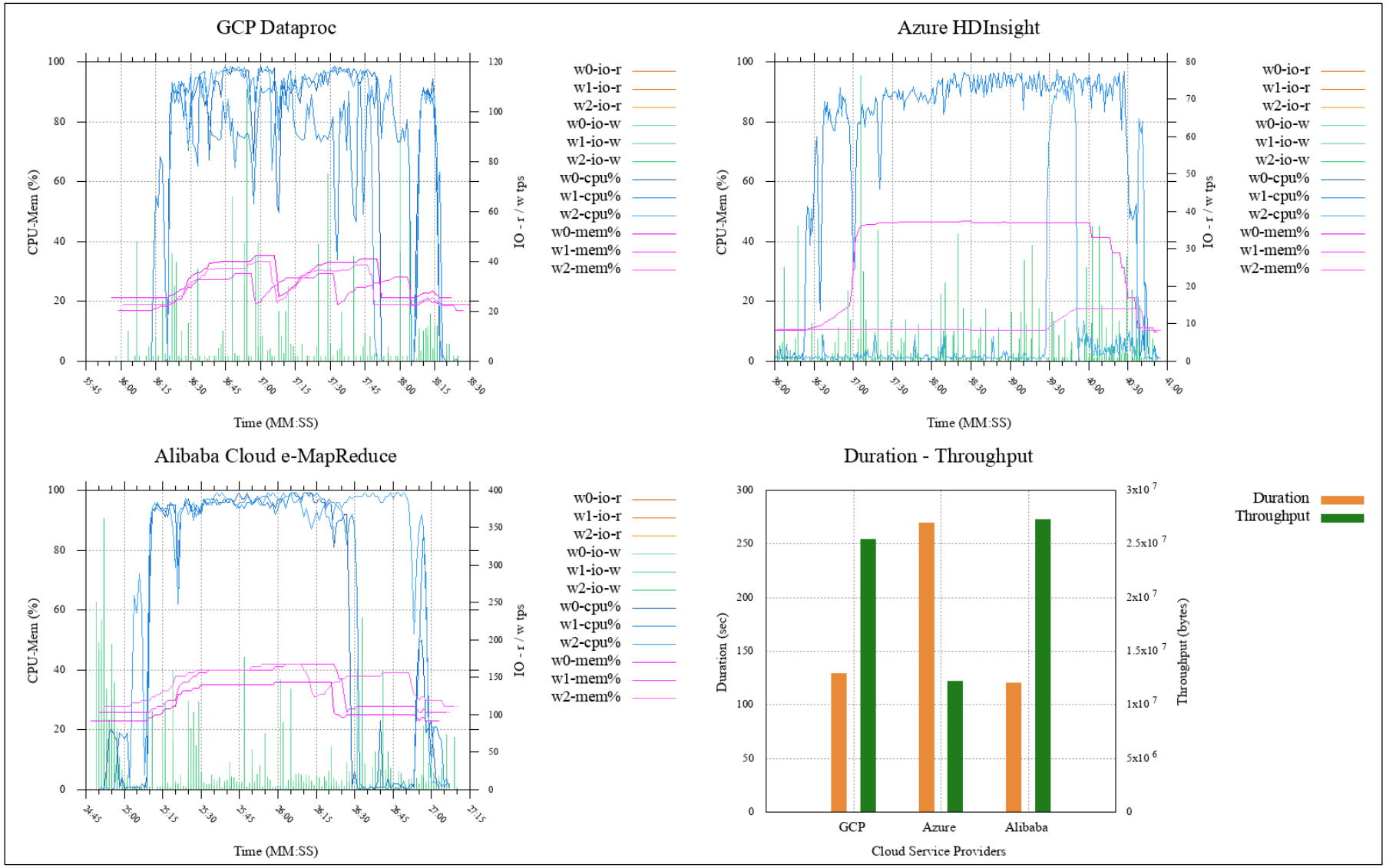
USE CASE 2 - Wordcount (Tiny; 32 KB)



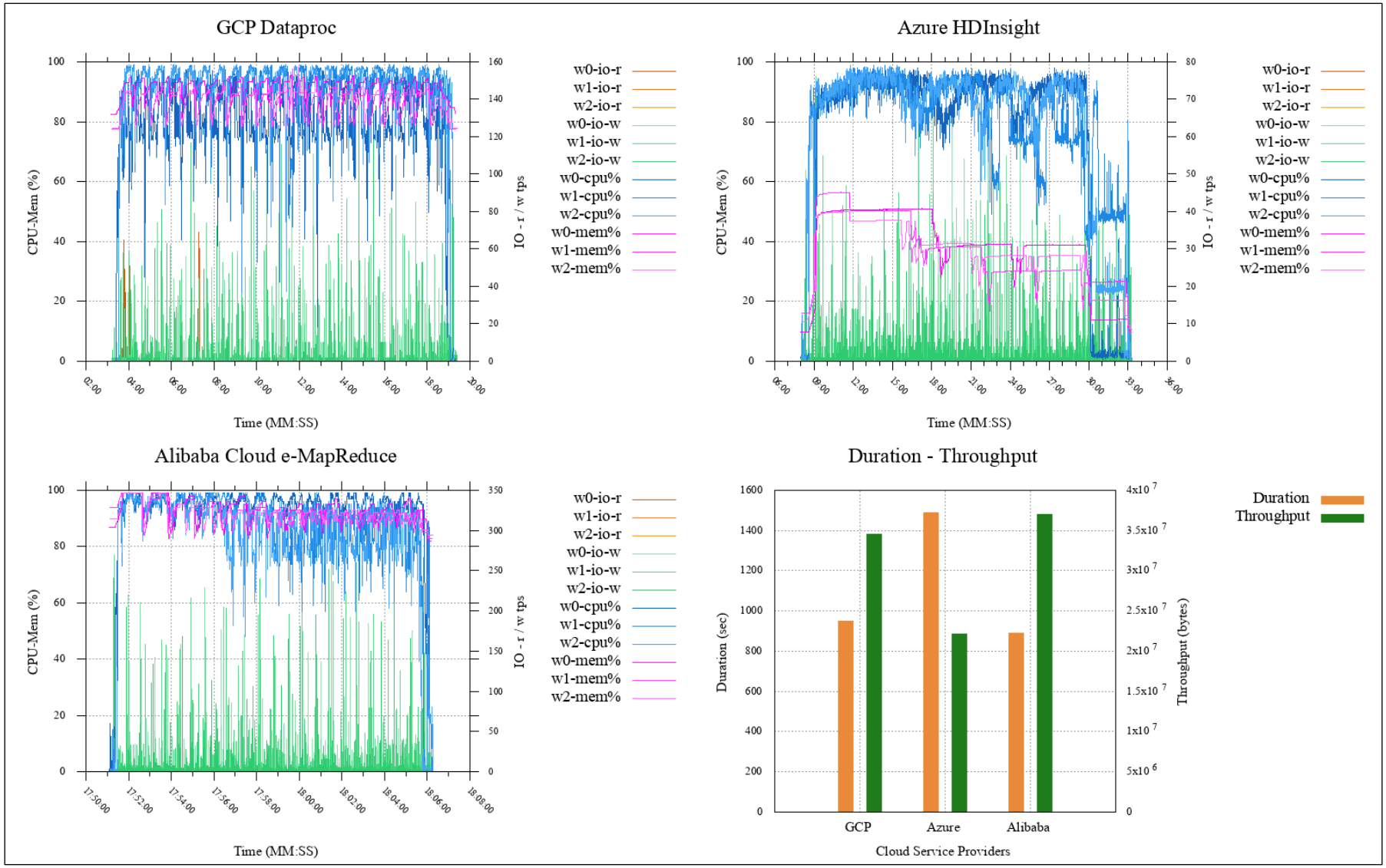
USE CASE 2 - Wordcount (Small; 320 MB)



USE CASE 2 - Wordcount (Large; 3.2 GB)



USE CASE 2 - Wordcount (Huge; 32 GB)



USE CASE 2 - Wordcount (Gigantic; 320 GB)