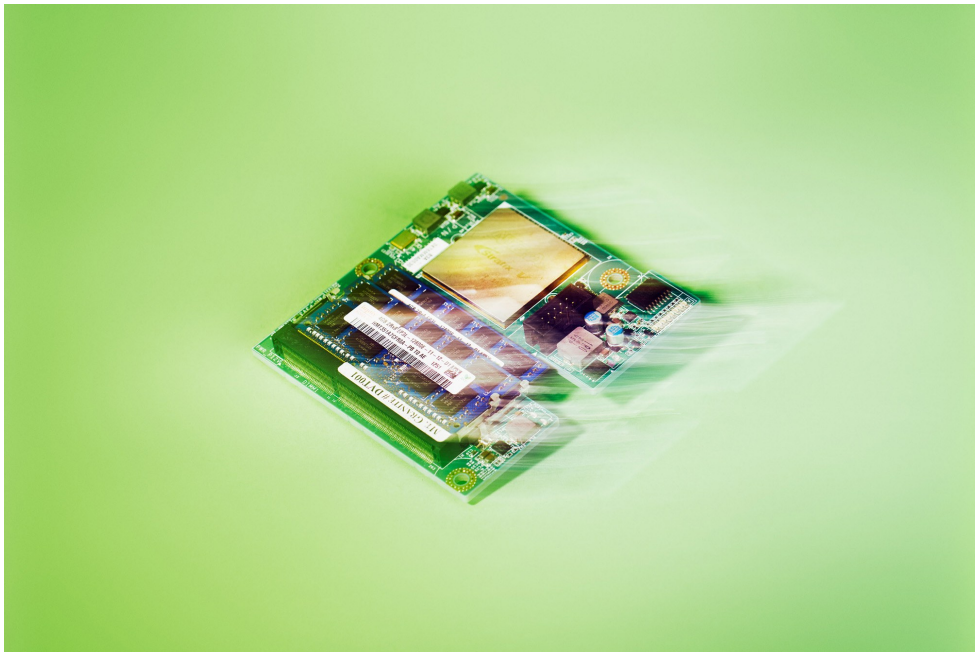
CADE METZ BUSINESS 09.25.16 7:00 PM

MICROSOFT BETS ITS FUTURE ON A REPROGRAMMABLE COMPUTER CHIP



Prototype hardware for Project Catapult, a six-year effort to rebuild Microsoft's online empire for the next wave of AI---and more. CLAYTON COTTERELL FOR WIRED

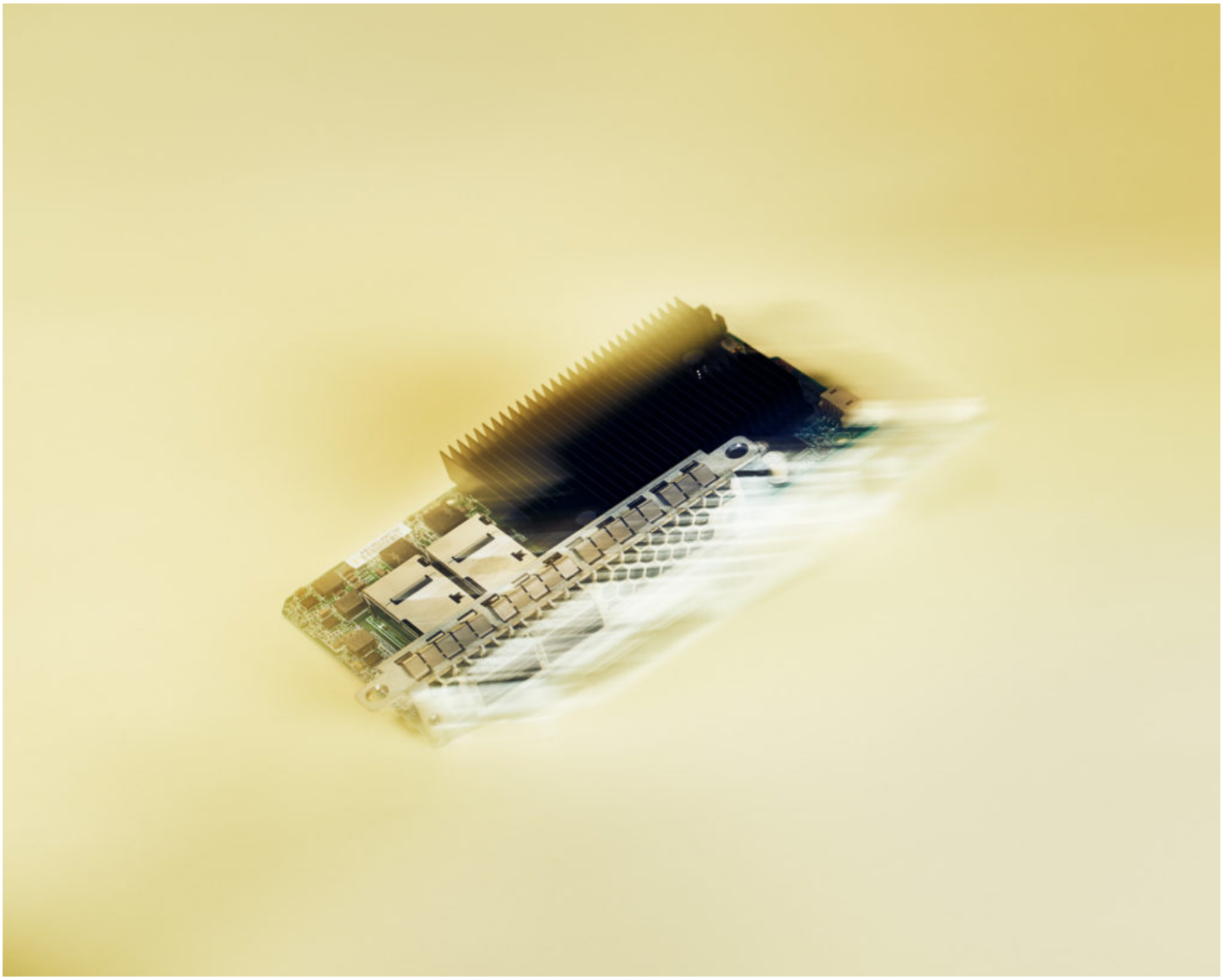
IT WAS DECEMBER 2012, and Doug Burger was standing in front of Steve Ballmer, trying to predict the future.

Ballmer, the big, bald, boisterous CEO of Microsoft, sat in the lecture room on the ground floor of Building 99, home base for the company's blue-sky R&D lab just outside Seattle. The tables curved around the outside of the room in a U-shape, and Ballmer was surrounded by his top lieutenants, his laptop open. Burger, a computer chip researcher who had joined the company four years earlier, was pitching a new idea to the execs. He called it Project Catapult.

Doug Burger. CLAYTON COTTERELL FOR WIRED

The tech world, Burger explained, was moving into a new orbit. In the future, a few giant Internet companies would operate a few giant Internet services so complex and so different from what came before that these companies would have to build a whole new architecture to run them. They would create not just the software driving these services, but the hardware, including servers and networking gear. Project Catapult would equip all of Microsoft's servers—millions of them—with specialized chips that the company could reprogram for particular tasks.

But before Burger could even get to the part about the chips, Ballmer looked up from his laptop. When he visited Microsoft Research, Ballmer said, he expected updates on R&D, not a strategy briefing. "He just started grilling me," Burger says. Microsoft had spent 40 years building PC software like Windows, Word, and Excel. It was only just finding its feet on the Internet. And it certainly didn't have the tools and the engineers needed to program computer chips—a task that's difficult, time consuming, expensive, and kind of weird. Microsoft programming computer chips was like Coca Cola making shark fin soup.



The current incarnation of Project Catapult. CLAYTON COTTERELL FOR WIRED

Burger—trim, only slightly bald, and calmly analytical, like so many good engineers—pushed back. He told Ballmer that companies like Google and Amazon were already moving in this direction. He said the world's hardware makers wouldn't provide what Microsoft needed to run its online services. He said that Microsoft would fall behind if it didn't build its own hardware. Ballmer wasn't buying it. But after awhile, another voice joined the discussion. This was Qi Lu, who runs Bing, Microsoft's search engine. Lu's team had been talking to Burger about reprogrammable computer chips for almost two years. Project Catapult was more than possible, Lu said: His team had already started.

Today, the programmable chips that Burger and Lu believed would transform the world—called field programmable gate arrays—are here. FPGAs already underpin Bing, and in the coming weeks, they will drive new search algorithms based on deep neural networks—artificial intelligence modeled on the structure of the human brain

—executing this AI several orders of magnitude faster than ordinary chips could. As in, 23 milliseconds instead of four seconds of nothing on your screen. FPGAs also drive Azure, the company's cloud computing service. And in the coming years, almost every new Microsoft server will include an FPGA. That's millions of machines across the globe. "This gives us massive capacity and enormous flexibility, and the economics work," Burger says. "This is now Microsoft's standard, worldwide architecture."

Catapult team members Adrian Caulfield, Eric Chung, Doug Burger, and Andrew Putnam CLAYTON COTTERELL FOR WIRED

This isn't just Bing playing catch-up with Google. Project Catapult signals a change in how global systems will operate in the future. From Amazon in the US to Baidu in China, all the Internet giants are supplementing their standard server chips—central processing units, or CPUs—with alternative silicon that can keep pace with the rapid changes in AI. Microsoft now spends between $5 and $6 billion a year for the hardware needed to run its online empire. So this kind of work is "no longer just research," says Satya Nadella, who took over as Microsoft's CEO in 2014. "It's an essential priority." That's what Burger was trying to explain in Building 99. And it's

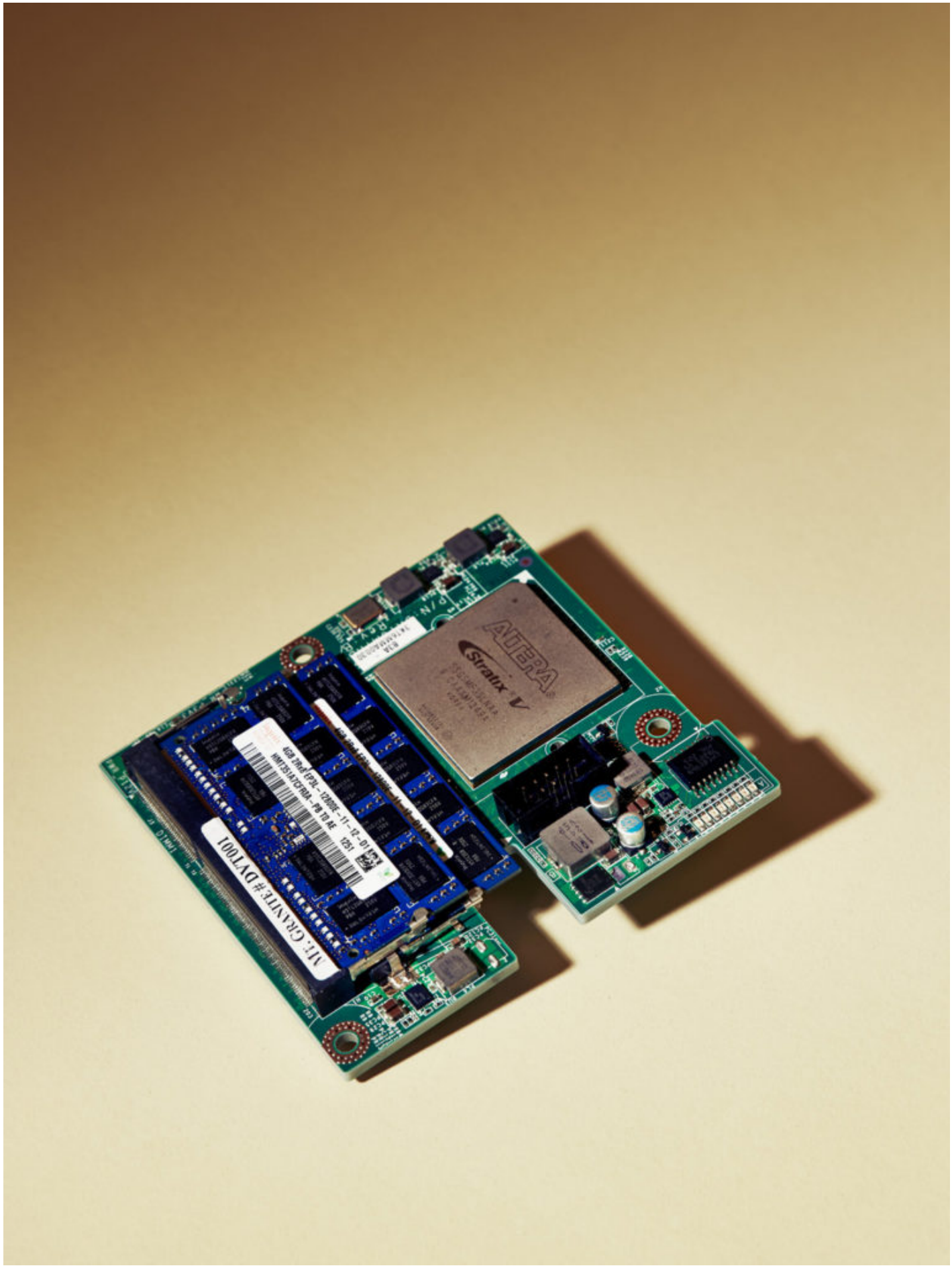
what drove him and his team to overcome years of setbacks, redesigns, and institutional entropy to deliver a new kind of global supercomputer.

A Brand New, Very Old Kind of Computer Chip

In December of 2010, Microsoft researcher Andrew Putnam had left Seattle for the holidays and returned home to Colorado Springs. Two days before Christmas, he still hadn't started shopping. As he drove to the mall, his phone rang. It was Burger, his boss. Burger was going to meet with Bing execs right after the holiday, and he needed a design for hardware that could run Bing's machine learning algorithms on FPGAs.

Putnam pulled into the nearest Starbucks and drew up the plans. It took him about five hours, and he still had time for shopping.

Burger, 47, and Putnam, 39, are both former academics. Burger spent nine years as a professor of computer science at the University of Texas, Austin, where he specialized in microprocessors and designed a new kind of chip called EDGE. Putnam had worked for five years as a researcher at the University of Washington, where he experimented with FPGAs, programmable chips that had been around for decades but were mostly used as a way of prototyping other processors. Burger brought Putnam to Microsoft in 2009, where they started exploring the idea that these chips could actually accelerate online services.


ALTERA
Stratix V
ALTERA#DVT001

Project Catapult Version 1, or V1, the hardware that Doug Burger and team tested in a data center on Microsoft's Seattle campus. CLAYTON COTTERELL FOR WIRED

Even their boss didn't buy it. "Every two years, FGPAs are 'finally going to arrive,'" says Microsoft Research vice president Peter Lee, who oversees Burger's group. "So, like any reasonable person, I kind of rolled my eyes when this was pitched." But Burger and his team believed this old idea's time had come, and Bing was the perfect test case.

Microsoft's search engine is a single online service that runs across thousands of machines. Each machine is driven by a CPU, and though companies like Intel continue to improve them, these chips aren't keeping pace with advances in software, in large part because of the new wave in artificial intelligence. Services like Bing have outstripped Moore's Law, the canonical notion that the number of transistors in a processor doubles every 18 months. Turns out, you can't just throw more CPUs at the problem.

But on the other hand, it's generally too expensive to create specialized, purpose-built chips for every new problem. FPGAs bridge the gap. They let engineers build chips that are faster and less energy-hungry than an assembly-line, general-purpose CPU, but customizable so they handle the new problems of ever-shifting technologies and business models.

At that post-holiday meeting, Burger pitched Bing's execs on FPGAs as a low-power way of accelerating searches. The execs were noncommittal. So over the next several months, Burger and team took Putnam's Christmas sketch and built a prototype, showing that it could run Bing's machine learning algorithms about 100 times faster. "That's when they really got interested," says Jim Larus, another member of the team back then who's now a dean at Switzerland's École Polytechnique Fédérale in Lausanne. "They also started giving us a really hard time."

The prototype was a dedicated box with six FPGAs, shared by a rack full of servers. If the box went on the frizz, or if the machines needed more than six FPGAs—increasingly likely given the complexity of the machine learning models—all those machines were out of luck. Bing's engineers hated it. "They were right," Larus says.

So Burger's team spent many more months building a second prototype. This one was a circuit board that plugged into each server and included only one FPGA. But it also connected to all the other FPGA boards on all the other servers, creating a giant pool of programmable chips that any Bing machine could tap into.

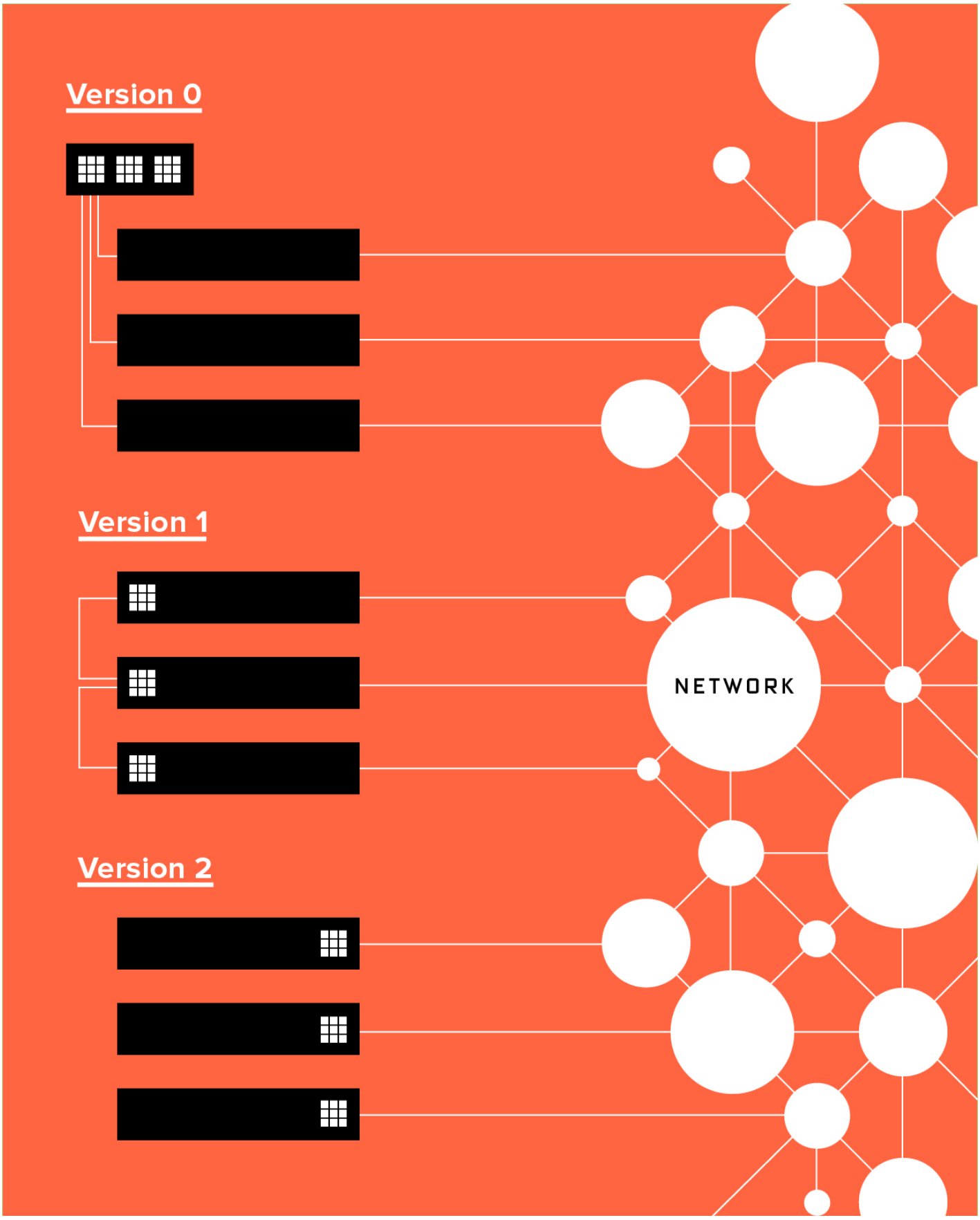
That was the prototype that got Qi Lu on board. He gave Burger the money to build and test over 1,600 servers equipped with FPGAs. The team spent six months building the hardware with help from manufacturers in China and Taiwan, and they installed the first rack in an experimental data center on the Microsoft campus. Then, one night, the fire suppression system went off by accident. They spent three days getting the rack back in shape—but it still worked.

Over several months in 2013 and 2014, the test showed that Bing's "decision tree" machine-learning algorithms ran about 40 times faster with the new chips. By the summer of 2014, Microsoft was publicly saying it would soon move this hardware into its live Bing data centers. And then the company put the brakes on.

Searching for More Than Bing

Bing dominated Microsoft's online ambitions in the early part of the decade, but by 2015 the company had two other massive online services: the business productivity suite Office 365 and the cloud computing service Microsoft Azure. And like all of their competitors, Microsoft executives realized that the only efficient way of running a growing online empire is to run all services on the same foundation. If Project Catapult was going to transform Microsoft, it couldn't be exclusive to Bing. It had to work inside Azure and Office 365, too.

The problem was, Azure executives didn't care about accelerating machine learning. They needed help with networking. The traffic bouncing around Azure's data centers was growing so fast, the service's CPUs couldn't keep pace. Eventually, people like Mark Russinovich, the chief architect on Azure, saw that Catapult could help with this too—but not the way it was designed for Bing. His team needed programmable chips right where each server connected to the primary network, so they could process all that traffic before it even got to the server.



The first prototype of the FPGA architecture was a single box shared by a rack of servers (Version 0). Then the team switched to giving individual servers their own FPGAs (Version 1). And then they put the chips between the servers and the overall network (Version 2). WIRED

So the FPGA gang had to rebuild the hardware again. With this third prototype, the chips would sit at the edge of each server, plugging directly into the network, while still creating pool of FPGAs that was available for any machine to tap into. That started to look like something that would work for Office 365, too. Project Catapult was ready to go live at last.

Larus describes the many redesigns as an extended nightmare—not because they had to build a new hardware, but because they had to reprogram the FPGAs every time. "That is just horrible, much worse than programming software," he says. "Much more difficult to write. Much more difficult to get correct." It's finicky work, like trying to change tiny logic gates on the chip.

Now that the final hardware is in place, Microsoft faces that same challenge every time it reprograms these chips. "It's a very different way of seeing the world, of thinking about the world," Larus says. But the Catapult hardware costs less than 30 percent of everything else in the server, consumes less than 10 percent of the power, and processes data twice as fast as the company could without it.

The rollout is massive. Microsoft Azure uses these programmable chips to route data. On Bing, which an estimated 20 percent of the worldwide search market on desktop machines and about 6 percent on mobile phones, the chips are facilitating the move to the new breed of AI: deep neural nets. And according to one Microsoft employee, Office 365 is moving toward using FPGAs for encryption and compression as well as machine learning—for all of its 23.1 million users. Eventually, Burger says, these chips will power all Microsoft services.

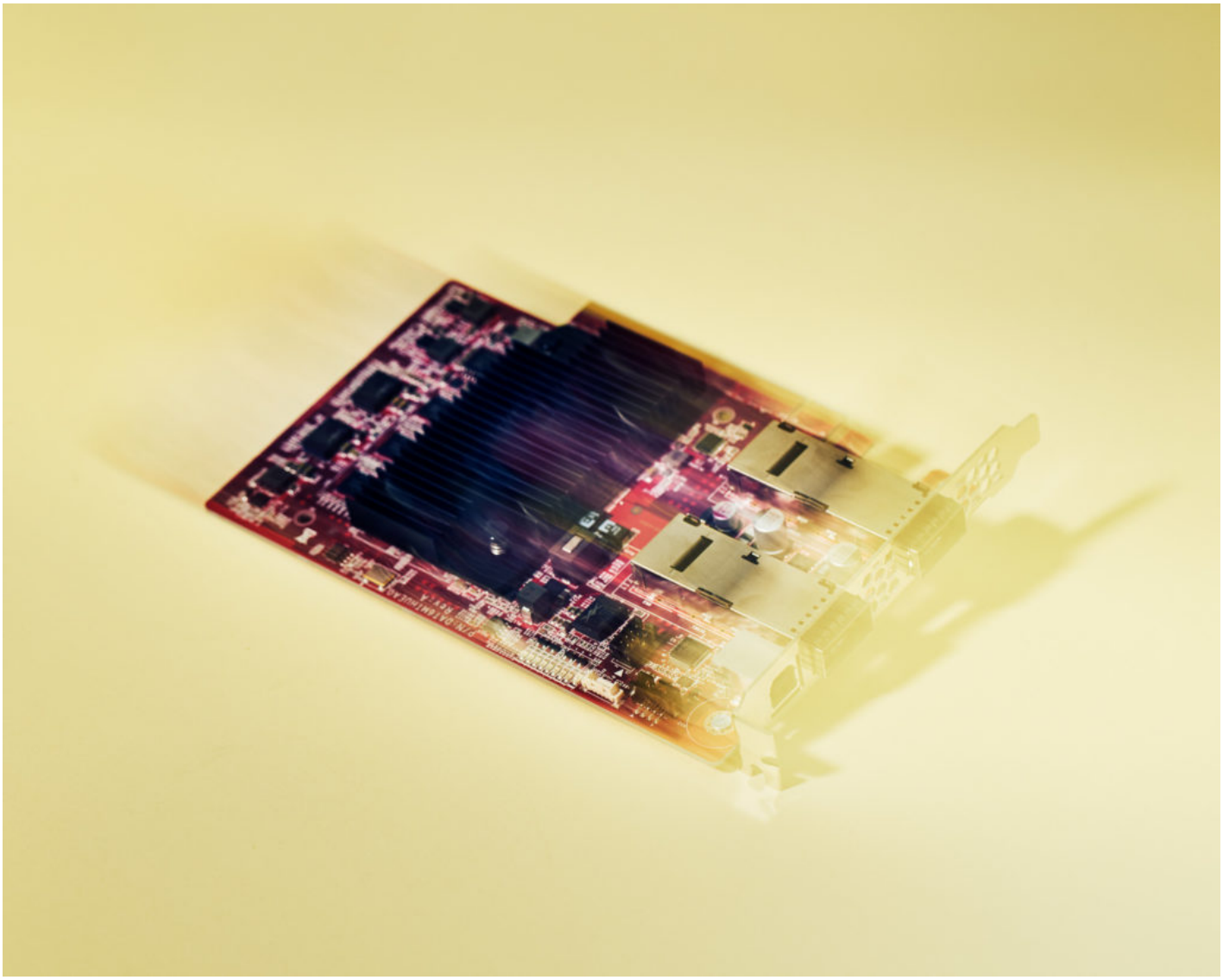
Wait—This Actually Works?

"It still stuns me," says Peter Lee, "that we got the company to do this." Lee oversees an organization inside Microsoft Research called NExT, short for New Experiences and Technologies. After taking over as CEO, Nadella personally pushed for the creation of this new organization, and it represents a significant shift from the 10-year reign of Ballmer. It aims to foster research that can see the light of day sooner rather than later—that can change the course of Microsoft now rather than years from now. Project Catapult is a prime example. And it is part of a much larger change across the industry. "The leaps ahead," Burger says, "are coming from non-CPU technologies."

Peter Lee. CLAYTON COTTERELL FOR WIRED

All the Internet giants, including Microsoft, now supplement their CPUs with graphics processing units, chips designed to render images for games and other highly visual applications. When these companies train their neural networks to, for example, recognize faces in photos—feeding in millions and millions of pictures—GPUs handle much of the calculation. Some giants like Microsoft are also using alternative silicon to execute their neural networks after training. And even though it's crazily expensive to custom-build chips, Google has gone so far as to design its own processor for executing neural nets, the tensor processing unit.

With its TPUs, Google sacrifices long-term flexibility for speed. It wants to, say, eliminate any delay when recognizing commands spoken into smartphones. The trouble is that if its neural networking models change, Google must build a new chip. But with FPGAs, Microsoft is playing a longer game. Though an FPGA isn't as fast as Google's custom build, Microsoft can reprogram the silicon as needs change. The company can reprogram not only for new AI models, but for just about any task. And if one of those designs seems likely to be useful for years to come, Microsoft can always take the FPGA programming and build a dedicated chip.



A newer version of the final hardware, V2, a card that slots into the end of each Microsoft server and connects directly to the network. CLAYTON COTTERELL FOR WIRED

Microsoft's services are so large, and they use so many FPGAs, that they're shifting the worldwide chip market. The FPGAs come from a company called Altera, and Intel executive vice president Diane Bryant tells me that Microsoft is why Intel acquired Altera last summer—a deal worth $16.7 billion, the largest acquisition in the history of the largest chipmaker on Earth. By 2020, she says, a third of all servers inside all the major cloud computing companies will include FPGAs.

It's a typical tangle of tech acronyms. CPUs. GPUs. TPUs. FPGAs. But it's the subtext that matters. With cloud computing, companies like Microsoft and Google and Amazon are driving so much of the world's technology that those alternative chips will drive the wider universe of apps and online services. Lee says that Project Catapult will allow Microsoft to continue expanding the powers of its global

supercomputer until the year 2030. After that, he says, the company can move toward quantum computing.

Later, when we talk on the phone, Nadella tells me much the same thing. They're reading from the same Microsoft script, touting a quantum-enabled future of ultrafast computers. Considering how hard it is to build a quantum machine, this seems like a pipe dream. But just a few years ago, so did Project Catapult.

Correction: This story originally implied that the Hololens headset was part of Microsoft's NExT organization. It was not.