
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

CMP216

Real-time compute workloads using Amazon EC2 storage optimized instances

Anuj Gupta (he/him)

Principal Solutions Architect
AWS

Karsten Ploesser (he/him)

Principal Solutions Architect
AWS

Derek Feriancek (he/him)

Senior Performance Engineer
Splunk Inc.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Learning objectives

Understand characteristics of real-time workloads

How these characteristics influence instance selection

EC2 storage optimized instances overview

TCO of upgrading to the new 4th-generation instance

Splunk’s journey from I3/I3en to Im4gn/Is4gen

Q&A and key takeaways

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What this session does not cover

- Deep dive into storage optimized instances architecture
  - You can attend the breakout session:
    CMP312: Run high-performance storage workloads on EC2 storage optimized instances
- How to enable your software to run on AWS Graviton

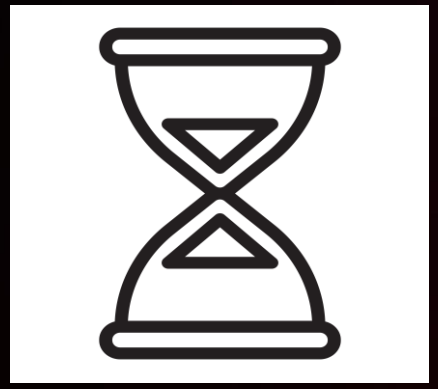
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Real-time workloads

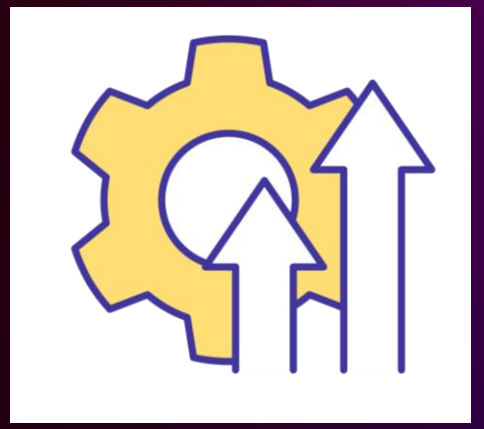
CHARACTERISTICS AND USE CASES



Event to system response within specific time constraints



Time constraints (less than 20 milli seconds)



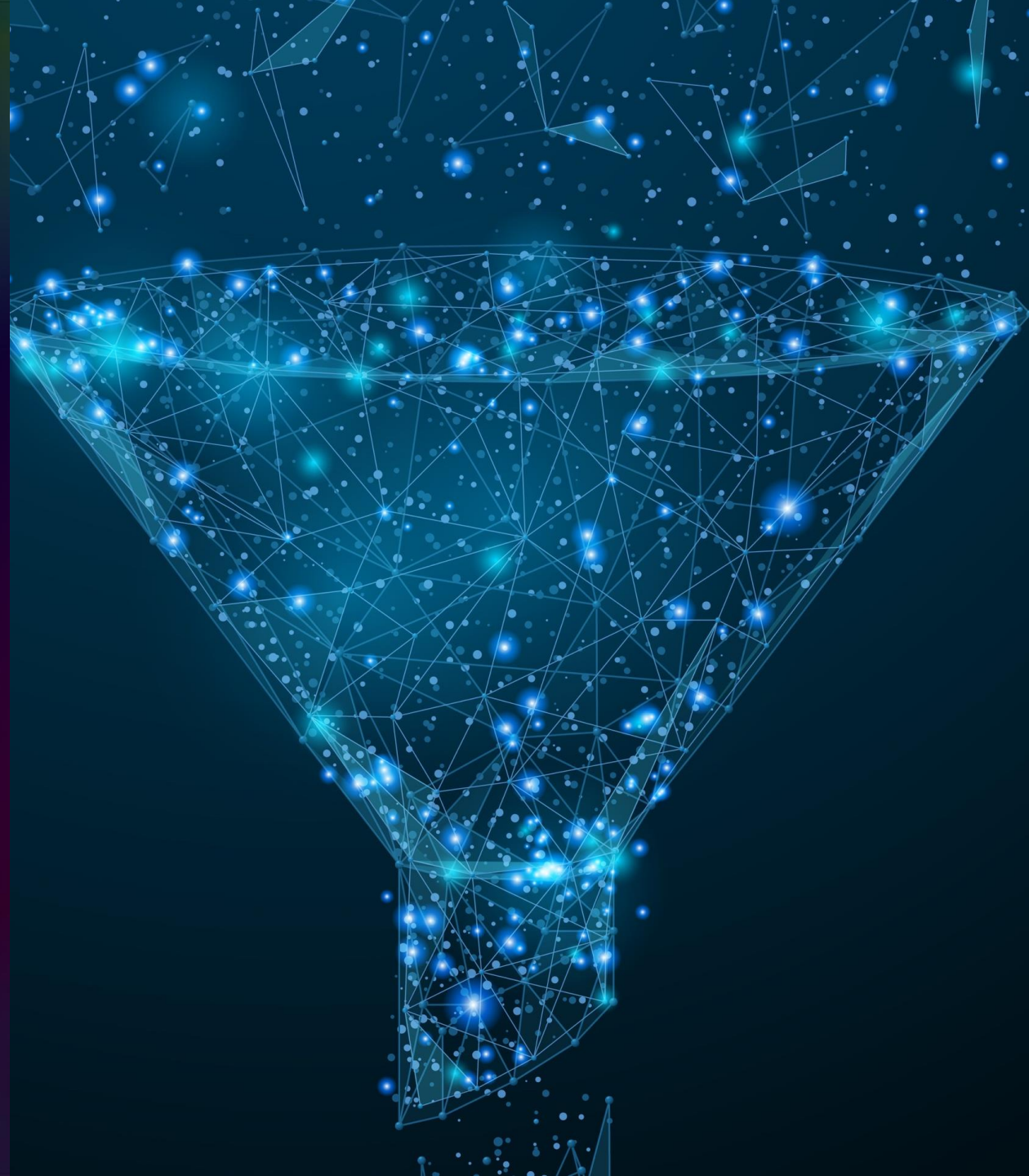
Time constraint should be met regardless of system load (millions of transactions per second)

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Real-time data processing workloads

- Consistent performance at petabyte scale
- High read and write disk bandwidth
- Low latency variability
- Highly fault tolerant



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

EC2 storage optimized instances

HIGH IOPS/THROUGHPUT, LOW LATENCY POWERED BY NITRO SSDS

	I4i	Im4gn	I3en	Is4gen
Processor	Ice Lake (intel)	AWS Graviton2 (aws)	Cascade Lake (intel)	AWS Graviton2 (aws)
vCPU:Mem (GB)	1:8	1:4	1:8	1:6
vCPU:Disk (GiB)	1:234	1:468	1:625	1:937
Mem:Disk (GiB)	1:29	1:117	1:78	1:156
$/TB (OD)	$0.366	$0.194	$0.181	$0.154
Target workloads	Small to medium datasets, NoSQL databases, transactional DBs, data analytics, search	Medium datasets, NoSQL databases, RDBMS, data analytics, and search engines	Large datasets, NoSQL databases, scale-out transactional databases, distributed file systems, data warehouse, indexing, and analytics workloads	Large datasets, NoSQL databases, search engines, data streaming, and large distributed file systems

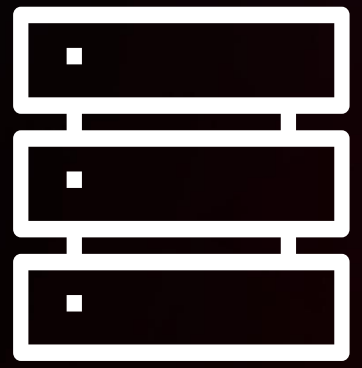
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

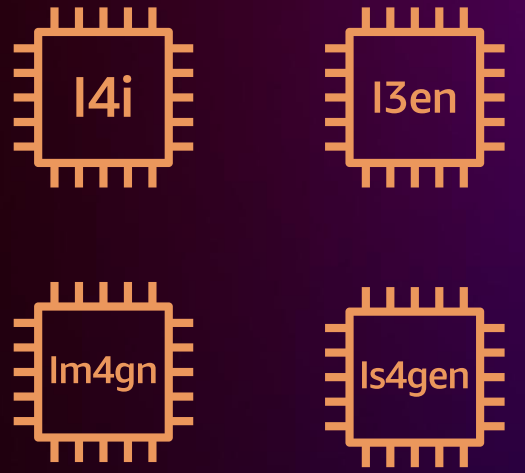
Identify the right instance for your workload



Workload simulation/
performance testing



Memory, disk,
and CPU ratios



Instance types and
cluster sizing

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example migration path (1:1 mapping)

	3rd-gen x86		4th-gen Graviton2	
Instance name	i3.8xlarge		im4gn.8xlarge	
vCPU	32	1:1	32	Same
Memory	**244 GiB**	2:1	**128 GiB**	Lower!
Storage	7,600 GB	1:2	15,000 GB	Higher
$/TB (RI)	$0.229	3:2	$0.123	Lower

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

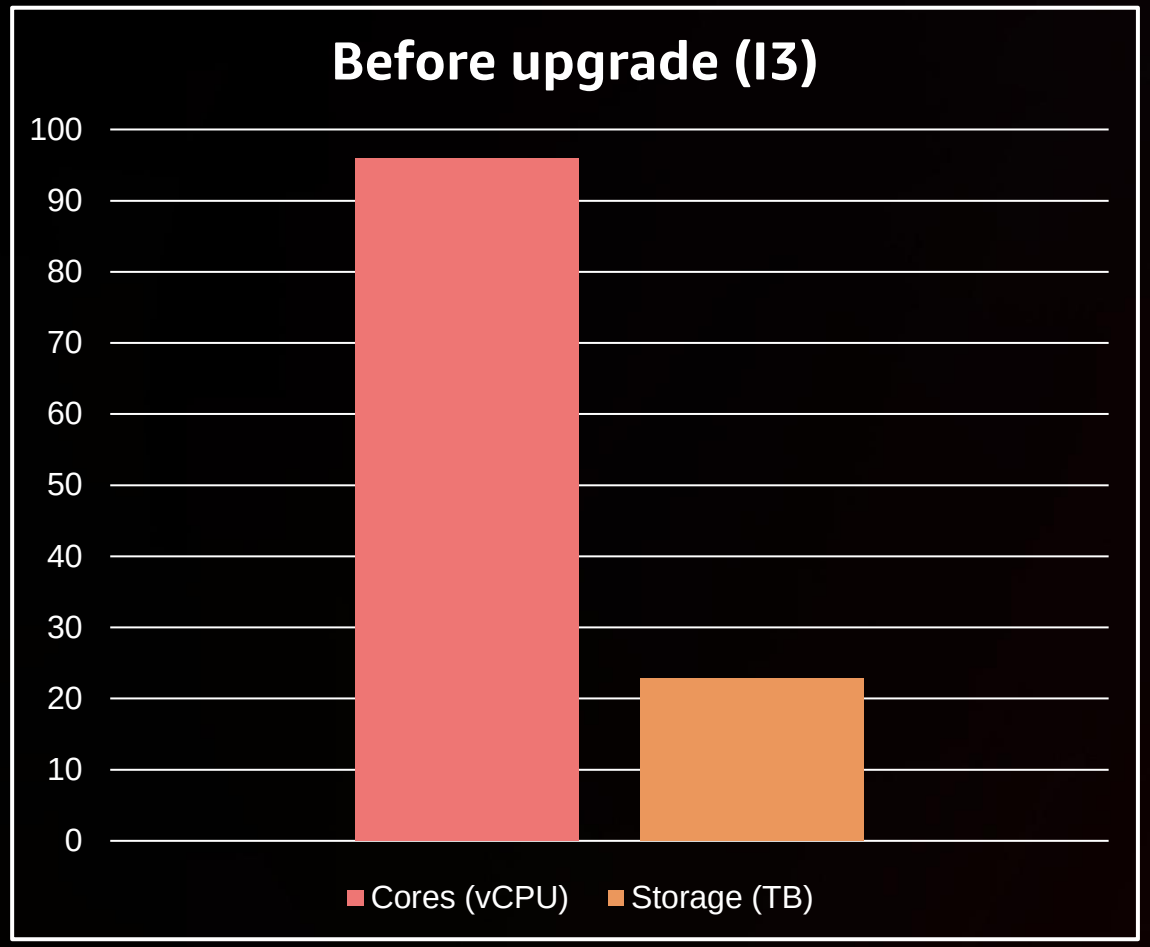
Refine your assumptions (moving to *n* :1)

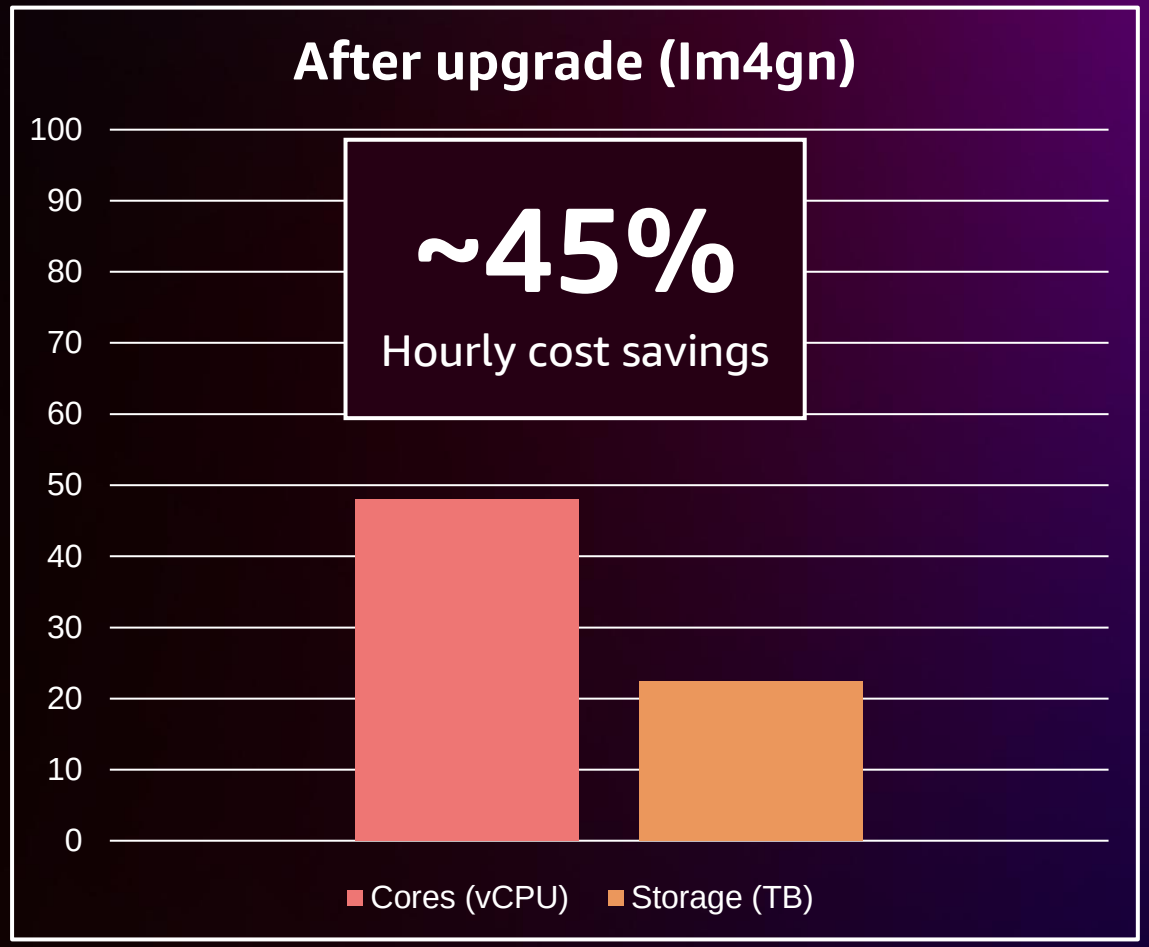
	3rd-gen x86		4th-gen Graviton2	
Instance name	i3.8xlarge		im4gn.4xlarge	
vCPU	32	2:1	16	Descale
Memory	**244 GiB**	4:1	**64 GiB**	Lower!
Storage	7,600 GB	1:1	7,500 GB	Same
$/TB (RI)	$0.229	3:2	$0.123	Lower

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Effects of upgrading to 4th generation (I4)

USE CASE: CLUSTERED DEPLOYMENT OF 3 X I3.8XLARGE NODES UPGRADED TO 3 X IM4GN.4XLARGE





aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Detailed TCO savings analysis

Combination of performance-related *scale-down* and *scale-in*

TCO attribute	Baseline	Conservative	Optimistic
Node size	I3.8xlarge	Im4gn.4xlarge	Im4gn.4xlarge
Cluster size	3	**4**	**3**
vCPU/node	32	16	16
DRAM/node	244	64	64
GB/node	7,600	7,500	7,500
Total vCPU	96	64	48
Total GB	22,800	30,000	22,500
Total $ (RI hourly)	$5.142	$3.733	$2.800
vCPU footprint		**33%**	**50%**
Cost savings		**27%**	**45%**

Mappings are the result of performance testing

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How Splunk uses Im4gn and Is4gen to turn data into insight and reduce cost

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why Splunk uses storage optimized instances

- Splunk has a heavy read/write load profile
  - Bandwidth and latency requirements
- Disk storage as a cache for working set of customers' data
  - Analogous to RAM and disk
- Hefty network and CPU usage
- Often CPU-limited

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How we evaluated Im4gn/Is4gen

- Evaluated a combo of scale-in and scale-down
- Customer representative benchmarks translate what AWS enhancements mean **for you**
- Calculated potential savings
- Optimized for key metrics
- Segmented customer base
  - CPU: Im4gn
  - Dense storage: Is4gen

Performance white paper

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Lessons learned

- Don’t overanalyze spreadsheets and ratios – test and be data-driven
  - Effective CPU strength to memory ratio
  - Memory issues with jemalloc
- Upgrade your tools
- Make use of telemetry
- Treat performance testing as iterative process

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Questions & answers

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways

1. Identify key characteristics of real-time workloads
2. Use workload simulation to measure performance
3. Pick the right instance for optimized price/performance
4. Reduce TCO by scaling in and/or scaling down

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Anuj Gupta
linkedin.com/in/anuj-gupta-24795310/

Karsten Ploesser
linkedin.com/in/karstenploesser/

Derek Feriancek
linkedin.com/in/derek-feriancek-336b68165/



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.