134,426 views | Mar 10, 2019, 12:27pm EDT

How An Acquisition Made By Amazon In 2016 Became Company's Secret Sauce

Janakiram MSV Senior Contributor
Cloud
I cover Cloud Computing, Machine Learning, and Internet of Things

When Avigdor Willenz invested $20 million in a startup founded by two of his ex-colleagues - Bilik (Billy) Hrvoye and Nafea Bshara – little did he know that it would get sold for a whopping $350 million to Amazon. Just in three years, this acquisition has turned Amazon Web Services into a formidable player in the hardware and chip market. This technology has emerged as a threat to Intel and AMD.

Annapurna Labs, an Israeli company, was named after one of the tallest peaks in the Himalayas. Billy and Nafea wanted to trek Annapurna just before launching the startup. Though the duo couldn't start the trek, they ended up calling their new-found company as Annapurna.

Annapurna Peak SOURCE: ANNAPURNA LABS

The timing couldn't have been better for Avigdor, Billy and Nafea. Amazon, which was seriously evaluating building custom chips for their cloud infrastructure services found Annapurna Labs a perfect target for acquisition. With the top engineering team at AWS becoming visibly unhappy with AMD's performance as an alternative to Intel, they walked away from the partnership with AMD.

There are many common traits between Annapurna Labs and Amazon. Both are incredibly secretive about their research and product roadmaps. Both were obsessed with customer adoption. Like most of the teams at Amazon, Annapurna Labs was a lean and mean company laser-focused on building a niche, world-changing technology that would challenge the segment leaders – Intel and AMD.

LinkedIn's 50 Best Startups To Work For In 2020

How To Create A Sales Plan While Dodging The World's Curveballs

Which RPA Software Are The Most Popular With Users?

Three years later, Annapurna Labs' investment emerged as one of the most successful and strategic acquisitions for Amazon. It gave AWS an edge against its arch-rivals, Microsoft and Google.

To appreciate the value Annapurna Labs brought to AWS, we need to understand the evolution of Infrastructure as a Service (IaaS).

The original avatar of cloud infrastructure was launching and accessing virtual machines on a pay-as-you-go pricing model. Spinning up VMs on-demand and shutting them down when the job was done was nothing short of magic.

Amazon EC2 was the pioneer in the IaaS market which revolutionized the way infrastructure was provisioned and consumed. It marked a milestone in computing by empowering multiple startups to go live and scale their business with almost no capital expenditure (CAPEX) spent on the infrastructure.

The original IaaS technology at AWS ran a highly customized version of Xen, an open source hypervisor that lets multiple virtual machines run on a single physical machine. Eventually, AWS found that Xen has many limitations that could slow down the rapid growth potential of Amazon EC2.

The engineers realized that there is an opportunity to optimize the infrastructure performance and cost by moving the software to a purpose-built hardware component. By offloading the hypervisor and networking stack to a specialized hardware accelerator called Application Specific Integrated Circuit (ASIC), Amazon EC2 would function at a higher speed and lower cost. Since these changes are made at the lowest level of the technology stack, it would be transparent to the customers, who will benefit from the enhanced performance.

The C3 instance family announced in 2013 saw the debut of custom chips in Amazon EC2. They were backed by a custom network interface that delivered faster bandwidth and throughput.

The decision to build homegrown, custom hardware for EC2 led to the partnership with AMD. James Hamilton, the man behind the massive data center rollouts at AWS, didn't find AMD delivering the expected performance.

A couple of years later, in 2015, the partnership opportunity with Annapurna Labs resulted in the launch of Amazon EC2 C4 instance family. Apart from offloading the network virtualization to custom hardware, these instances were supported by an ASIC optimized for storage services. But both C3 and C4 were still running the traditional hypervisor on top of an Intel Xeon processor.

What started as an alliance with Annapurna Labs quickly graduated into an acquisition. In 2016, Amazon announced that it is acquiring the Israeli startup for an undisclosed sum.

In 2017, after two years of intense collaboration between Amazon EC2 engineers and Annapurna Labs, AWS announced the C5 family of EC2 instances. These new instances offered a 25% price/performance improvement over the C4 instances, with over 50% for some workloads. This launch marked a milestone in the history of Amazon EC2. It replaced Xen with a highly optimized KVM hypervisor tightly coupled with an ASIC that did the heavy lifting of virtualization. This combination delivered blazing fast virtual machines that almost matched the performance of a bare metal server. Amazon continued to support and enhance other instance types that still run on Xen.

The powerful combination of wafer-thin KVM software tightly coupled with an ASIC is dubbed as Project Nitro. The number one goal of Project Nitro's engineering team was to provide performance that is indistinguishable from bare metal, and they did succeed in meeting it.

At re:Invent 2017, AWS announced the most anticipated feature of Amazon EC2 – Bare metal instances. The secret sauce behind the bare metal instances is Project Nitro. It also enabled Amazon to support a variety of hypervisors including Xen, KVM, and even vSphere.

Project Nitro became the core technology building block to run VMware on AWS. It paved the way to the historical partnership between VMware and AWS, which were fierce competitors in the cloud market.

The next logical step for Annapurna Labs was launching a custom CPU which they did at re:Invent 2018. AWS Graviton processors, built around Arm cores from custom-built silicon, are designed from the ground up for scale-out workloads.

Amazon EC2 A1 instances are powered by AWS Graviton that runs on a technology stack with no dependency on Intel x86 architecture. This marked the official entry of Amazon into the silicon industry challenging the dominance of Intel and AMD.

The innovations from Annapurna Labs don't stop at the network, storage, and virtualization accelerators. They are now enabling AWS to take the lead in training and running machine learning and artificial intelligence models in the cloud.

Andy Jassy, CEO at AWS, announced AWS Inferentia at re:Invent 2018. This service is powered by Annapurna ASIC chips that accelerate the inference of machine learning models trained using TensorFlow, Caffe2 and ONNX. The chip is expected to become available later this year.

Custom ASIC for AWS Inferentia SOURCE: AMAZON

AWS Inferentia makes Amazon's cloud the cheapest to run machine learning inferences. It competes against Google's AI accelerator called TPU and Microsoft Azure's FPGA.

With cloud providers poised to build custom silicon, Annapurna Labs is turning out to be Amazon's secret weapon. It will help AWS deliver innovative services that offer the best in class infrastructure to run modern workloads.

Follow me on Twitter or LinkedIn. Check out my website.

Janakiram MSV

Janakiram MSV is an analyst, advisor and an architect at Janakiram & Associates. He was the founder and CTO of Get Cloud Ready Consulting, a niche cloud migration and... **Read More**