Made in the Cloud!

Understanding the Case for IC Hardware Development in the Cloud

Authors

Bryan Dickman,
Valytic Consulting Ltd.

Joe Convey,
Acuerdo Ltd.

Teng-Kiat Lee,
Synopsys, Inc.

Sandeep Mehndiratta,
Synopsys, Inc..

Introduction

While software developers have been building and delivering software in the cloud for many years now,

How much IC hardware is "made in the cloud"?

Run faster, run leaner, innovate sooner, save money, understand your engineering platform costs better and get best-in-class Business Continuity[1] thrown in for free…

Why wouldn't you want to move all your compute-heavy IC design engineering jobs to the cloud right now? Why isn't everyone doing it?

These are heady claims and big questions to which there are not too many clear answers just yet, but the picture is most certainly improving, and consequently more companies are taking the plunge and moving complex IC design engineering tasks to the cloud.

The objective of this whitepaper is to look at the possibilities in the light of realistic engineering challenges, rather than from a Cloud Service Provider (CSP) viewpoint, and to consider the viewpoints from the perspectives of organizations of different sizes and maturity. As many engineering directors will know, the CFO does not take kindly to never ending requests to add more capacity to the on-premises (we will use "on-prem" from herein) compute estate and dreams of the day when just one OPEX line describes the true cost of developing complex IC products rather than a composite of people, IT, EDA, services, facilities, and other related running costs. At the same time, there is a levelling-up opportunity for those smaller organizations to compete against the larger ones without the need for huge initial infrastructure investments.

Cloud delivers high availability, reliability, scalability, performance, and affordability that are becoming harder to match with on-prem solutions for most companies today.

On-Prem

When we talk about "on-prem" systems, these can take several forms depending on how you have provisioned your infrastructures. For some, on-prem is an on-site server room full of purchased or leased servers, storage, and networking, managed by a local IT team. For start-ups this may be no more than a few racks of IT kit. For larger businesses with multiple locations in multiple geographies, each site may have local capacity, and they may be able to share organizational capacities across geographies, thus achieving some level of Business Continuity. Eventually, growing infrastructure needs may expand to data centre level with either fully managed on-site or off-site data centres, or third party managed off-site Colos[2].

Why Cloud? Why Now?

For a long time, IC design teams have resisted a move to cloud for hardware development based on several myths, but 3 key drivers are now coming together as a perfect storm: -

1. The SysMoore[3] era is driving systemic complexity and hyper-convergent design flows, which in turn require exponentially more compute and EDA resources.
2. AI entering design tools and workflows is driving further multiplicative requirements for flexible (i.e., unlimited) access to compute and EDA resources.
3. CSPs have scaled HPC-optimized infrastructure, availability, affordability, and capacity to handle these scaled-out workloads.

The convergence of these 3 drivers means that now is the right time to figure out your cloud strategy, and that will depend if you are a small or large organization, just starting out, or heavily invested in corporate on-prem systems.

So, what about the myths that have been holding this cloud adoption back?

The Security Myth

A key issue which originally stalled any serious moves was the semi-industry's reluctance to place valuable IP onto someone else's hardware with limited ability to sue for liability and limited trust in the security capabilities of the cloud providers. This is even though most organisations have long since adopted Microsoft 365 and other SaaS cloud solutions such as Salesforce, to run the operations of their businesses, meaning highly confidential assets such as legal, commercials and operations being essentially entrusted to "the cloud".

This motivator to keep all IP in-house and "safe" is understandable; the IP is the crown jewels for most semiconductor companies after all, but it became apparent quite quickly that the CSPs have invested far more than any single company is likely to invest in security, so "safest" usually turns out to be the cloud option.

The Predictability (Availability) Myth

The second issue is concern for predictability of availability. Hardware engineering teams don't trust that cloud can deliver the required capacity when they need it. They prefer the predictability of knowing what capacity they have available in their internal compute and storage estate and being assured that it will be there when they need it because it is not being shared with an unknown pool of users. This would be fine in a world of perfect capacity forecasting where it is certain that all engineering workloads will fit within the given capacities, but in practice, internal engineering teams are constantly competing for shared on-prem resources and someone must take a priority call on which projects win when a crunch arises. Of course, you can invest and expand your on-prem to meet burst capacity requirements, but capacity expansion is typically non-agile, and in one direction only – so the on-prem estate only ever grows, it never shrinks when demand wanes. In fact, capacity availability is what the cloud excels at thanks to the huge expansion in capacities globally by the large CSPs.

The Affordability Myth

The third big myth is around cost. On-prem IT teams always believe they can deliver capacity more cheaply than cloud, and owning hardware is essentially more cost effective than renting it over time, or at least the costs are better understood and more controllable. They worry that cloud's effectively infinite capacity will encourage engineering teams to be lazy and consume compute and storage in an uncontrollable fashion with runaway costs. However, they will need to provision enough on-prem capacity to cope with peak demands meaning expensive on-prem hardware may be sat idle for a percentage of time, with poor overall utilisation. This might be an on-prem cost that is overlooked.

The Ease of Use Myth

The fourth myth is around migration of IC development workflows from on-prem to cloud. When workflows have deep dependencies on the architecture of the on-prem estate, lifting and shifting to cloud seems like an insurmountable effort and cost barrier. So, easier to assume that…

…if it's not broken, don't fix it!

In fact, investing effort to migrate your existing workflows to the cloud has other benefits. As jobs become better encapsulated and less dependent on the target platform, you can progress to an environment where workflows are portable and can be run both on-prem or in-cloud in a seamless fashion.

But What's Wrong with On-Prem?

The answer to that depends on who is asking.

From a Business Risk View…

For a large corporate organisation to move away from a well understood and reliable investment model is a business change risk. They are likely to have large legacy investments in their on-prem engineering platform to underpin that consistent delivery. It's all about predictable outcomes and guaranteed roadmap delivery to support quarterly revenue targets. Investors expect revenue and profitability growth, which is solidly underpinned by a resilient delivery machine and known quality that cannot be put at risk. In these circumstances, it's a brave engineering manager that pushes for change. "It's my job to worry about delivery; someone else can worry about the money when we need more compute capacity; our job is to deliver (whatever the cost…)". Someone in engineering management will care about predictable OPEX costs and will worry about cloud costs that can just runaway. "Can we trust engineers not to launch zombie jobs that will cost a fortune?".

For a leaner start-up organization, the business risk is characterized by a large investment cost barrier for on-prem which may be at odds with a business strategy to be agile and competitive, establish market share quickly, and secure future investment. They really don't have the time, expertise, space, and budget to grow their own platform. It's about total focus on innovation and fastest delivery of first successful product to support further funding rounds and eventual purchase, or IPO.

Of course, there are examples of everything else in between. Larger engineering operations may allow teams working in small acquisitions, for example, to act as a testbed to experiment with cloud with known risk and manageable outcomes should it succeed or fail.

The Engineering View….

Luckily this picture is changing, as more engineering teams realise that

Cloud deployment is not an all or nothing endeavour.

Teams can start small, explore productivity and innovation opportunities, and manage risks at the same time. They don't need a big-bang event. They can run many smaller evaluations and proofs of concept and show what success looks like for workflows running in the cloud, one workflow at a time. As they do this, they may be freeing up capacity in an overloaded on-prem estate for workloads that are harder to move over or need to be managed internally for other reasons.

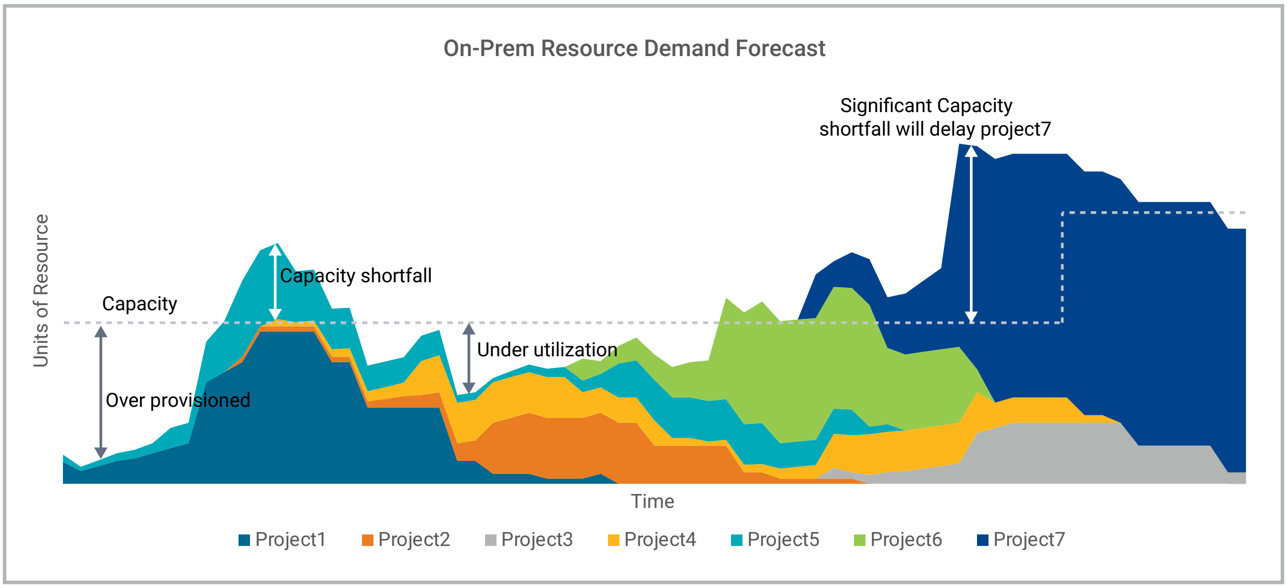
For smaller and medium companies, it is a different story. Of course, cost matters, but investment rightly goes into innovation and first product delivery. They would prefer to avoid excessive start-up investment costs and focus on reaching market earlier and at lower costs. They simply may not have the space to accommodate infrastructures nor the people to build and operate them. It's not their core skillset and neither do they want it to be so.

What's the fastest and most scalable way to deliver engineering production capability?

For this reason, Pay-As-You-Go (PAYG) models are very appealing, but this needs to extend to both infrastructure and IC design tooling. What's more, you can also think of PAYG as Pay-As-You-GROW i.e., you can instantly grow your consumption to meet demand. It's a budgetary exercise and not an operational deployment issue.

The Fixed Capacity, Fixed Architecture Problem

Recall the predictability (availability) myth. This is very much a big-company problem. You have multiple engineering teams competing for the same shared on-prem resources, and you have the operational challenges of forecasting demand and aligning the capacities to meet the forecast. If you fail to do this well, your product roadmap is in jeopardy, or you may have over-spent on infrastructure that is poorly utilised to meet peak demands.



For the smaller organization, the forecasting may be less complex than above, but the ability to forecast may be much less, due to lack of historical data, or the less predictable nature of doing something new for the first time. In a cloud scenario, the capacity line shown above translates to a cost or budget line, but the good news is that you can flex up and down around that budget line so long as your average and total costs come in on budget. When they don't, the mitigation is "only" a budget issue i.e., it is not a budget AND an operational deployment issue to increase capacities. You can effectively flex capacities instantaneously within budget constraints.

Another problem with architecting and building your own infrastructure is that hardware quickly goes out of date as servers and storage solutions are constantly improving in performance and capacity. This implies an additional cost of regularly refreshing the hardware. Again, something that the smaller start-up business might struggle with, but would be able to exploit fully as a cloud adopter.

EDA license costs vary according to tools, but when running those high-value tools, it makes sense to optimize the license investment by executing the most expensive jobs on the fastest available hardware. For example, many modern engineering workflows are evolving to exploit GPUs, AI acceleration, and non-Intel server architectures for speed and power efficiencies and cost reduction. CSPs can evolve and scale-up their service offerings at a much faster rate than most companies, big or small, operating their own infrastructures.

Barriers to Cloud

Having explored perceived business risks and how they alter according to the size of company, we also need to look at some of the practical issues preventing hardware engineering teams (in both large and small organizations) from successful cloud adoption.

Existing IC design, verification, and implementation workflows may have evolved with an unknown number of assumptions and dependencies on the compute and storage environment, making it difficult to lift and shift to cloud. On top of that there may be tools licensing assumptions that also prevent you from using the exact same workflow in an unconstrained cloud environment.

Can I use my EDA tools in the cloud?

The likelihood is that existing contracts may not have the provision to use your existing EDA tool licences in the cloud (the Bring-Your-Own-License or BYOL model), so whoever manages EDA contracts is going to have to talk to the vendor to establish the art of the possible. All the main EDA vendors offer multiple models to use licences in the cloud, however there may be some contract changes necessary and you may in fact want to be able to do both things, continue to operate some of your licences on-prem as before, while operating others in a cloud environment.

Again, this may be easier as the smaller start-up, where you can configure your licensing model as appropriate to cloud from the start. Here, the on-demand OPEX nature of cloud compute opens the possibility to run 10X/100X/1000X bigger workloads for RTL simulation for example, something only large companies could do till now. But affordability and availability of licenses for the right EDA tools becomes a limiting factor. Users need a range of flexible solutions, that allow them to make choices about which CSP to use, and PAYG (or metered) licensing models to control license costs. After all,

there's no point having access to infinite compute in the cloud if you are then limited by EDA licenses.

Data and Storage

Don't forget that cloud is not just about compute, it's also about storage. For some engineering workflows the storage cost can be the dominant factor. Take circuit simulation for example; FastSPICE workflows can generate significant volumes of characterization data. As a rule, creating large datasets in the cloud is not an issue in terms of availability, but cloud storage can be expensive. Not everything that is stored in the cloud needs to be in high-performance tier 1 storage. Move less frequently accessed data into one of the slower tier 2 or tier 3 storage mediums provided by your CSP. There is also a cost associated with data transfer. It's usually free to upload, but downloads are metered and costs can escalate.

Cloud users need to carefully consider both storage consumption and compute consumption. Execute all your data processing in the cloud wherever possible, to minimize the volumes of data that need to be shifted in and out. Don't retain large volumes of intermediate generated data, that can be reproduced easily. Perform all your data analytics in the cloud, something that the cloud is increasingly capable of doing thanks to the emergence of modern big data analytics.

Controlling Costs

The need for demand forecasting does not go away as you switch to cloud. What does go away are the technical and physical barriers to capacity expansion, as short-term demands can be more easily met, and you need only pay for the services you consumed.

No need to maintain expensive capacity on stand-by.

However, as mentioned earlier, costs could spiral with runaway consumption if engineering teams no longer feel constrained by fixed capacity and see great opportunities to improve engineering quality and time to market without consideration for the cost ROI calculation. Therefore, built in utilization analytics and budgetary management are essential parts of any cloud-based IC design workflow.

Having an excellent understanding of capacity requirements before experimenting in cloud is highly advisable. It may be possible to plan cloud requirements on this basis. What is less obvious is what the cost will be to provide that access and how it will compare to existing on-prem costs.

It is sometimes the case that organizations don't have a true picture of the total build and operate costs for an on-prem capability. Cost models can ignore some elements of the operational costs such as staffing costs and any lost productivity time that may arise from availability outages due to failures, maintenance, and upgrades. Also, the cost of mitigating against reliability and business continuity risk which implies the need for duplication and redundancy in the systems. Since on-prem estates are usually heterogeneous mixes of old and new hardware, fast and slow compute, small memory, and large memory systems; it is not always understood if the workload/tool combination is operating on the most cost-efficient node. On-prem IT teams therefore invest much time and effort into real-time and operational analytics to measure the utilisation and efficiency of their on-prem investments. Also, careful planning is required to schedule regular maintenance cycles that minimize the availability impact of the systems.

Cost models for cloud are also not that simple. There are many dimensions to the cost such as the choice of provider, choice of services (compute and storage), choice of add-on services such as bigdata analytics, pricing models based on PAYG, pre-purchase plans, or spot-pricing models which can discount services significantly at times of low demand.

The great thing about cloud is that all the costs associated with running a job can be accounted for on a single OPEX line. Remember that costs must be considered as the combination of compute, EDA tools licences, storage, and data transfer.

As mentioned earlier, smaller companies and start-ups may not want to have capex costs associated with on-prem compute and will build cloud-only access to tools into the financial structures from the word go. The financial re-engineering required for a large company to do is difficult, making a full-scale move to cloud a non-starter in most cases. They will opt for a hybrid measure, with known costs.

Migration Strategies

The Lift-and-Shift Challenge

If you look at the compute consumption for IC design today, verification using simulation is typically the dominant workflow. Why is this? The key dilemma of verification is its somewhat open-ended nature for a modern high-complexity ASIC or IP Core. How do you know when you are done and how do you know that you have found all the bugs? You don't! And a consequence of modern constrained-random verification methodologies is that there is always the possibility to run infinite verification cycles, although good use of coverage and other verification metrics is the normal way to curtail this. Hence verification typically consumes the most compute and simulation license resources. However, regression testing and deep soak testing tends to be highly scalable in nature as it is typically characterized by many independent short jobs that can be run in parallel. The time to complete the job is governed mainly by the width of the capacity available, and not the execution time of the longest job. Moving your simulation payloads to the cloud can be a game-changer, by exploiting the massive parallelization that is on offer with cloud.

Other IC design workflows are also very consumptive. Take FastSPICE for circuit simulation. Although less influenced by random seeding and therefore more deterministic, FastSPICE typically is very compute and data intensive as it crunches through all the standard characterization corners for an IC circuit, and this makes it another good candidate for shifting to cloud where capacities are less constrained and the end-to-end time to get the job done can be dramatically reduced.

This "lifting and shifting" of workflows to the cloud necessitates those workflows are encapsulated in a way that they can be sent to the cloud with all the job dependencies packaged up with the job. For teams looking to port in-house workflows, this may mean some analysis of the existing on-prem workflows to establish the I/O requirements and file dependencies of each job, and then possibly some re-architecting of the workflows to make them cloud-ready. Further, you will have to decide where to perform the interactive elements of the workflow such as debug. Will you run both batch and interactive workloads in the cloud, or only run batch in the cloud and use on-prem for all interactive debug jobs? Response times being the main consideration for interactive debug activities such as waveform analysis.

Adoption Models

Smaller organizations and start-ups are not generally encumbered with the investment legacy that a larger established organization is likely to have. This leaves them several fast-start and agile options to establishing a working cloud model for their IC design workloads.

BYOL/BYOC

There are 2 scenarios here that should not be confused. Bring your own License (BYOL), where users can use their existing on-prem licenses with their chosen CSP. That may be the case for the established or larger business that already has established tools licensing agreements for their on-prem environment, and they now want to be able to use those same tools in a cloud environment. Users are still license limited according to their license investment level. Bring your own Cloud (BYOC) is a model which may be much more attractive to those users that have already established cloud capabilities with a chosen CSP, and where demands can be very bursty and peaky in nature. Licenses are not limited as it is a PAYG consumption model similar to the CSP compute model. Metering and analytics are used to bill for usage and to allow for effective consumption budgeting. Other possible pricing models may be available that allow users to pre-plan and pre-purchase license hours at alternative cost points.

SaaS

Software as a Service (SaaS) is further alternative where the user no longer needs to be concerned with the set up and running costs of using a public CSP since the tools vendor is hosting the application on the vendors chosen cloud service. This is a familiar model for many modern applications that your organization is already using today; think Office365, Salesforce, Business Intelligence tools, and many others. These are all SaaS solutions that are running on cloud services, where the user is not exposed to the complexities of the cloud service. The IC development team only care about running a workflow such as simulation for example. Effectively this becomes "Simulation as a Service". Again, the charging model could be metered/PAYG, or it could be based on a given capacity of licenses. For the smaller start-up, this option might represent the lowest cloud adoption barrier, and fully removes the cloud management aspect from the equation.

Hybrid Cloud

For larger organizations with a pre-existing investment in on-prem capabilities, a hybrid approach is a popular strategy. Peak demands can be met by bursting capacity into the cloud for suitable workloads, while less portable workloads continue to be run on-prem with no impact. This more gradual migration to cloud means that headroom becomes available in the on-prem estate for those workflows that require more time and effort to “cloudify”. Established workflows may have evolved to contain many built-in (and possibly unknown) dependencies on the on-prem environment, for example, accessing many distributed NFS files, libraries, scripts, and tools as they execute. In the hybrid-cloud model you need to decide which workflows you are going to port to the cloud and which ones will remain on-prem. Eventually, you might evolve to a point where all workflows work equally well on-prem or on-cloud, at which point your users don’t need to care where their jobs run. They launch a job and whether it ends up on-prem or on-cloud should not matter. At this point the on-prem estate can be characterized as an on-prem or private cloud and the hybrid nature of the compute environment is abstracted away from the consumer.

Opportunities to Do Things Differently

Innovation is the lifeblood of any R&D effort in the semi-industry. The pace of change has been relentless and has been pushed by tools innovation and customers presenting vendors with engineering challenges that would have seemed preposterous only 5 years ago.

The advent of cloud could turn out to be one of those major inflection points in history which allow engineers to improve personal productivity, performance, turnaround time, and accuracy. Essentially, breaking free of the constraints placed on innovation by limited compute can offer opportunities to do things in a different way and it becomes a leveller for small organizations to compete with larger ones.

Many IC design engineering teams are accustomed to the constraints of capacity. This affects the overall quality of the final product and the time to market. What opportunities are missed by not being able to deliver highest possible quality in a market-winning timeframe because there simply aren’t enough compute resources available to accelerate the delivery? Take RTL verification for example. What would you do differently if you could turnaround a full regression in a few hours, thanks to massive scale-out, rather than waiting until Monday morning to get the results of the over-weekend regression? And if that regression has gone rogue for some reason, you may have to restart and wait for several more days to get a result. At the end of the day, the main resource constraint you are operating under is the people one. Engineer time is the most valuable and the most limiting factor. Engineers being blocked waiting for lengthy batch compute jobs to complete is not an effective use of your engineering talent. With capacity and availability constraints lifted, talented engineers can focus on the things that they do best, innovating!

Earlier we said, “why cloud, why now?” and characterized the present situation as a perfect storm in which growth in systemic complexity drive compute and EDA demands exponentially, and modern AI-enhanced EDA technologies further increase this demand. Both CSPs and EDA vendors offer usage models that make cloud adoption affordable and scalable for IC hardware developers; whether you are a small or medium sized organization with some or no on-prem investment, or a large established organization with significant sunk costs in infrastructure but need to be able to meet demand expansion requests or peak usage levels in a more sustainable way as the business grows.

What’s more, if the competition is already exploiting this vast resource to deliver products faster, you have no choice but to do the same.

[1] https://en.wikipedia.org/wiki/Business_continuity_planning

[2] https://en.wikipedia.org/wiki/Colocation_centre

[3] https://www.synopsys.com/glossary/what-is-sysmoore.html

SYNOPSYS®
Silicon to Software™

©2022 Synopsys, Inc. All rights reserved. Synopsys is a trademark of Synopsys, Inc. in the United States and other countries. A list of Synopsys trademarks is available at synopsys.com/copyright.html. All other names mentioned herein are trademarks or registered trademarks of their respective owners.

03/25/22.CS865930248-Cloud WP.

Pub: March 2022