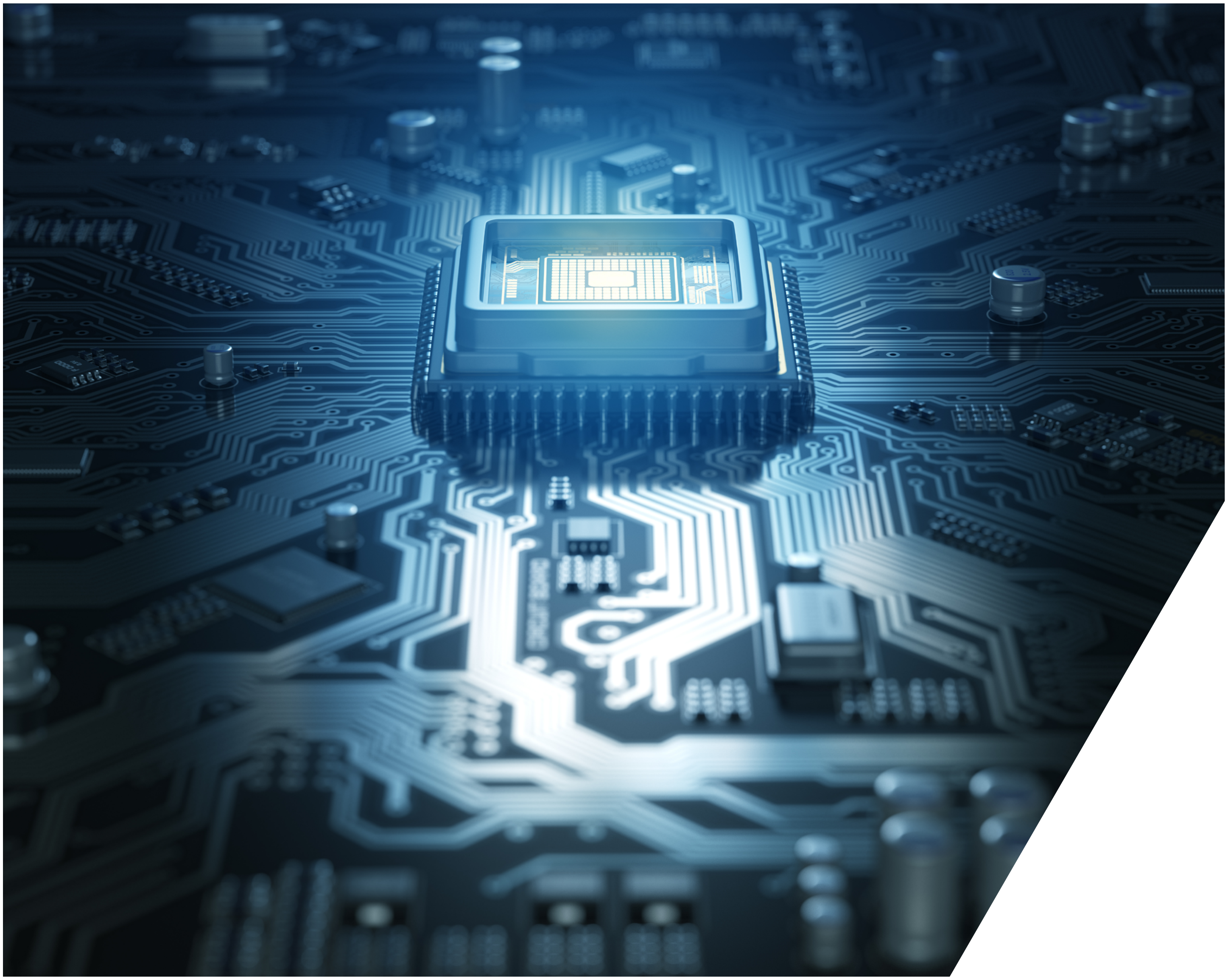
CUSTOMER STORY



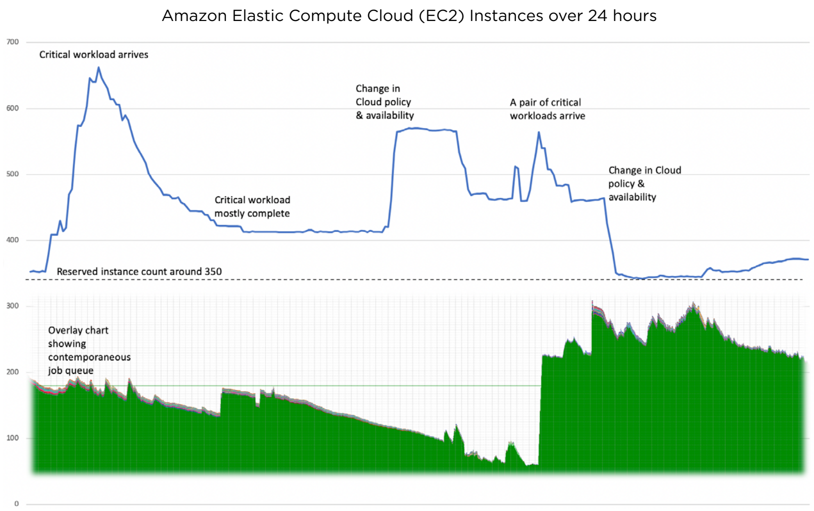
RAPID CHIP DESIGN IN THE CLOUD

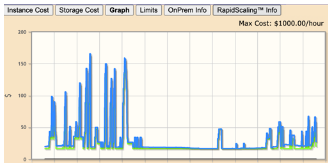
LICENSE-FIRST APPROACH TO SCHEDULING ENHANCES RESOURCE MANAGEMENT

About the Customer

Annapurna Labs was established as a fabless chip start-up in 2011, focused on bringing innovation to the fast-growing cloud infrastructure. Four years after its inception, Annapurna Labs was acquired by Amazon Web Services (AWS). Since then, Annapurna Labs has accelerated its innovation and developed a number of products that benefit cloud customers, including AWS Nitro technology, Inferentia custom Machine Learning chips, and AWS Graviton2 processors, based on the 64-bit Arm Neoverse architecture purpose-built cloud server.

> “Altair’s license-first approach to scheduling enabled Annapurna Labs to enhance its resource management. It not only gave us more control over resource usage and cost, we dramatically improved productivity and time-to-product through the Continuous-Integration development flow.
>
> Nafea Bshara,
> Annapurna Labs





LEFT: Around 350 Reserved Instances are scaling to 660 using On-Demand and Spot Instances responding to variable workload over 24 hours. Jobs queue is represented in the green overlay chart.

RIGHT: Cost profile of the compute resources shows how Rapid Scaling tracks the needs of the workload.

Their Challenge

As a chip design company, time-to-market and engineering efficiency are the most critical and expensive metrics upon which to focus. With this in mind, the team at Annapurna Labs selected the Altair Accelerator™ job scheduler for their front-end and back-end workflows. The team was managing workloads on a number of dedicated Amazon Elastic Compute Cloud (EC2) instances and they could occasionally scale up by manually adding new On-Demand instances. However, the process was not automated and led to high touch, inefficiency, forgotten unused compute resources, and either under-scaling or excessive scaling. As a feature within Accelerator, Rapid Scaling unused compute resources was developed with Annapurna Labs to add structure and efficiency to scaling AWS compute resources, shorten time to results, and change the development model to Continuous Integration.

Our Solution

In addition to automatically starting new instances only when there is demand, Rapid Scaling looks at the speed at which demand is being processed and stops scaling up if the speed is good enough. This means demand can be satisfied in 10 minutes. The license-first approach to scheduling allows Accelerator to efficiently differentiate workloads waiting for licenses versus workloads waiting for hardware. Only if a workload is waiting for hardware does it make sense to request AWS instances. All resources are freed after they have been idle for one minute.

Many features were added in cooperation with Annapurna Labs, including configurable selection of instance type, Spot Instance support, protection against various errors like saturation of instance types, size of /tmp, fine control of the number of jobs that can be executed on each new instance, and many others.

Electronic design automation (EDA) jobs can be short and the spin-up time for an instance is comparable to those jobs' runtime. The ability to understand workload speed and spin-up costs enables Rapid Scaling to avoid overshoot. Amazon EC2 offers the broadest and deepest choice of instances, built on the latest compute, storage, and networking technologies and engineered for high performance and security. Rapid Scaling allows job resource requests to map to the most appropriate instances.

While AWS exhibits high elasticity in some cases, a particular instance type may not be available. Rapid Scaling understands how to select a backup instance type if the first choice is not available. After the workload surge has passed, idle instances terminate. This flexibility maps nicely into AWS notions of Reserved, On-demand, and Spot instances.

Results

With the installation of Rapid Scaling, Annapurna Labs has reduced its cost by at least 50%. Additionally, with Rapid Scaling now part of Annapurna Labs's chip development Continuous Integration flow, they are seeing faster incremental development and continuous regression. Annapurna Labs keeps tighter control on costs and benefits from a detailed view into resource usage by projects and users.

Rapid Scaling is available on the AWS Marketplace: **View AWS Marketplace**