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

CMP315

Turbocharge your workloads using the high-frequency x86 processor on AWS

Sudeep Sharma
Principal Product Manager
AWS

Allan Carter
Principal Specialist SA
AWS

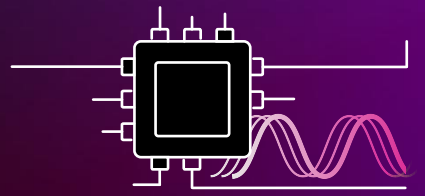
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

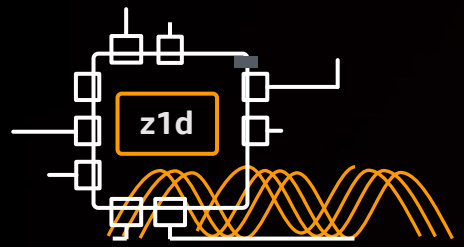
- High-frequency family overview
- New high-frequency instance
- Dive deep into EDA workloads
- Save costs using high-frequency instances

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

High-frequency family overview



2018



Frontend EDA

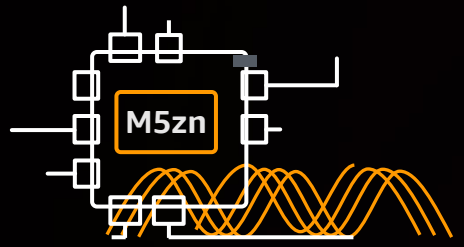


Financial modeling



4.0 GHz all core Turbo, 1:8 compute-to-memory ratio

2020



Gaming

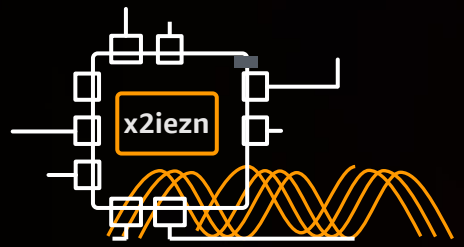


Simulation modeling

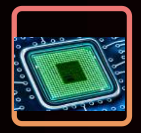


4.5 GHz all core Turbo, 1:4 compute-to-memory ratio

2022



Backend EDA



Relational databases



4.5 GHz all core Turbo, 1:32 compute-to-memory ratio

aws

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

Customer quotes



"We used Amazon EC2 M5zn instances and compared to C5d instances, we experienced performance improvements of 25% for data indexing and 20% on query/search. With these improvements, ChaosSearch is able to reduce the total number of instances for customers at scale, enabling them to save on their overall costs." -Thomas Hazel, ChaosSearch founder and CTO



"NXP Semiconductors, N.V. announced they are moving their EDA to the cloud. We worked with AWS to run our Calibre and Spectre workloads on Amazon EC2 X2iezn instances, which feature the fastest Intel Xeon Scalable processors in the cloud, with an all core turbo frequency of up to 4.5GHz. We measured 10-15% higher performance using X2iezn instances compared to our on-premises, Xeon Gold 6254 with max turbo frequency of 4.0GHz."
-Maarten Kelderman, Director R&D IT Design Infrastructure

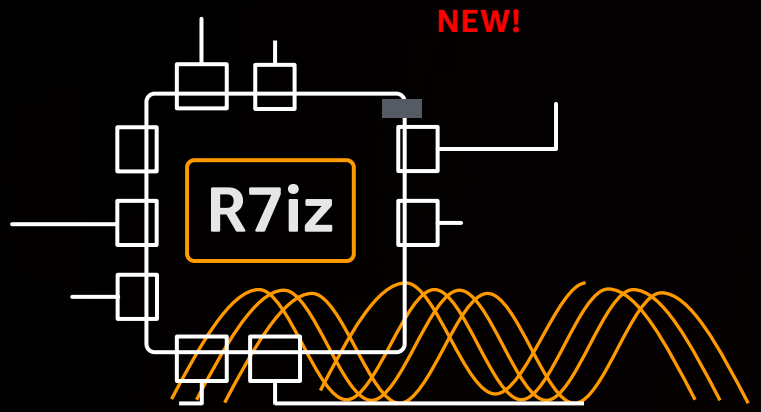


"Synopsys customers leverage our tools in the cloud to accelerate delivery of next-generation products. Synopsys IC Validator is one such tool that provides highly scalable high-performance physical verification signoff. On Amazon's new EC2 X2iezn instances, we achieved 15% performance improvement, scalability to 1000s of cores, and 30% better efficiency using IC Validator's unique elastic CPU management technology versus R5d instances."
-Christen Decoin, VP Business Dev, Custom Design and Manufacturing Group, Synopsys Inc.

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

R7iz – z1d's generational refresh

EC2's first Sapphire Rapids Platform and x86 DDR5 platform

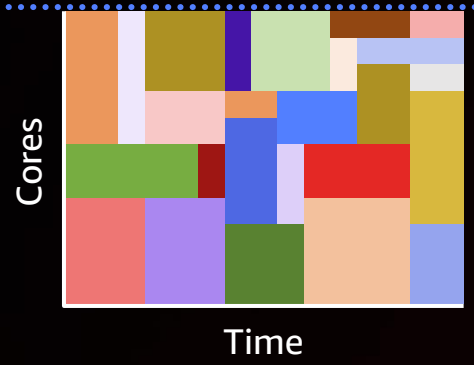


- Custom 4th Generation Intel Xeon Scalable Processor (Sapphire Rapids) offering up to **3.9 GHz** all core turbo frequency
- Highest performance per vCPU among x86 EC2 instances
  - Up to **20%** higher SpecInt17 performance vs z1d
  - Up to 40% higher performance for databases and audio video processing
- **2.6x** larger sizes
  - up to 32xlarge (128vCPUs) vs. 12xlarge (48vCPUs) for z1d
  - Up to 1024 GB vs. 384 GB for z1d
- Two bare metal sizes
- **2x** networking and EBS bandwidth:
  - Up to 50 Gbps of networking BW
  - Up to 40 Gbps EBS BW

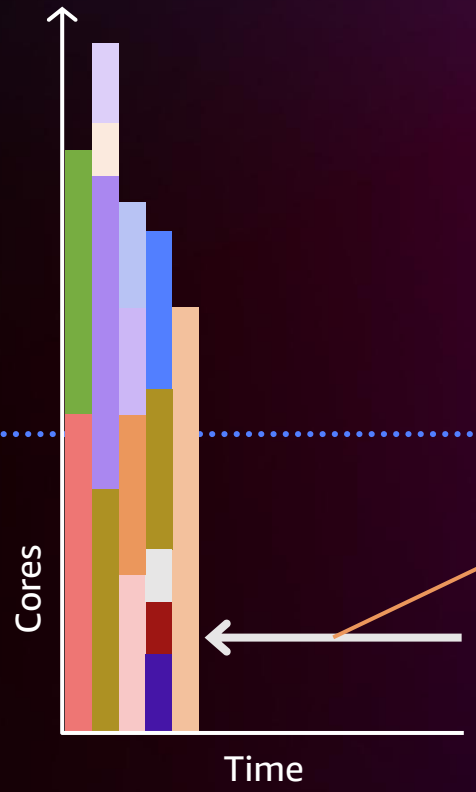
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cloud is a natural fit for semiconductor design

Fixed data center capacity limit



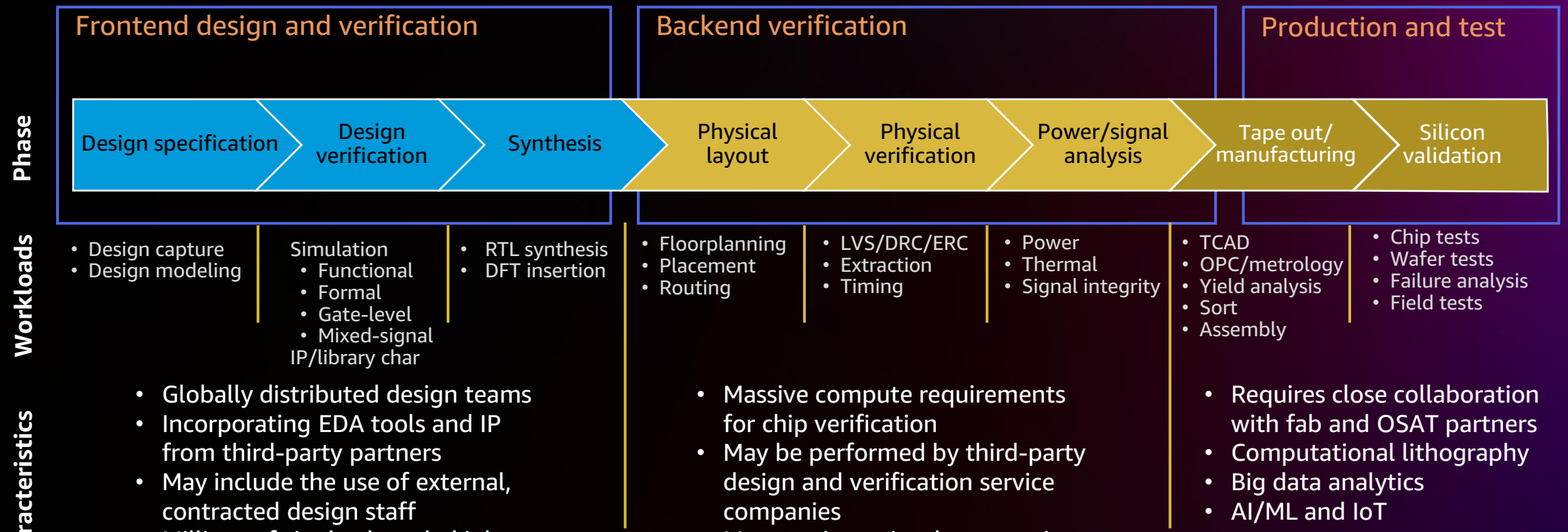
- Static capacity
- Long queue times
- Engineering bottlenecks



- Variable capacity
- No waiting for resources
- Jobs complete faster

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

Semiconductor design flow



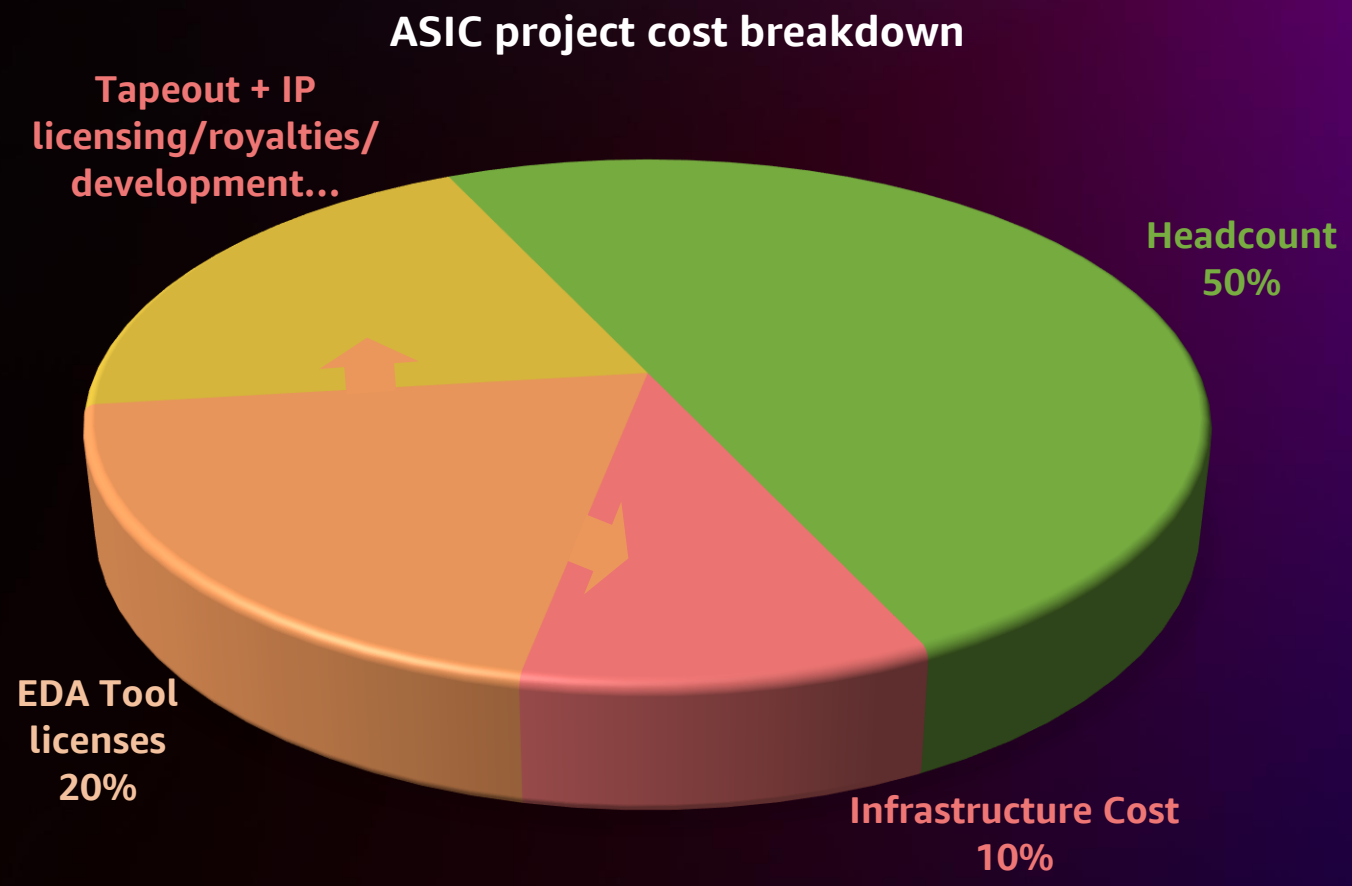
Semiconductor flows require high-performance computing and storage and orchestration of many diverse tasks with complex dependencies

aws

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

Typical 5nm project cost breakdown

- Advanced node ASIC design
- 18- to 24-month project cycle time



Optimizing EDA job execution can have a sizable impact on EDA Tool ROI and overall project TTM and cost

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

EDA license and productivity optimization

- EDA licenses are a significant expense in silicon design environments
  - Often much greater than annual spend on servers
  - Intel – *"Typically, we spend $4 for EDA licenses for every $1 spent on servers"* (https://intel.ly/3hdiA9k) March 2, 2022
- Optimizing EDA tool performance is critical in order to reduce costs and improve productivity
- Key mechanisms for achieving this are
  - Running on optimal instances (latest generation processors with well-matched core/memory)
  - Eliminating time waiting for compute to free up

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

EDA efficiency model on-premises

On-premises infrastructure is **mix of processor generations**, which can have a **major impact** on EDA application performance

Intel has published per-core EDA relative performance across 4 generations of compute for EDA

Intel whitepaper
https://bit.ly/3FHQJIu

Example calculation: EDA spend = $10M, CPU distribution as shown in table

Assumptions: (1) 100% core utilization; (2) No scheduler wait time; (3) No servers "down"; (4) EDA jobs are CPU intensive

	Current gen	N-1 gen	N-2 gen	N-3 gen
Core distribution	30%	30%	20%	20%
Perf vs current gen	100%	90%	65%	53%
Equivalent current gen cores	30%	27%	13%	11%

Total current gen equivalent cores = 81%
Total EDA value realized = $8.1M
Loss of $1.9M due to running on older-generation servers

Cost eliminated (or reversed) when running on latest processors on AWS

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

Thank you!

Sudeep Sharma
svsharma@amazon.com

Allan Carter
cartalla@amazon.com



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

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