

Scaling Your Chip Design Flow

Using Google Cloud to accelerate your chip design process

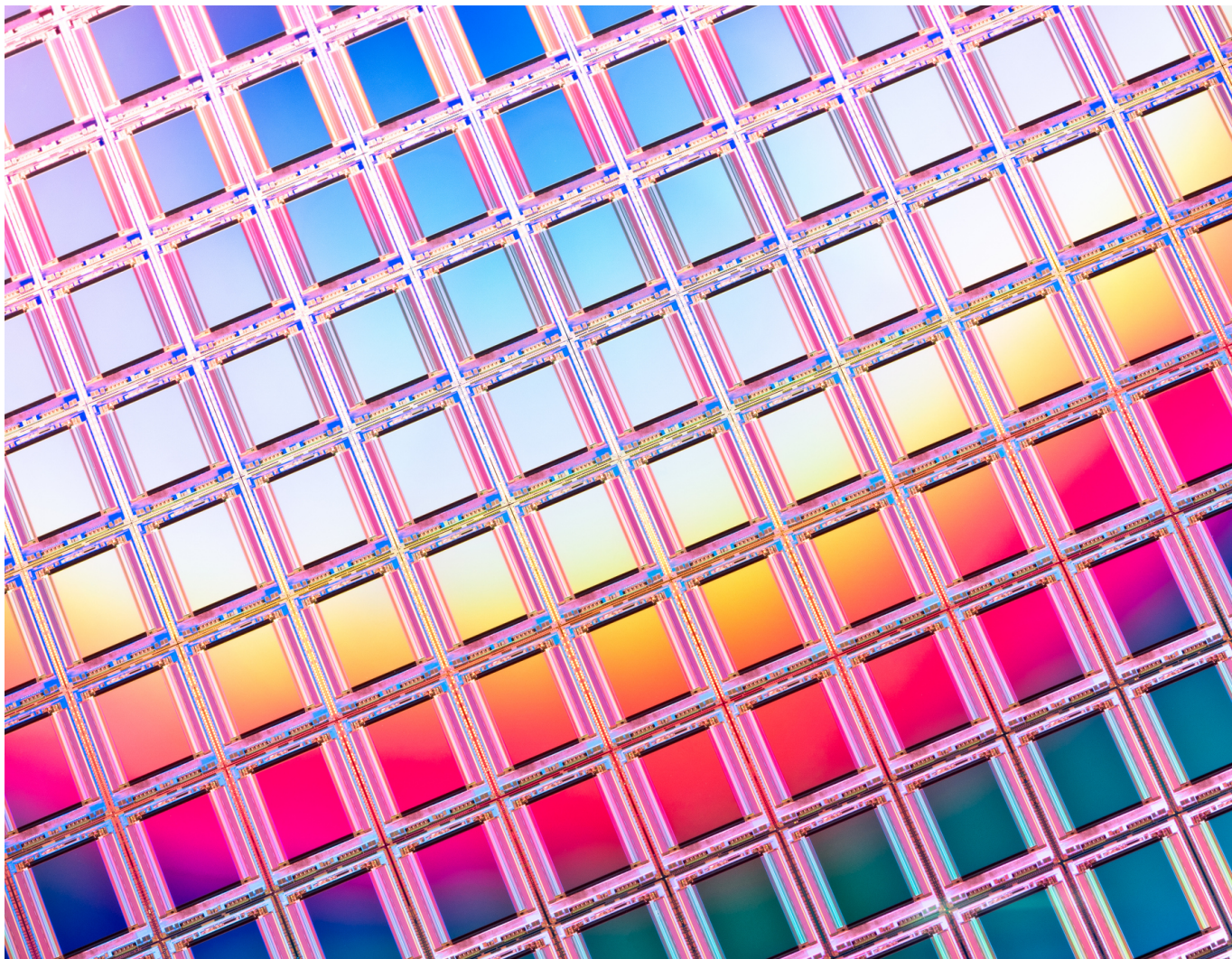


Table of contents

Executive summary.....	1
Chapter 1: Chip design process.....	3
Chapter 2: Designing chips faster.....	6
Chapter 3: How Google Cloud accelerates chip development	9
Best in Class Compute	
Best in Class Storage	
Best in Class Network	
Best in Class Security	
Best in Class Big Data and AI/ML	
Chapter 4: Reference Architecture for EDA Workloads	19
Chapter 5: Making a difference - where it matters	23
Open Source and Innovation	
Sustainability	
Ecosystem	
Cost Optimization	
Conclusion	25
References	26

Executive Summary

As companies undergo digital transformation - a process when an organization embraces new technologies to redesign and redefine its relationships with customers, employees and partners - cloud-driven infrastructure modernization stands out. Companies around the world, from a variety of industries are re-platforming to scale to customer demand, enable higher internal productivity or accelerate innovation. Nokia, the European company well-known for the world's first mobile call in 1991 and several noteworthy mobile innovations recently announced that it will migrate its entire on-premise infrastructure to cloud, in keeping with its cloud-first strategy. This strategic step enables Nokia to expand collaboration and innovation capabilities in its numerous data centers worldwide as it continues to grow its vast product portfolio of 5G, software, mobile and technology products.

The advancements in personal computing, mobile computing, 5G and AI/ML, are enabled by semiconductor electronics. The semiconductor industry has enjoyed steady growth over the years; it is currently valued at hundreds of billions of dollars, and expected to grow robustly in the coming decades. Technologies such as AI, cloud-computing, 5G, Advanced Driver Assistance Systems (ADAS) are all expected to contribute to the growth and limitless opportunities.

The modern day chip design process that accelerated with the invention of the MOSFET¹ over 60 years ago is the place where electronics begins. This process has evolved and benefited significantly from access to powerful computing infrastructure. Gordon Moore's prediction for the number of transistors doubling on chips every 18 to twenty four months held up for a long time. Chip designers were able to keep up with this timeline by retooling and using faster machines. However, technology nodes (connected to the size of transistors in the library used for fabricating the chip and referenced in nanometers or 'nm') are becoming smaller and smaller. For instance, chips were at 28 nm just a few years ago but are now at 7nm and decreasing further (Figure 1), driven by the demands of the several hundred industry use cases in mobile, computing, AI/ML, and others. With decreasing device sizes, the potential for complexity grows exponentially and the essential process of verifying and validating functionality, chip operation variability, timing, low-power requirements, semiconductor manufacturing rules among other checks becomes enormously compute intensive. Lack of access to high performance computing capabilities ultimately delays chip design and production, which can result in significant business losses.

The total cost of designing and manufacturing chips, from the skilled designers to the complex electronic design automation (EDA) tools to the ultra expensive manufacturing costs has ensured that a small number of companies operate in this space, and competition is fierce among them. In this golden age for semiconductors where there will be insatiable demand for AI, automotive, cloud and edge hardware, semiconductor companies who want to succeed will have to continue to design chips faster than their competitors.

¹ Metal-oxide-Semiconductor Field-effect-Transistor, the predominant device used in digital and analog integrated circuits

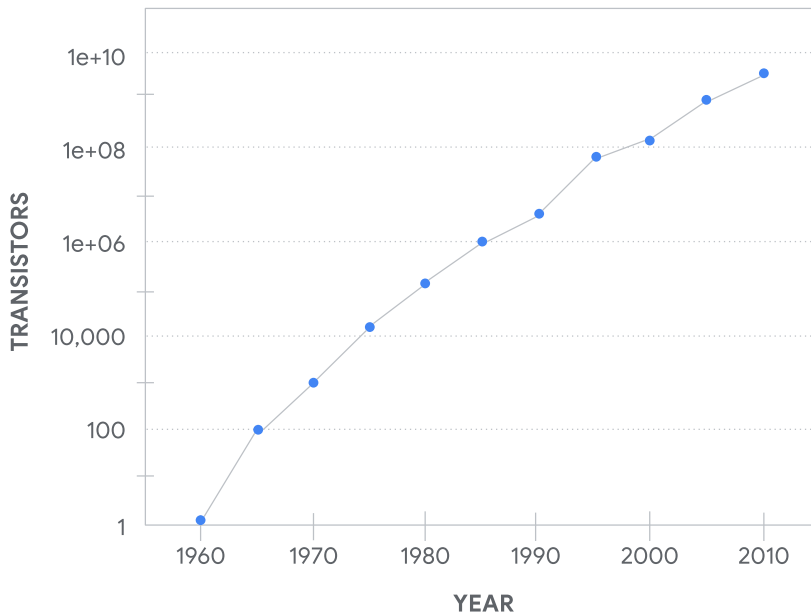
Transistors per die

Figure 1: Increasing transistor density on silicon die
Source: *Computer System Designs: System-on-Chip*, Michael J Flynn and Wayne Luk

Chip design on cloud, where the traditional on-premise datacenter is augmented or replaced by public cloud, provides companies a proven option to accelerate their end-to-end chip design cycle. Public clouds offer elasticity, flexibility, security and scalability not met easily by most on-premise data centers. Several industries have already seen the value of leveraging cloud-based data centers and are reaping business success by 1) delivering products faster, 2) ensuring less downtime including handling unforeseen and unexpected downtimes, 3) reducing costs and 4) enabling innovation.

This white paper details the constantly growing complexity of the chip design process including specific challenges in the workflow, and illustrates with real-world examples how cloud computing provides a way to successfully address bottlenecks and design at scale.

Chip design on cloud, where the traditional on-premise datacenter is augmented or replaced by public cloud, provides companies a proven option to accelerate their end-to-end chip design cycle. Public clouds offer elasticity, flexibility, security and scalability not met easily by most on-premise data centers.

Chapter 1

Chip design process

Like any project, a chip design project starts with a requirements specification. Circuits can be digital (operating at a few defined levels of voltage), analog (operating over a continuous range) or a mixture of both device types. A vast majority of chips designed are digital. Examples include microprocessors, network processors or special Application Specific Integrated Circuits or ASICs.

A simplified version of a typical flow in the digital design flow process is shown below (Figure 2):

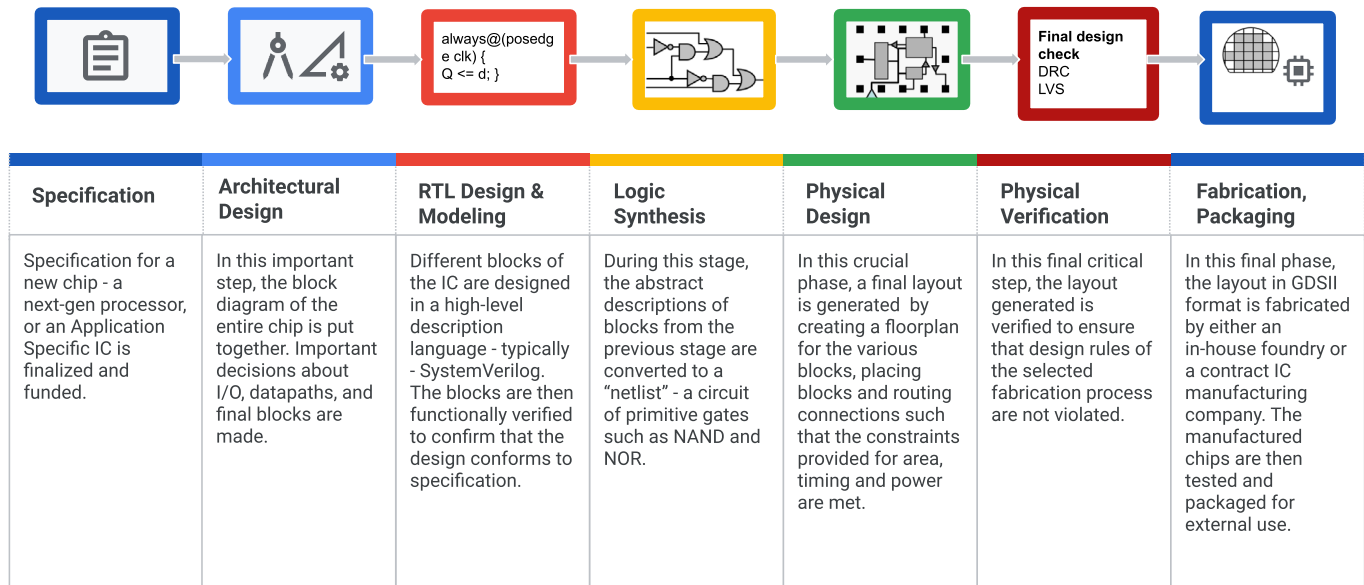


Figure 2: Overview of Digital Integrated Circuit Design Process

Each of the steps specified in Figure 2 can include several EDA tools. Most EDA tools tend to be from commercial enterprises, though it is common for chip design companies to have several home-grown utilities for ensuring a smooth flow between EDA tools in different phases.

"RTL Design & Modeling", "Physical Design", "Physical Verification" and "Fabrication & Packaging" steps are further highlighted below, due to their intensive compute needs.

RTL (Register transfer level) Design and Modeling includes the coding of the design blocks in a high-level hardware description language such as [SystemVerilog](#). Once the design is coded, the next step is to verify that the design conforms to specification. This is one of the most time consuming steps in the design process, accounting for more than half the time needed in the entire design cycle. Mature RTL verification flows include static analysis (checks for design integrity without use of test vectors), formal property verification (mathematically proving or falsifying design properties), dynamic simulation (test vector based simulation of actual design) and emulation (a complex system that imitates the behavior of the final chip, especially useful to validate functionality of the software stack). Large amount of compute is needed for dynamic simulation, although RAM requirements are moderate. Similarly, formal property verification is an [embarrassingly parallel](#) solution and can benefit significantly from availability of inexhaustible compute resources.

Physical Design involves several steps and iterations to ultimately provide a design that meets the constraints specified for power consumption, clock speed and chip area. A physical design process starts with the netlist generated from the Logic Synthesis process (see Figure 3), in addition to the library information relevant to the manufacturing process to be used for the chip. As shown in Figure 3, the complex process of physical design involves steps such as partitioning, creating a clocking network for the chip and connecting the various components together. A successful flow requires several iterations between the steps to meet requirements. Most steps require large high-memory multicore machines and several days per block to reach the expected result. Further, with the advent of ML-enabled engines in EDA tools, tools require access to specialized hardware to build ML models during the flow.

Physical Verification is generally the last step for many companies, before the design is released for manufacturing. In this important step, design rules specified by the manufacturing process are validated. Most EDA tools in the physical design and verification step tend to be distributed and perform better with access to more compute resources.

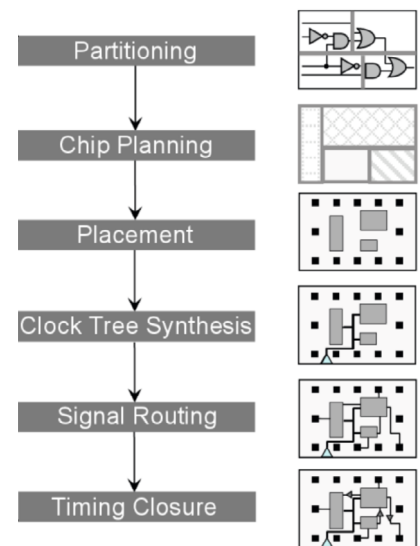


Figure 3 Typical steps in physical design; Source: [VLSI Physical Design: From Graph Partitioning to Timing Closure](#), Andrew B. Kahng, et al

Fabrication and Test involve taking the taped-out² design, processing it to improve manufacturability and fabricating the required design on silicon wafers. When it comes to chip manufacturability, it is important to mitigate issues that impact yield. Yield, in the semiconductor manufacturing context, is the amount of total input transformed into usable product. Semiconductor yield is an aggregate number that includes several types of yield during various stages of production. For example, one type of yield (“die yield”) refers to the number of die (individual design units) that continue to the final testing stage. The impact of yield on a chip vendor is well understood by reviewing the classic yield ramp curve (Figure 4). Yield issues can be process-related, environmental or electrical, and they are impacted by feature size, number of metal layers, and wafer size among a multitude of other issues and factors. Manufacturing processes include hundreds of steps, with each step producing vast amounts of data. This crucial step in the overall product design lifecycle requires large compute, and ability to store volumes of diverse data such that insights from such data are used to make process/parameter changes to accelerate yield improvement.

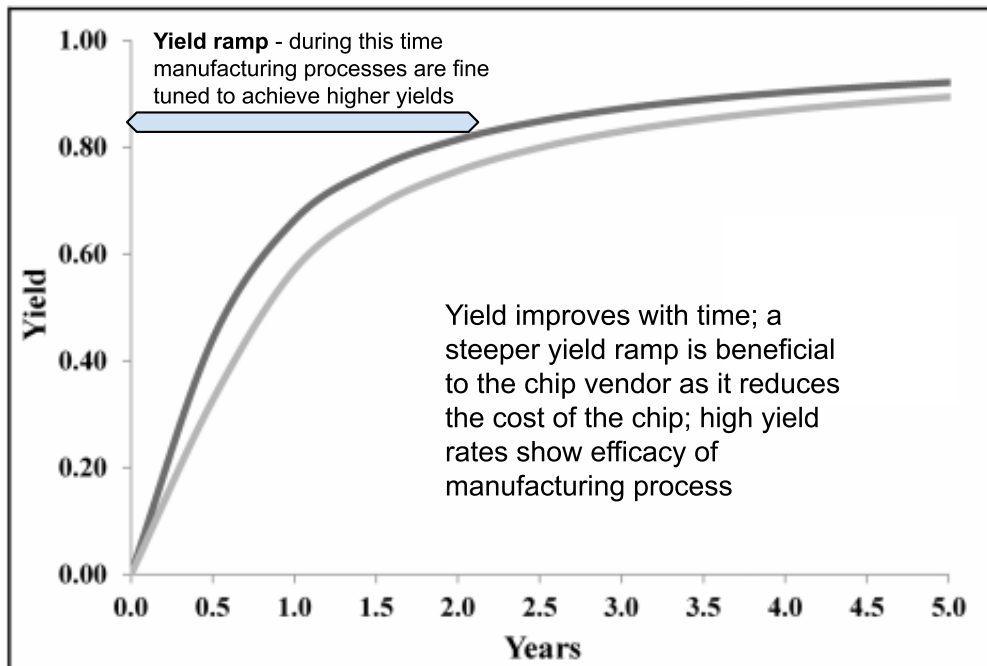


Figure 4: Yield Curve for memory chips, Source: Based on “Yield Learning Curve Models in Semiconductor Manufacturing” by Israel Tirkel

² A term from the early days of electronics when the schematic or circuit design was manually written to tape, disk or CD. In recent times, it is the step where the design team hands off the final chip layout file for fabrication

Chapter 2

Design chips faster

The chip design community is focused on enabling optimizations to design chips faster, such as providing reusable IP (intellectual property blocks), debug utilities to find root cause faster or improved EDA algorithms to improve time to results. Designer productivity solutions have been mostly around better EDA tools, new tools or new ways to improve flows. However, current optimizations are not enough. Each generation and new node requires significantly more processing and analysis, leading to a growing gap between design complexity and designer productivity.

A typical chip can take anywhere from nine to eighteen months from conception to delivery. This time has stayed constant over the years, in part due to improved EDA tools and more powerful computers, but also due to the size of chip design teams which have grown considerably over the past few years to meet time to market needs.

The vast majority of chip design happens on “bare-metal” (machines with no virtualization software) machines located at company premises, or at off-site company-owned data centers. Due to the lead times needed to acquire updated hardware and get access to compute when needed, designers have been unable to leverage the productivity offered by having access to a modern infrastructure that can scale as needed, shrink when not needed and provide the most powerful hardware needed to run complex EDA tools. Tools used for the various steps outlined in the previous sections are compute and memory intensive with peak needs during many phases of the process. The standard process of setting up new design centers at global locations is also impacted by the time needed to build on-premise design centers.

Each generation and new node requires significantly more processing and analysis, leading to a growing gap between design complexity and designer productivity.

Four main issues are significant challenges to productivity:

Elasticity:

In a typical on-premise data center, peak demand is often difficult to meet in a cost effective manner. Often times, the need for additional compute is short-lived, and on-premise datacenters are not built to provide infrastructure for ultra short periods of time. A more prevalent technique is to plan for the additional compute well in advance (typically 8 to 12 weeks) and have it available when the additional compute is needed. Manual provisioning is a significant drain on productivity. Further, once the chip is taped out, on-premise datacenters suffer from significant under-utilization.

Access to state-of-the-art compute:

Like in all industries, EDA tools are constantly trying to leverage the benefits of state-of-the-art hardware architecture. In fact, it is typical for new versions of EDA tools to perform better when a chipset with a larger cache is provided, or if a multi-core capability that allows scaling to several threads is supported. Similarly, several EDA tools perform better with a high-performance network between shared storage and compute. On-premise data centers have upgrade cycles that are 3 to 5 years long - often leading to an environment where the best tool performance is not achieved.

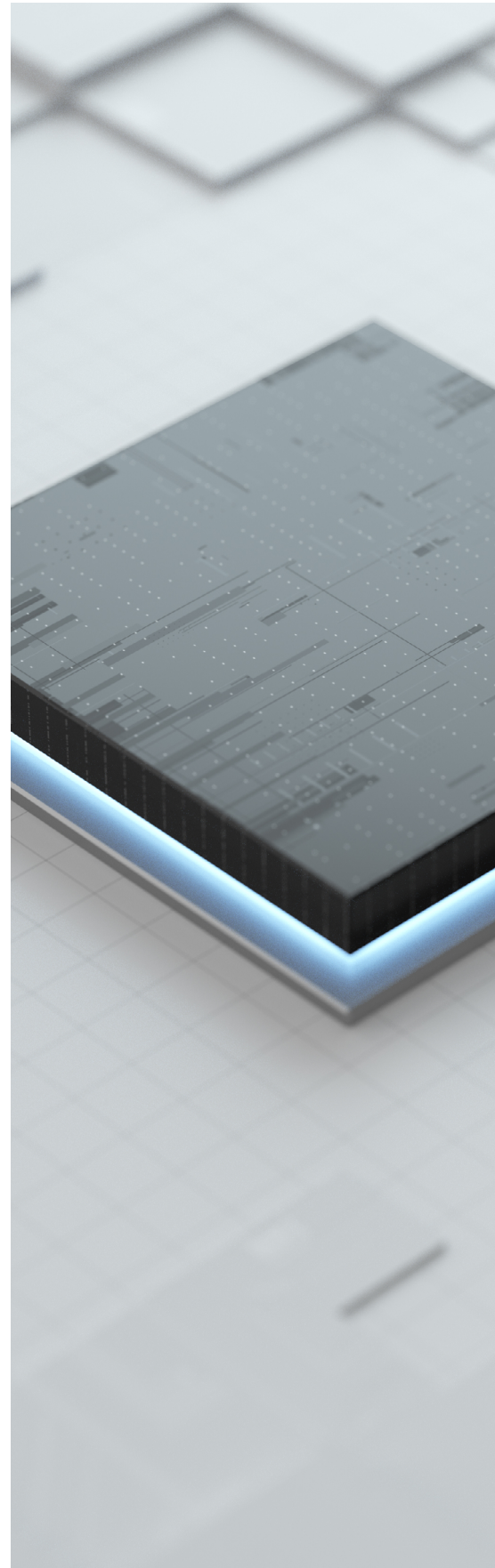
Data-driven insights:

Chip design companies design several generations of a chip, create similar chips and can benefit from sharing knowledge acquired when building new chips. Similarly fabrication plants operate highly complex equipment that generate volumes of data. Learning from data would be vastly beneficial in chip design and manufacturing. However, this requires a systematic process, where successful recipes, tool parameters, bug reports, tool options and similar metrics are saved. An infrastructure that can support the logging and analysis of chip design and manufacturing data is prohibitively expensive to build for a large, multi-site design team, and this prevents design teams from applying advanced analytics, AI/ML models and Big Data techniques consistently and successfully in chip design.

It is typical for new versions of EDA tools to perform better when a chipset with a larger cache is provided, or if a multi-core capability that allows scaling to several threads is supported.

Agility:

The number of skilled engineers needed to design and verify a chip has been growing and companies are increasingly looking to leverage global locations for access to highly skilled engineers and a diverse employee base. Sometimes, starting a new design center is key to business execution success. However, only a few design houses with deep roots in global locations have been successful in such efforts. This ability to start new design centers in a short amount of time is daunting for most companies.



Chapter 3

How Google Cloud accelerates chip development

Chances are that you have already used one of the many Google services (eg: Search, Gmail, Chrome, YouTube) with over a billion users each. Google Cloud is built on the same robust global infrastructure that securely and reliably delivers these services to the world. We offer a complete platform addressing the compute, storage, networking, security, and AI/ML needs of a successful cloud enhanced EDA strategy.

One of the basic services of a cloud is infrastructure as a service (IaaS). In this service, customers customize the infrastructure they need, acquire it almost instantly when needed and have the capability to discard it when not needed. IaaS provides customers with best-in-class compute when they need it - with virtually no limits. Google Cloud is one of the prominent cloud providers with the industry's best IaaS offerings (described more in 'Compute').

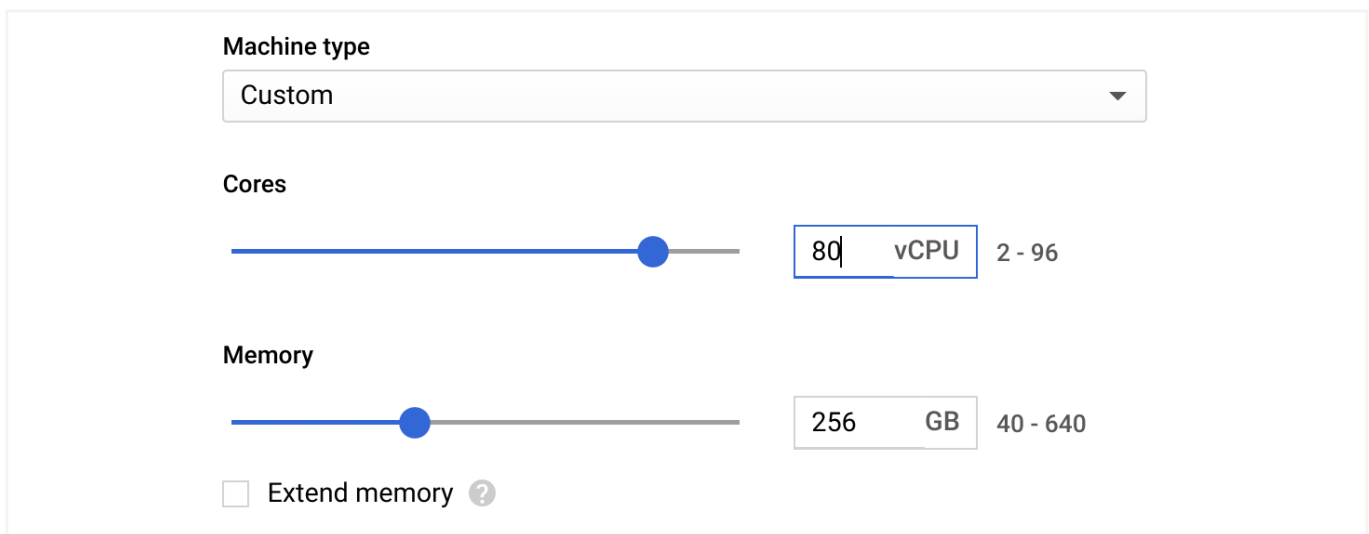
Cloud also offers several types of storage. It is a common misconception that storage on cloud is expensive. On the contrary, cloud offers several tiers of storage, where correct selection of storage for tasks can make the total cost of storage very reasonable for most data center needs. Google Cloud offers low latency storage solutions that can meet the requirements of typical chip design customers.

AI and ML are ubiquitous in every industry, and every company is looking to leverage its data such that it can increase its impact. Google Cloud's Data Management and AI/ML solutions are the most comprehensive in the industry with out-of-box support for several types of data analytics and ML models to help customers build data-driven solutions quickly.

Google Cloud's Data Management and AI/ML solutions are the most comprehensive in the industry with out-of-box support for several types of data analytics and ML models to help customers build data-driven solutions quickly.

Best in Class Compute

Compute is one of the most important components of an EDA workflow infrastructure. Often customers in an on-prem model make decisions to procure compute resources based on the peak. These capex investments are slow to add and frequently underutilized and over-provisioned after peak demands are met. Google Cloud offers unparalleled flexibility in terms of Compute offerings to run the EDA workloads in the most optimized way. Google Cloud offers pre-configured VM in a variety of shapes and offers custom shapes for the ultimate flexibility.



The screenshot shows the 'Machine type' selection interface. At the top, a dropdown menu is set to 'Custom'. Below this, there are two sliders for configuration:

- Cores:** A slider is positioned at 80 vCPU. The range is indicated as 2 - 96.
- Memory:** A slider is positioned at 256 GB. The range is indicated as 40 - 640.

At the bottom, there is a checkbox labeled 'Extend memory' with a question mark icon next to it.

Figure 5: Custom Machine Shape Selection

Google Cloud accelerates EDA workloads with high-performance virtual machines using Google Compute Engine (GCE). Google Cloud has the industry's fastest startup time for provisioning virtual machines. These virtual machines easily integrate with other Google Cloud services such as storage, AI/ML, and analytics. Key benefits include [confidential computing](#), [predictive autoscaling](#), [live migration for VMs](#), [sole-tenant nodes](#), [HPC VM images](#), [GPU accelerators](#), [custom machine type sizing](#), [per second billing](#), [committed-use discounts](#) (up to 57% savings), and [preemptible VMs](#) (savings up to 80%).

A summary of machine types is provided below:

- General purpose (E2, N1, N2, N2D) machines provide a good balance of price and performance.** N2 machine types are the second generation general-purpose machine types based on Cascade Lake CPUs with a base frequency of 2.8 GHz and a sustained all core turbo of 3.4 GHz. Pre-configured shapes are available between 2 to 80 vCPUs and 0.5 to 8 GB of memory per vCPU. These VMs could be used for some of the front end workloads and infrastructure for tools with medium performance requirements.

- **HPC machine images** for machines allow HPC workloads like the ones in EDA to provision machines that come preconfigured with settings to get the most optimized performance for highly parallel and/or tightly coupled loads. HPC machine images are another way in which Google Cloud is able to offer infrastructure pre-tuned for EDA workloads.
- **Compute optimized (C2) machines offer high-end vCPU performance for compute-intensive workloads.** C2 machine types offer up to 3.8 GHz sustained all-core turbo on Intel Cascade lake processors to support compute intensive front end design applications. C2 family of VMs offer a variety of shapes ranging from 4 to 60 vCPUs with local SSD.
- **Memory optimized (M2) machines offer the highest memory and are great for in-memory databases.** For memory intensive backend design applications like static timing analysis and physical synthesis/verification, memory optimization families (M1 and M2) offer higher memory to VCPU ratios (ranging from 14 to 28GB per VCPU). M2 Ultramem is available on the Cascade Lake platform and offers VMs that support memory size up to 12TB.
- **Accelerator optimized (A2) machines are machines for very demanding applications.** A2 machines are based on the A100 GPU.
- **Custom Machine type machines offer unlimited flexibility for CPU and memory selections.** With Custom Machine types, users can select the exact amount of VCPU and memory that a job requires (Figure 5). This eliminates any VM level overprovisioning and customers only pay for the resources they use.
- **Sole-tenant nodes.** EDA tools communicate with a license server to ensure that the user has the required licenses. License server latency can impact tool performance and it is important to ensure that licenses are served as quickly as possible. Google Cloud's sole tenant nodes are excellent choices for license server installations.

HPC machine images are another way in which Google Cloud is able to offer infrastructure pre-tuned for EDA workloads.

Best in class Storage

Traditionally, EDA customers, both on-prem and in-cloud, tend to use File Storage solutions and Network File System (NFS) protocol for exporting file shares to instances running the EDA jobs. When it comes to EDA workloads characterization, a broad range of I/O profiles can be observed. It is common for workloads to require a deep and wide directory structure with thousands of files per directory. The I/O profiles can vary and depend on software tools in use and the workflow design stage, but the following assumptions can be taken into account when trying to generalize these workloads:

- During front-end design phases, when a large number of jobs run in parallel, I/O patterns tend to be mostly random, very metadata intensive, accessing a large number of small files.
- During back-end design phases, the I/O patterns become more sequential, with fewer number of read/write jobs running simultaneously and accessing much larger files. These kind of workloads will drive an increase in overall storage throughput performance requirements.

The common practice for EDA workflows migration to cloud is to try to maintain parity with on-prem storage system/solution design. By using this approach, customers can migrate their applications more quickly and without requiring modifications to their workflows.

For EDA workloads running at scale with mixed I/O patterns, storage systems are one of the major candidates for becoming the bottleneck. Aspects of planned or unexpected increases of performance and capacity must be considered and addressed during the early stages of storage solution planning, providing EDA customers with a highly efficient and highly scalable environment. There are several different kinds of storage resources available in Google Cloud. Each of these tools has recommended usage patterns and price/performance characteristics and can be a critical component of EDA workloads toolbox in the cloud.

Aspects of planned or unexpected increases of performance and capacity must be considered and addressed during the early stages of storage solution planning.

Shared file systems in Google Cloud

NFS solutions are the norm for EDA workloads and users can choose from among the many options available for implementing NFS systems. Options include using the native Google Cloud Filestore, third-party storage solutions from the Google Cloud Marketplace, or rolling out a custom NFS solution directly on GCE.

Locally-attached storage - LocalSSD

Each GCE VM instance can have locally attached NVMe drives added on request. Users can attach a number of multi-GB partitions up to 9TB total NVMe storage per instance. NVMe performance is very fast for read and write operations; details for IOPS/throughput for read and write can be found in the latest documentation available in the references section at the end of the paper.

Network-attached storage - Persistent Disks (PDs)

By default each VM instance has a PD bootdisk attached. Users can increase the size and performance characteristics of the attached PD or attach additional PD drives for a total of up to 64TB per instance.

VM instances can attach either HDD or SSD PDs and performance scales with size up to 100k read and 30k write IOPS and with max sustained throughput of 1.2MB/s read and 800MB/s write.

Google Cloud Filestore

Google Cloud Filestore provides managed NFS volumes and makes them available to GCE resources. There are multiple tiers of size and performance for these volumes, BASIC_HDD, BASIC_SSD, and HIGH_SCALE. You can create managed volumes that range in size anywhere between 10TB up to 320TB with performance that scales roughly linearly.

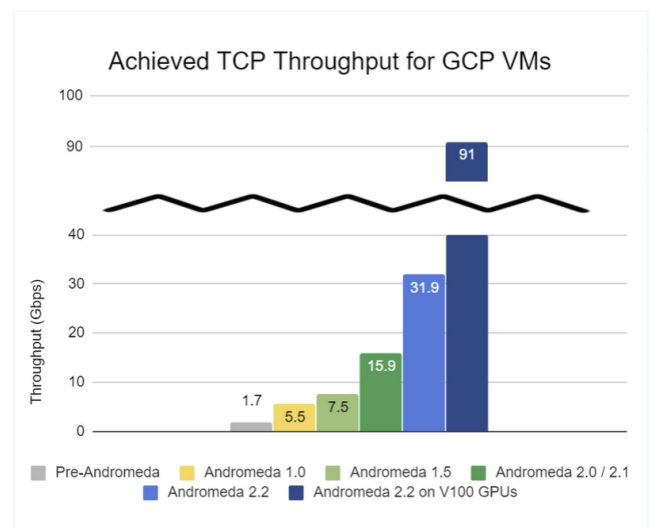
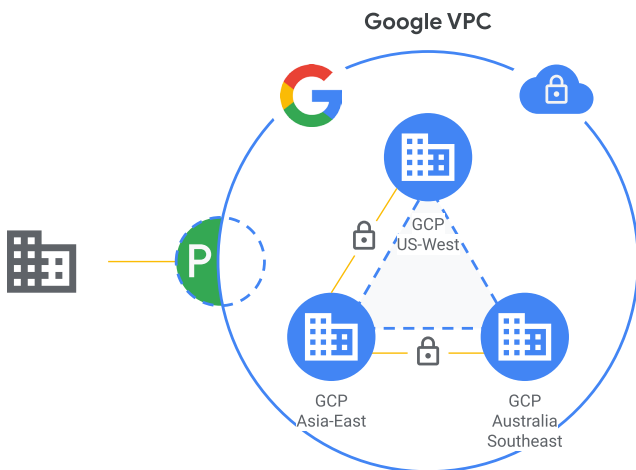


Figure 6: Google's high-performance network substrate. Source: Google, LLC

Object storage - Google Cloud Storage (GCS)

Google Cloud Storage is the cheapest and largest-scaling storage available. It's a wide parallel file system that can be used for extremely high-throughput distributed reads and writes.

GCS buckets can be regional or multi-regional, which places the replication burden on Google for redundancy. They can also be helpful for archival usage patterns and various intermediate and long-term archive storage. The time-to-first-byte read latency can be high for object storage, but the low cost can still make this very attractive for long-running multi-day jobs with bursty read profiles.

In addition to options above, Google Cloud's several partners provide managed services that work well for EDA workloads.

Best in Class Network

EDA workloads demand massive amounts of compute and storage. However the underlying network plays a critical role and it is imperative that the network provides adequate bandwidth, throughput and performance to enable EDA jobs.

Google runs a global, unified, Google owned and operated fiber optic network connecting all data centers providing unmatched speed and resilience. Google Cloud is available in 200+ countries and territories. Google Cloud's networking infrastructure is the same stack used by Google's renowned services, such as Search, YouTube and Gmail. Enhanced with the release of Andromeda[1] in 2014, Google's Software Defined Networking (SDN) based substrate for network virtualization, Google Cloud has been able to provide end-to-end performance with enhanced functionality including distributed denial of service protection, transparent service load balancing, access control lists and firewalls.

What this really means is that, when a customer creates a virtual private cloud (VPC) on Google Cloud, they create a global network.

Google Cloud's networking infrastructure is the same stack used by Google's renowned services, such as Search, YouTube and Gmail.

Google Cloud is the first major public cloud to offer a tiered cloud network. With the premium tier, which is the default network offering, traffic is delivered over Google's well-provisioned, low latency, highly reliable global network. This network consists of an extensive global private fiber network with over 100 points of presence (POPs) across the globe.

The ultimate goal of the network is to maximize bandwidth to compute engine VM's. The latest generation of Andromeda - Version 2.2 - provides an 18X increase in VM-to-VM bandwidth and 8X reduction in latency without introducing any downtime. The maximum network egress bandwidth is raised to 32 Gbps for same zone VM-to-VM traffic on all common VM types. Further, Google Cloud users also benefit from improved performance isolation through the use of hardware offloads. This enables the Compute Engine guest VM to bypass both the hypervisor and the software switch, and talk directly to the Network Interface Card (NIC).

The SDN layer is unique in that it allows for zero-copy payload transfer from the VM memory directly to the NIC and also allows for improving performance and efficiency under the hood without requiring the use of SR-IOV or other specifications that tie a VM to a physical machine for its lifetime.

Cloud interconnect

Most semiconductor customers require private connectivity from their data centers to the cloud.

Cloud Interconnect extends the on-premises network to Google's network through a highly available, low latency connection. Customers have the choice of dedicated or partner interconnect. Dedicated Interconnect connects directly to Google whereas the Partner Interconnect connects to Google through a supported service provider. In both cases traffic does not traverse the public internet.

The latest generation of Andromeda - Version 2.2 - provides an 18X increase in VM-to-VM bandwidth and 8X reduction in latency without introducing any downtime

Cloud Interconnect extends the on-premises network to Google's network through a highly available, low latency connection.

Best in Class Security

Security is often one of the biggest concerns for anyone looking to the cloud. One of the biggest concerns semiconductor companies have is security, to make sure design IP (Intellectual property) is protected.

From the beginning Google has worked to make its services both secure and reliable. The underlying infrastructure doesn't rely on any single technology to make it secure, rather, security is built through progressive layers that deliver true defense in depth. Starting from the bottom, Google builds its own custom hardware. The same goes for software, including low-level software and the OS. Further, hardware is designed to include components specifically for security - like Titan, which is a custom security chip. All of this rolls up to custom data center designs, which include multiple layers of physical and logical protection.

Data by default is automatically encrypted at rest and in transit and also distributed for availability and reliability. Communications over the Internet to the public cloud services are all encrypted. Further, since encryption is one of the most expensive components of packet processing, Andromeda 2.2 utilizes special-purpose network hardware in the NIC to offload that encryption, freeing the host machine's CPUs thereby running the guest vCPUs more efficiently.

With Google Cloud, customers have full control and sole ownership of their data. By providing Access transparency logs, Google Cloud is the only major cloud to provide near real-time logs of Google administrator access on Google Cloud. Customers can also choose to manage their own encryption keys using Google Cloud's Cloud Key Management Service. Furthermore, with Cloud Data Loss Prevention (DLP), customers can identify, redact, and prevent sensitive or private information from leaking outside of an organization.

Global	Americas		Europe, Middle East and Africa		Asia Pacific	
ISO 27001	USA HIPAA	Canada Personal Information & Electronic Documents Act	Europe GDPR	Spain Esquema Nacional de Seguridad	Australia Australian Privacy Principles	Japan FISC
ISO 27017	HiTrust		EU Model Contract Clauses			My Number Act
ISO 27018	FedRAMP		Privacy Shield			
SOC 1	FIPS 140-2		TISAX	South Africa POPI	Australian Prudential Regulatory Authority Standards	Singapore MTCS Tier 3
SOC 2	COPPA	Argentina Personal Data Protection Law	Germany BSI C5		IRAP	OSPAR
SOC 3	FERPA		Switzerland FINMA	UK NCSC Cloud Security Principles		
PCI DSS	NIST 800-53			NHS IG Toolkit		
CSA STAR	NIST 800-171					
MPAA	Sarbanes- Oxley					
GxP	SEC Rule 17a-4(f)					
Independent Security Evaluators Audit	CFTC Rule 1.31(c)-(d)		France HDS			
	FINRA Rule 4511(c)					
	HECVAT					

Best in class Big Data and AI/ML

Big Data, AI and ML are proven technologies to optimize flows. Chip design and manufacturing flows generate lots of data - but systematic collection, analysis and insights is difficult [2]. GCP offers several scalable solutions to build robust data collection and analysis frameworks, such as Cloud SQL and Cloud Spanner for relational database services, Bigquery for scalable data warehousing, and Bigtable for wide-column databases.

Almost every design flow uses databases, but most are home-grown efforts using open source technologies. Such solutions are often brittle and have difficulty scaling to multiple sites and increased usage. With a service like Cloud SQL, users can migrate their backend database servers to managed services like CloudSQL and improve engineer productivity.

AI/ML is also enabling innovative approaches to problem solving in many domains including semiconductors. In chip design and manufacturing ML techniques are being used successfully to get the better quality of results faster. ML techniques are also being used in semiconductor wafer testing with success. Semiconductor organizations have realized the potential of introducing AI/ML in flows and have created organizations to enable AI/ML in flows.

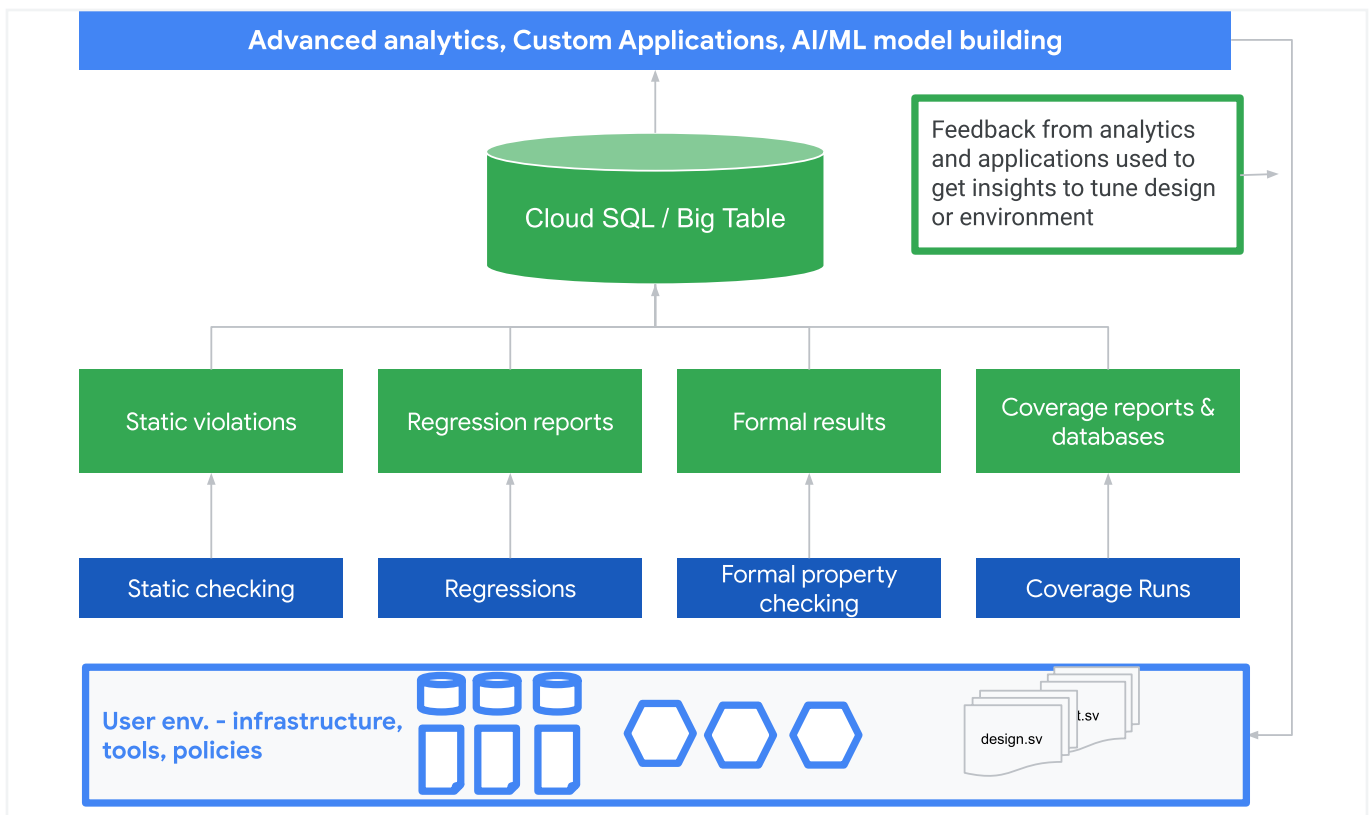


Figure 7: Typical verification flow leveraging GCP native database services

Google Cloud's many solutions spanning databases, datastores, AutoML, ML pipelines among others provide a robust and proven framework for implementing AI/ML solutions. In a recent engagement, a manufacturing company successfully used GCP's AutoML Vision product to build an effective classification model that improved detection of errors in wafers. Figure 7 on the previous page shows how Big Data and ML techniques can be used in mature verification flows to identify bottlenecks, detect anomalies and improve debug productivity.

Google Cloud's many solutions spanning databases, datastores, Auto-ML, ML pipelines among others provide a robust and proven framework for implementing AI/ML solutions.



Chapter 4

Google Cloud reference architecture for EDA workloads

When it comes to selecting the reference architecture, several options exist. Every user situation is unique, and the guidelines below are helpful in determining the architecture. First, some basic deployment models are described, followed by a reference architecture that supports the model.

Burst Model	Determination could be made at runtime whether a tool or flow is run in Cloud or on-prem. Similar setup exists in Cloud and on-premises to enable a seamless runtime environment. Setup example.
Hybrid Model	A subset of projects or flows are run all-on in Cloud. Other projects stay on-prem. Infrastructure setup is similar to All in Cloud. Setup example.
All-on-cloud	All infrastructure (compute, storage, networking, license server) is in Cloud with no dependency on on-prem infrastructure. Setup example.

A typical Burst-mode operation is as follows:

- Engineers login as usual, and submit a job to the on-prem job scheduler
- If the job scheduler finds that the on-prem cluster cannot meet the compute requirement, it sends the request to the cloud via a resource connector
- Job executes on the cloud, and the results are returned to the user for next steps

An architecture capable of supporting All-on-Cloud mode is shown below in Figure 8. Further, a complete example with scripts to demonstrate this mode of operation can be found [in this repo](#).

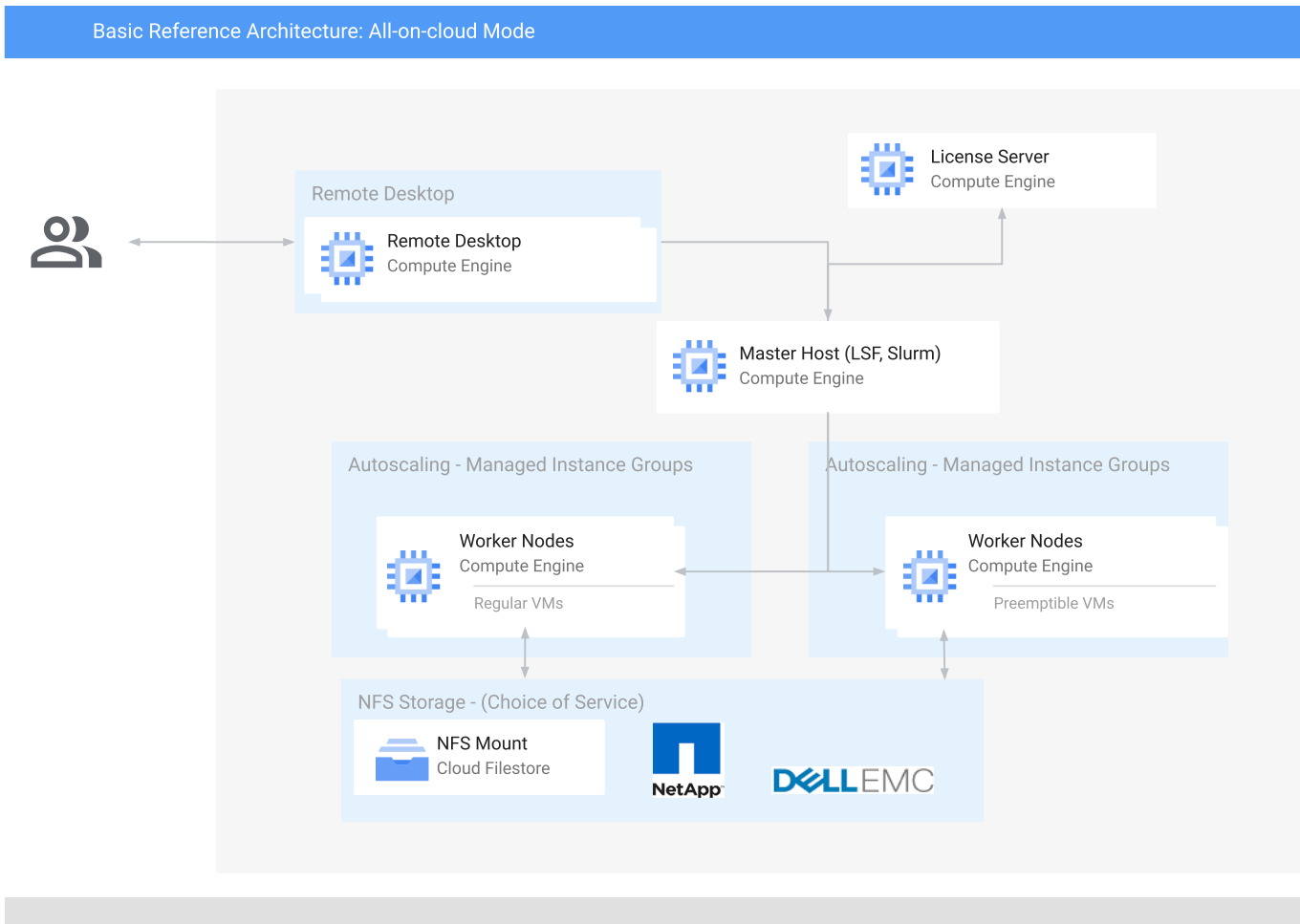


Figure 8: Basic reference architecture for all-on-cloud mode

The architecture shown in Figure 9 can be enhanced by adding Big Data and ML capabilities.

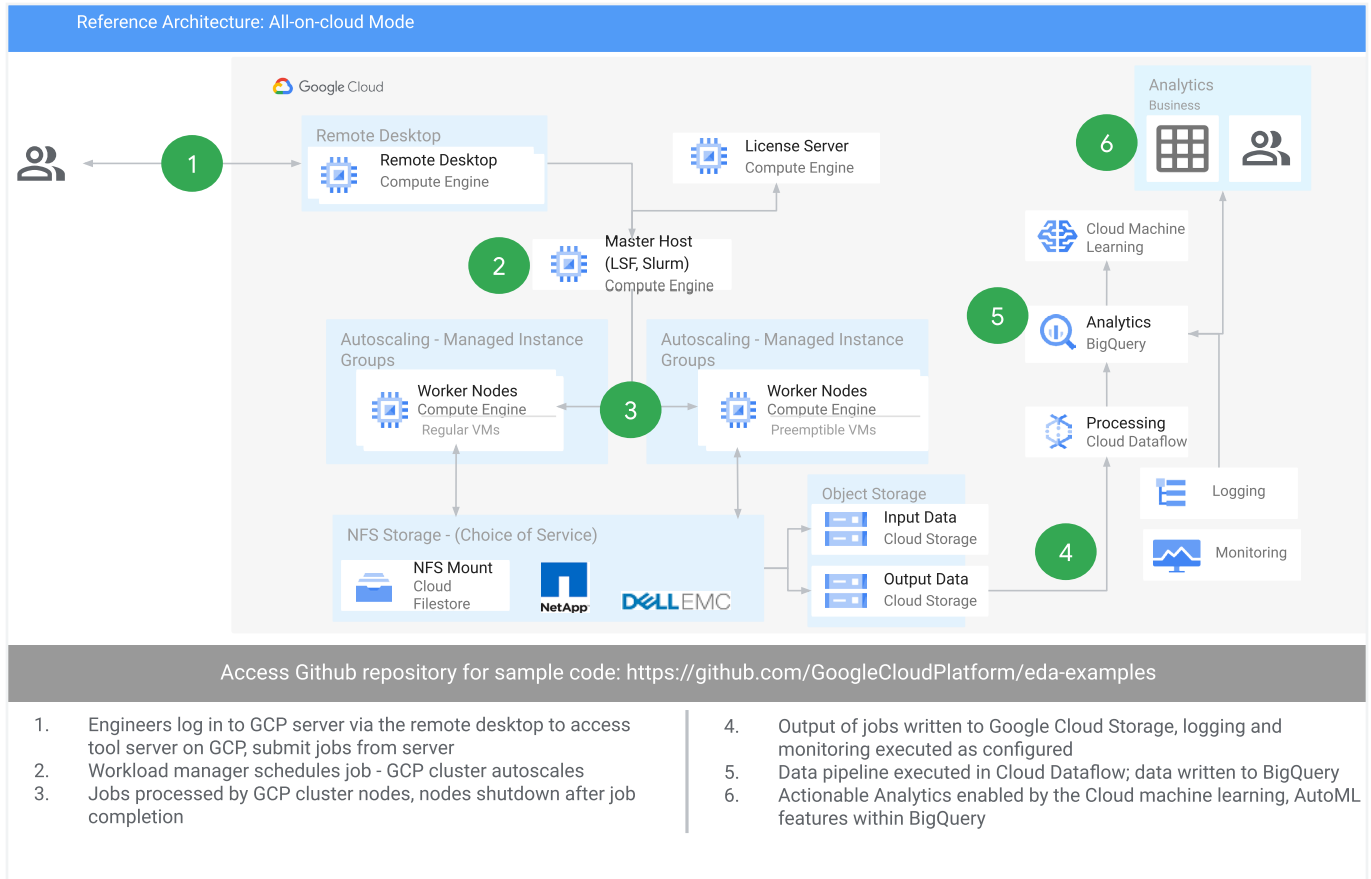


Figure 9: Reference architecture for all-on-cloud mode

A sample architecture capable of meeting Burst model is shown in Figure 10.

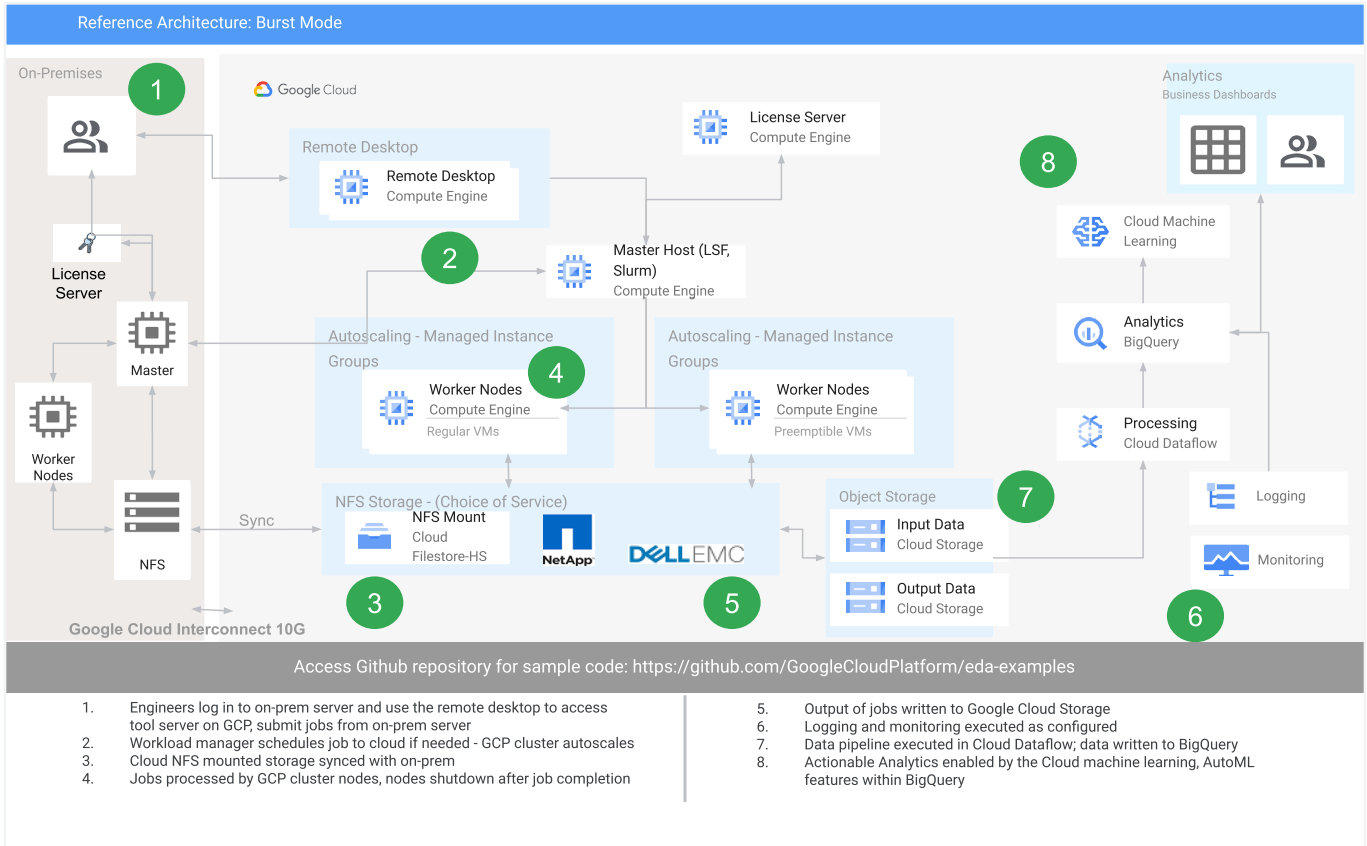


Figure 10: Burst mode reference architecture

Chapter 5

Making a difference — where it matters

Open Source and Innovation

Google's commitment to open source and innovation is proven. TensorFlow, the world's foremost deep learning framework was open sourced by Google in 2015 and continues to be used in a majority of deep-learning enabled applications in production. Kubernetes and KubeFlow are other examples of Google's successful open source contributions.

At Google, we believe that innovation is the key to business success and leadership. Google Cloud consulting services team is committed to working with customers to define, build and execute co-innovation projects that have measurable impact for business.

Sustainability

Google operates the cleanest cloud in the industry. We believe in sustainability and have over a decade of investment in operating responsibly. We have been carbon neutral since 2007 and matched 100% of our electrical consumption with renewable energy purchases since 2017. We recently also committed to fully decarbonize our electricity supply by 2030 and operate on clean energy, every hour and in every region.

You have many things to consider in the cloud platform you choose—its price, security, openness, and products. We believe you should consider the environment too. When you choose Google Cloud, your digital footprint is offset with clean energy, making your compute impact net zero. By moving compute from a self-managed data center or collocation facility to Google Cloud, the net emissions directly associated with your company's compute and data storage will be zero.

We have been carbon neutral since 2007 and matched 100% of our electrical consumption with renewable energy purchases since 2017.

Ecosystem

Semiconductor chip design and manufacturing eco-system is large, vibrant and evolving. At Google Cloud, we believe in building strong partnerships and nurturing them for long-term success. We have ongoing partnerships with major EDA vendors, storage and infrastructure vendors, and foundries. We actively participate in industry events to grow awareness about cloud technologies for chip design.

Cost Optimization





In addition to the custom VM machines, GCP offers two strong mechanisms to optimize the EDA infrastructure for cost.

Preemptible VMs

With PVMs, GCP offers the same VM for a fraction of the cost (upto 80% discount). PVM can be stopped (preempted) by compute engine and is only available for a maximum of 24 hours. With fault tolerant architecture, PVMs provide a very cost effective solution for running batch, checkpointed and high throughput EDA jobs.

Rightsizing Recommendations

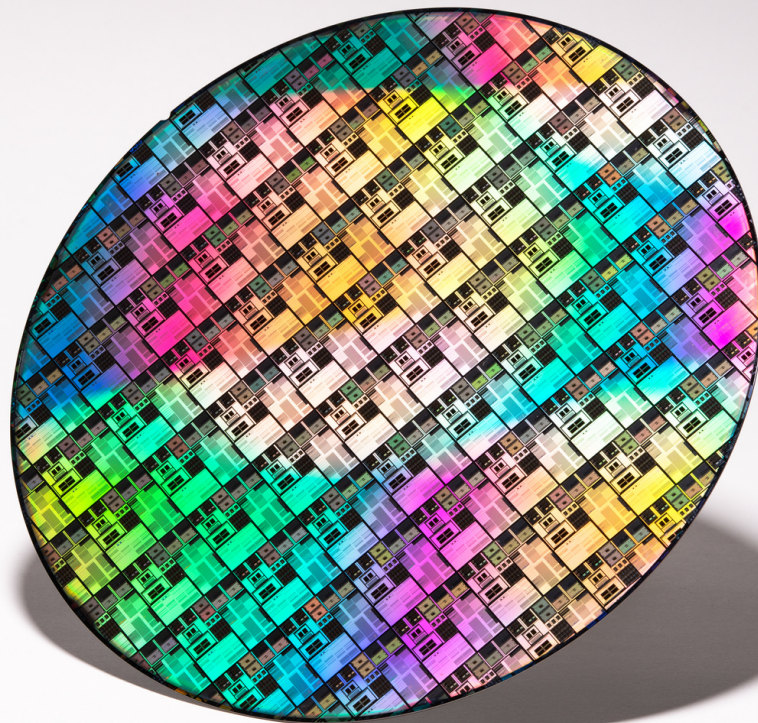
GCP offers a built-in mechanism for additional optimization recommendations after you start running your jobs. Rightsizing insights provide recommendations for updating the VM shape if the system detects under utilized VMs.

<input type="checkbox"/>	Name [^]	Zone	Machine type	Recommendation
<input type="checkbox"/>	<input checked="" type="checkbox"/> instance-10	europa-west1-c	4 vCPUs, 26 GB	 Save \$102 / mo
<input type="checkbox"/>	<input checked="" type="checkbox"/> instance-11	europa-west1-c	1 vCPUs, 0.6 GB	 Increase perf.
<input type="checkbox"/>	<input checked="" type="checkbox"/> instance-12	asia-east1-a	1 vCPUs, 0.6 GB	 Increase perf.
<input type="checkbox"/>	<input checked="" type="checkbox"/> instance-2	europa-west1-c	4 vCPUs, 15 GB	 Save \$77 / mo

This provides an additional mechanism for admins to constantly keep the cost under control for the EDA infrastructure.

Conclusion

Infrastructure modernization is the cornerstone of the digital transformation journey. Semiconductor design and fabrication is a complex and compute-intensive process with several challenges that impede the ability to design faster. Google Cloud's solutions in Compute, Networking, Storage, Data Management and AI/ML enable the semiconductor industry to scale their operations by getting secure access to the latest infrastructure when needed, and differentiate their flows with ML-based techniques.



References

Articles

1. M. Dalton, D. Schultz, J. Adriaens, A. Arefin, A. Gupta, B. Fahs, D. Rubinstein, E. C. Zermeno, E. Rubow, J. A. Docauer, J. Alpert, J. Ai, J. Olson, K. DeCabooter, M. de Kruijf, N. Hua, N. Lewis, N. Kasinadhuni, R. Crepaldi, S. Krishnan, S. Venkata, Y. Richter, U. Naik, and A. Vahdat. Andromeda: Performance, Isolation, and Velocity at Scale in Cloud Network Virtualization. In NSDI, 2018.
2. S. Obilisetty, "Digital intelligence and chip design," 2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), Hsinchu, 2018, pp. 1-4, doi: 10.1109/VLSI-DAT.2018.8373256.

Links

1. <https://www.fiercееlectronics.com/electronics/synopsys-google-cloud-team-cloud-based-functional-verification>
2. <https://cloud.google.com/blog/products/ai-machine-learning/ai-and-machine-learning-improve-manufacturing-visual-inspection-process>
3. <https://cloud.google.com/blog/products/networking/google-cloud-networking-in-depth-how-andromeda-2-2-enables-high-throughput-vms>
4. <https://cloud.google.com/solutions/filers-on-compute-engine>
5. <https://cloud.google.com/compute/docs/nodes/sole-tenant-nodes>
6. <https://cloud.google.com/blog/products/storage-data-transfer/introducing-lustre-file-system-cloud-deployment-manager-scripts>



Google Cloud