AWS Whitepaper

Semiconductor Design on AWS

Semiconductor Design on AWS: AWS Whitepaper

Table of Contents

Semiconductor Design on AWS

Publication date: **March 12, 2021**

Abstract

Semiconductor and electronics companies can significantly accelerate their product development lifecycle and time to market by taking advantage of the near-infinite compute power, storage, and resources available on Amazon Web Services (AWS). This whitepaper presents an overview of the semiconductor design flow, a migration path for moving design and verification workflows to AWS, and the AWS architectural components needed to optimize semiconductor design workloads on AWS.

Introduction

The workflows, applications, and methods used for the design and verification of semiconductors, integrated circuits (ICs), and printed circuit boards (PCBs) have been largely unchanged since the invention of computer-aided engineering (CAE) and electronic design automation (EDA) software. The computing requirements, however, have dramatically increased as device geometries have shrunk and electronics systems and integrated circuits have become more complex. CAE, EDA, and emerging workloads such as computational lithography and metrology have driven the need for massive-scale computing and data management in next-generation connected products.

IT and EDA support organizations face the challenge of providing the infrastructure required to run workflows in a way that meets schedule and budget requirements. They must invest in increasingly large server farms and high-performance storage systems to enable high quality, fast turnaround of workflows. A lot of overall design time is spent verifying components. Workflows like the characterization of intellectual property (IP) cores, functional verification, and timing analysis have a spiky demand and limit engineering productivity. This creates a need to have enough compute capacity to minimize the time that engineers wait for results, but can result in underutilization of resources between runs of the workflows. New and upgraded IC fabrication technologies have increased peak compute and storage requirements, challenging organizations to find ways to meet the needs of silicon development teams while managing costs.

Semiconductor companies today use AWS to deploy infrastructure on an as-needed basis with pay-as-you-go pricing. The massive scale of AWS enables them to run CAE and EDA workflows as quickly as possible with no upfront capital expenditures or long term commitments. They benefit from a more rapid, flexible deployment of CAE and EDA infrastructure to run the complete IC design flow, from register-transfer-level (RTL) design to the delivery of GDSII files to a foundry for chip fabrication. AWS gives them access to the latest compute, storage, network technologies, and higher-level services that enable them to meet the ever-increasing demands of semiconductor design workloads so they can innovate faster.

Companion implementation guide

Although this whitepaper provides some level of detail for running semiconductor design workflows on AWS, it does not dive deep into specific architectural details. Instead, AWS provides a companion guide that gives detailed architectural guidance for building your semiconductor design infrastructure on AWS. The companion implementation guide can be found here: Run Semiconductor Design Workflows on AWS.

Semiconductor design overview

Semiconductor design workloads comprise workflows and a supporting set of software tools that enable the efficient design of microelectronics, and in particular, semiconductor ICs. Semiconductor design and verification relies on a set of commercial or open-source tools, collectively referred to as EDA software, which expedites and reduces time to silicon tape-out and fabrication. EDA is a highly iterative engineering process that can take months, and in some cases years, to produce a single integrated circuit.

The increasing complexity of integrated circuits has resulted in an increased use of preconfigured or semi-customized hardware components, collectively known as IP cores. These cores (provided by IP developers as generic gate-level netlists) are either designed in-house by a semiconductor company, or purchased from a third-party IP vendor. IP cores themselves require EDA workflows for design and verification, and to characterize performance for specific IC fabrication technologies. These IP cores are used in combination with IC-specific, custom-designed components, to create a complete IC that often includes a complex system-on-chip (SoC), making use of one of more embedded CPUs, standard peripherals, input/output (I/O), and custom analog and/or digital components.

The complete IC itself, with all its IP cores and custom components, then requires large amounts of processing for full-chip verification, including modeling (simulating) all the components within the chip. This modeling, which includes hardware description language (HDL) source-level validation, physical synthesis, and initial verification (for example, using techniques such as formal verification), is known as the *front-end design*.

The physical implementation, which includes floor planning, place and route, timing analysis, design-rule-check (DRC), and final verification, is known as the *back-end design*. When the back-end design is complete, a file is produced in GDSII format. The production of this file is known as *tape-out*. When completed, the file is sent to a fabrication facility (called a *foundry*), which may or may not be operated by the semiconductor company, where a silicon wafer is manufactured. This wafer, containing perhaps thousands of individual ICs, is then inspected, cut into dies that are themselves tested, packaged into chips that are tested again, and assembled onto a board or other system through highly automated manufacturing processes.

All of these steps in the semiconductor and electronics supply chain can benefit from the scalability of the cloud.

Benefits of the AWS Cloud

The benefits of running your semiconductor workloads in the AWS Cloud include improved productivity, high availability and durability, matching your compute resources to your workload requirements, and accelerating your upgrade cycle.

Improved productivity

Organizations that move their workloads to the cloud can see a dramatic improvement in development productivity and time to market. AWS bills by the second for compute resources, so the cost of a batch of jobs is the same whether they run serially or in parallel. By scaling horizontally, a simulation regression can run all of its jobs in parallel, and complete in the time it takes to run the longest job. A check-in regression might complete in 5 minutes instead of 30 minutes, which reduces the time a developer needs to wait to verify a change.

Tape-outs are often gated by the time required to verify late changes. Sign-off verification can be completed in days instead of weeks, and compress the development schedule. The scalability of AWS improves productivity and reduces development time by reducing the time it takes to make changes, find bugs, synthesize logic, analyze timing, characterize libraries, and perform all the tasks required to complete a semiconductor design at no additional cost. These extreme levels of parallelism are common in the AWS Cloud in industries with large computing requirements.

High availability and durability

The AWS Global Infrastructure spans multiple locations worldwide. These locations include Regions and Availability Zones. Each AWS Region is a separate geographic area around the world, such as Oregon, Virginia, Ireland, and Singapore. AWS Regions are designed to be completely isolated from other Regions. Each Region has at least two Availability Zones, each including one or more data centers, that are fully interconnected with a low latency network. As of this writing, AWS has 24 Regions and 77 Availability Zones around the world. This design achieves the greatest possible fault tolerance and stability.

AWS gives you the ability to place resources, such as compute capacity and data storage, in multiple Availability Zones within a region with single millisecond latency between Availability Zones. You can protect against failures and ensure you have enough capacity to run your most compute-intensive workflows by taking advantage of multiple Regions and multiple Availability

Zones. This large global footprint enables you to position computing resources near your IC design engineers in locations where low-latency performance is important. For updated information, see the AWS Global Infrastructure page.

Match compute resources to workload requirements

AWS offers many different configurations of virtual and bare metal servers (such as cores, memory, storage, and network bandwidth) known as Amazon Elastic Compute Cloud (Amazon EC2) instances. Customers choose instance types that match the compute needs of their jobs. Combined with the on-demand nature of the AWS Cloud, you can get precisely the compute infrastructure you need, for the exact job you need to run, and for only the time you need it.

Amazon EC2 instances are available in many different sizes and configurations. These configurations are built to support jobs that require both large and small memory footprints, high core counts of the latest generation processors, and storage requirements from high input/output operations per second (IOPS) to high throughput. By right-sizing the instance to the unit of work for which it is best suited, you can achieve higher performance at lower overall cost. You no longer need to purchase cluster hardware that is entirely configured to meet the demands of just a few of your most demanding jobs. Instead, you can launch instances that match the characteristics of your desired workload into dynamically scalable compute clusters that are uniquely optimized for specific workloads or stages of chip development.

For example, consider a situation where you're developing a critical IP core and need to perform gate-level simulations for a few weeks. In this example, you might need to have a cluster of 100 compute servers (representing over 2,000 CPU cores) with a specific memory-to-core ratio and a specific storage configuration. With AWS, you can deploy and run this cluster, dedicated and purpose-built only for this task, for only as long as the simulations require, and then terminate the cluster when that stage of your project is complete.

Consider another situation in which you have multiple semiconductor design teams working in different geographic regions, each using their own locally-installed EDA IT infrastructure. This geographic diversity of engineering teams has productivity benefits for modern chip design, but it can create challenges in managing large-scale EDA infrastructure; for example, to efficiently use globally-licensed EDA software. By using AWS to augment or replace these geographically separated IT resources, you can pool all of your global EDA licenses in a smaller number of locations using scalable, on-demand clusters on AWS. As a result, you can more rapidly complete critical batch workloads, such as static timing analysis, DRC, and physical verification.

Accelerated upgrade cycle

Another important reason to move EDA workloads to the AWS Cloud is to get access to the latest processor, storage, and network technologies. In a typical on-premises installation, you must select, configure, procure, and deploy servers and storage devices with the assumption that they remain in service for multiple years. Depending on the selected processor generation and time of purchase, performance-critical production workloads might be running on hardware devices that are already multiple years and multiple processor generations out of date. When you use AWS, you can select and deploy the latest processor generations within minutes, and configure your EDA clusters to meet the unique needs of each application in your chip design workflow.

Path to migration

This whitepaper has provided an overview of the industry and benefits of running on AWS. From here, it provides an overview of the path to migrating workflows to AWS, from connecting to AWS, defining proof-of-concept (POC) data and computing requirements, to running an entire production workflow. Additionally, it provides a reference architecture diagram that illustrates the process and provides a high-level view of how you can run your workflows on AWS.

Connect to AWS

AWS offers multiple connection options for infrastructure administration, user access, and data movement. Although you can quickly and easily connect to AWS over a simple internet connection, AWS recommends that customers connect with either AWS Site-to-Site VPN or AWS Direct Connect. AWS Direct Connect is a network service that provides a private, secure, and low-latency connection from your data center, office, or co-location environment to AWS. In many cases it can reduce your network costs, increase bandwidth and throughput, and provide a more consistent network experience than internet-based connections. Both Services can enable secure connections, and AWS Direct Connect can provide a high-bandwidth connection suitable for large amounts of data ingress and egress.

Select and prioritize workloads to move to AWS

You will need to select and prioritize the workloads that you want to migrate to AWS. You may want to migrate a low-risk workflow first to use as a POC. After a successful POC, workloads that are time-sensitive or are constrained by the lack of adequate, on-premises compute resources are good candidates. Frequent data synchronization can be challenging and costly, so AWS suggests starting with workloads that have minimal runtime dependencies on on-premises data.

You may want to prioritize migrating workloads that use on-premises resources that are required for other purposes. AWS has customers that move both front-end and back-end workloads. Some customers run sign-off on AWS because it is critical to their schedule, and they lack the on-premises infrastructure needed to satisfy the resource requirements of these workloads. You might require only one workload to move to AWS, such as a bursty scale-out workload like IP characterization.

Another consideration is the data shared between workloads. You may benefit from keeping workloads that share data in the same location, either on-premises or in AWS.

Define storage requirements and transfer data

After workloads are selected for migration, analyze what data is required by the workload, and the capacity and performance requirements of the storage subsystem. This evaluation process can significantly delay the process of migrating to AWS, so it is recommended that you begin the process of identifying workload data dependencies early.

You also need to analyze whether data can be migrated to the cloud once, or if it needs to be synchronized with the on-premises storage. In addition to increasing workflow turnaround time, copying large volumes of data between on-premises sites and AWS has cost implications. Moving data into AWS is free, but moving data out is not. For these reasons it is desirable to minimize data transfer back and forth between on-premises sites and AWS.

AWS provides several ways to migrate on-premises data to the cloud that will initially store the data in Amazon Simple Storage Service (Amazon S3). Amazon S3 is a secure, high-performance object storage service that provides 99.999999999% of durability, fine grained access control, up to 25 Gbps transfer to EC2 instances (per instance), cross-region replication, data tiering, and more. See Amazon S3 Features and Amazon S3 FAQs for more information.

Although tools and semiconductor flows require POSIX-compliant file systems and do not currently support object storage, S3 can be used as the back-end for creating an AWS-managed, POSIX-compliant file system using Amazon FSx for Lustre. Having the data in S3 enables agility and fast failure. An S3 bucket can be the golden repository of data, and be used to quickly create new file systems or transfer data to EC2 instances so that different storage solutions can easily and quickly be created and evaluated. The data in S3 can also be used for disaster recovery to quickly restore data from job or system failures.

The method of transfer largely depends on how much data you need to move. If you have a relatively small amount of static data and you have a fast, reliable internet connection, you may be able to use your internet connection. Leveraging AWS Direct Connect allows for high-bandwidth ingress and egress to AWS. AWS DataSync makes it simple and fast to move large amounts of data between on-premises storage and Amazon S3, Amazon Elastic File System (Amazon EFS), and Amazon FSx for Lustre over the internet or Direct Connect.

If you have large amounts of library, design, or simulation data that requires an initial one-time transfer, consider using AWS Snowball. AWS Snowball Edge supports 80TB of usable storage, and has a rich feature set that provides edge services and clustering abilities. For additional information, see the “When to use Snowball” section of the AWS Snowball FAQs.

Enable user access and remote visualization

After connections have been established and data is migrated, user access and tool installation can be performed. User access is enabled either with terminal-based ssh sessions or graphical remote desktops. A remote desktop provides the same user experience on AWS as on the user's on-premises infrastructure. Tool engineers and chip designers can submit jobs, manage data, and use GUI-driven interactive tools (such as layout, place and route) in a familiar environment that creates a smooth transition to the cloud.

For remote visualization, you can use existing on-premises commercial solutions such as NoMachine NX or OpenText Exceed TurboX. Optionally, AWS offers a remote desktop service called NICE DCV. NICE DCV is a secure, cloud-native solution that provides a robust user interface for engineering and physical design teams, and performs well over varying network conditions. NICE DCV is provided at no additional cost; you just pay for the EC2 instance that it is running on.

Although the details are beyond the scope of this document, you will also need to install tools and other applications that are required by the engineers on your team. This may involve creating, mounting, and configuring several file systems that are accessible across all Amazon EC2 instances. It is also important to work with your ISV or tool provider to ensure your licensing agreements allow you to run the necessary EDA tools on AWS.

Build scale-out infrastructure

At this point in the process, engineers should have the prerequisites to start launching jobs and analyzing results. The AWS infrastructure that will be launched is a scalable, elastic environment that is capable of running full SOC design workflows.

The next step is to deploy a resource management and orchestration service that enables engineers to submit workloads into a scalable, elastic environment capable of running full SOC design workflows. The top commercial schedulers and workload management systems commonly found in silicon design environments provide options for AWS integration. This allows these tools to dynamically scale the EC2 computing resources based on demand in the queues. If your current scheduler does not support AWS integration or you are looking for a turnkey solution, AWS recommends using the official AWS Solution, Scale-Out Computing on AWS, to automate the process of building your environment. The solution deploys a web-based user interface (UI) and automation tools that enable you to create your own job queues, auto-scaling compute resources, shared file systems, remote desktop sessions, and other components you need to run silicon design

workloads at scale in AWS. As part of launching the solution, customizations should be made to include the previously-installed tools and POC data, as these will be needed to launch jobs.

Run POC workflow

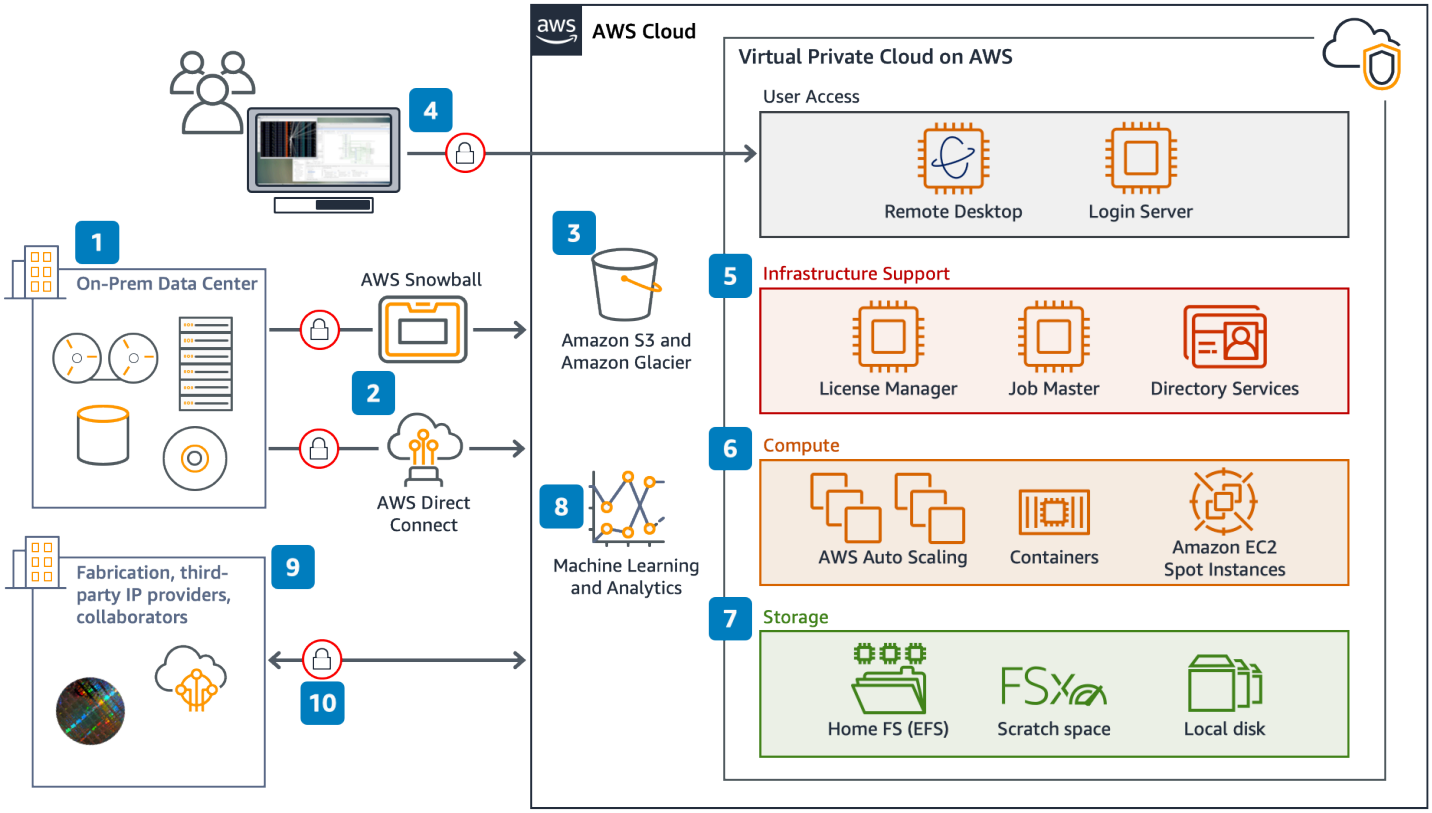
After the environment is launched and engineers are able to access both tools and data, the POC workload should be run. AWS recommends starting with a simple test to verify functionality and results. From there, you should scale-out to a large number of cores and begin to take advantage of the elasticity of AWS. With on-premises systems you have a fixed number of resources; on AWS you have nearly infinite resources that enable you to scale jobs quickly and significantly reduce your runtime. The results of the POC workload should be compared with on-premises results, to both validate results (consistency) and analyze runtimes.

Migrate additional workloads to AWS

After the POC is successful, the next step is to start migrating the workloads that you identified earlier. Update your migration plan based on key learnings from the POC. After you have gone through the process of moving one workload to AWS, much of the initial work you performed can be leveraged for your remaining workload migrations.

Reference architecture

The previous section provided a path to migration. This section provides an annotated reference architecture diagram that reinforces these concepts. For a deep dive into the architecture and specific AWS Services that should be used, refer to Run Semiconductor Design Workflows on AWS.



Reference architecture diagram depicting semiconductor design on AWS

Architecture diagram descriptions

1	Determine what data is needed for proof of concept or test.	**6**	AWS compute is flexible and robust, more than capable of running semiconductor design workflows.
2	Transfer data into AWS via **AWS**	**7**	Store tools and job data on Amazon

Architecture diagram descriptions

	Snowball, **AWS Direct Connect**, or using several other AWS services.		EFS, Amazon FSx for Lustre, and local disk. Optionally, move long-term data storage to Amazon S3.
3	Transferred data is stored in **Amazon S3** buckets. You can access data stored in Amazon S3 from an **Amazon EC2** instance or nearly any AWS service.	**8**	Once your data is in AWS, you can leverage other services, such as data lakes, AI/ML, and analytics.
4	Users access their environment through a remote desktop session or command line (ssh).	**9**	Isolating environments leads to enhanced security and limits third parties to only the data they need.
5	All of the infrastructure needed for semiconductor design workflows is available on AWS.	**10**	Encryption is everywhere and can be enabled with your encryption keys.

Collaboration enabled by AWS

Across the semiconductor industry, collaboration is part of the design process, fabrication, and product manufacturing. AWS enables you to securely collaborate with third-party IP providers, EDA tool vendors, foundries, and contract manufacturers. For example, you might have a requirement to work with a third-party IP provider or contract engineering team to create or validate a portion of your system-on-chip (SoC). Using AWS for collaboration makes it possible to segregate roles and data, lock down the environment to only authorized users, and monitor activity in the environment.

When trying to create similar collaborative environments in your on-premises data center, you might have the ability to isolate users and groups through existing network policies; however, you are still allowing external access to your internal infrastructure, and the collaboration environment is not scalable. On AWS, you can set up a completely separate, secure, and scalable environment that enables you to isolate access to only what is needed for the collaborative effort. This can be accomplished in several ways, but usually starts with a separate Amazon Virtual Private Cloud (Amazon VPC) with specific security settings for the level of security and access required. The setup of these secure chambers is covered in detail in Run Semiconductor Design Workflows on AWS.

Resources

There are numerous resources to help you get start with running your semiconductor workloads on AWS. Here are a few links that serve as top-level starting points:

- For a comprehensive list of resources, see the Semiconductor and Electronics on AWS Resources page. On this page you will find:

- Whitepapers
- Blogs and articles
- Reference architectures
- Videos and webinars
- Technical tools and training

- For infrastructure setup and automation, see Scale-Out Computing on AWS.
- For addition information on how AWS can help run your workflows, see the official Semiconductor and Electronics on AWS website.

Conclusion

Semiconductor design teams worldwide benefit from the agility, performance, and security of the cloud for their most critical semiconductor design workflows. AWS can significantly accelerate product development lifecycles and reduce time to market by providing near-infinite compute and storage resources. AWS enables secure, fast-to-deploy environments for design collaboration. Other advantages include access to a wide range of analytics capabilities, including AI services such as yield/failure analysis, and EDA workflow optimization.

Contributors

Contributors to this document include:

- David Pellerin, Head WW Semiconductor, Amazon Web Services
- Mark Duffield, WW Tech Leader, Semiconductors, Amazon Web Services
- Allan Carter, WW Solutions Architect, Semiconductors, Amazon Web Services
- Nafea Bshara, VP/Distinguished Engineer, Amazon Web Services

Document revisions

To be notified about updates to this whitepaper, subscribe to the RSS feed.

Change	Description	Date
Initial publication	Whitepaper first published	March 12, 2021

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.