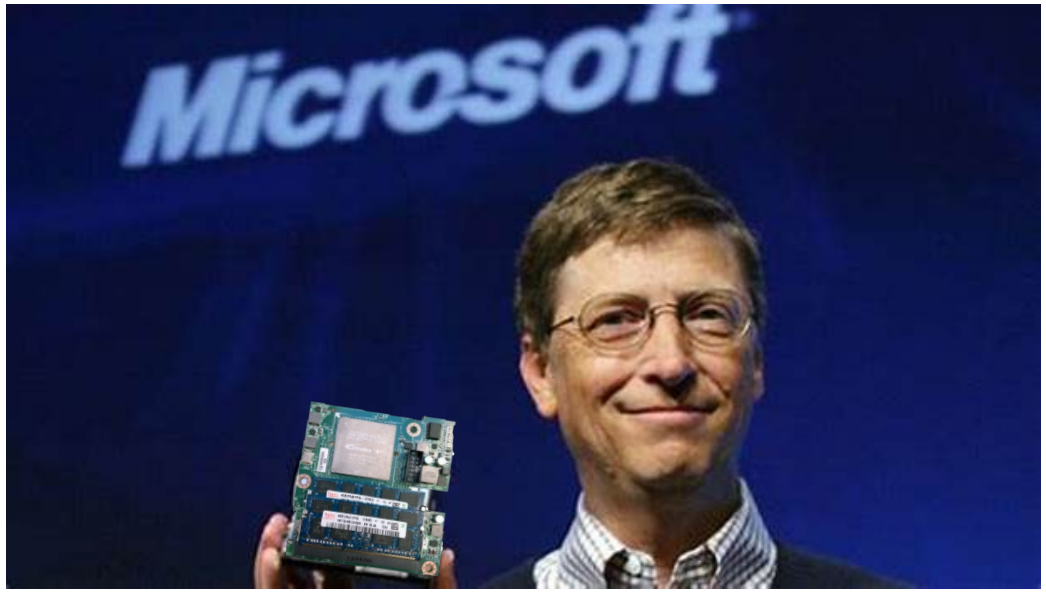
NANDA 2024:

Think Different: The Unique Role of FPGAs in a GPU-Dominated World

Sept 9, 2024

Andrew Putnam
Partner GM, Cloud & AI Hardware Engineering
Microsoft Azure

The AI Era

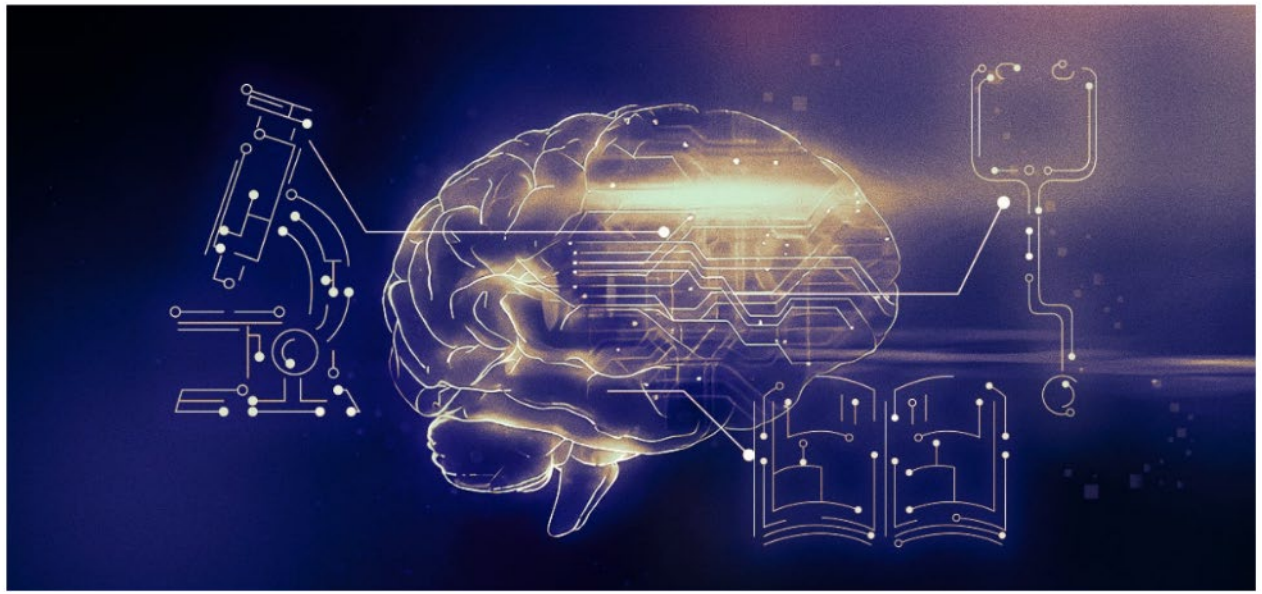


A NEW ERA

The Age of AI has begun

Artificial intelligence is as revolutionary as mobile phones and the Internet.

By **Bill Gates** | March 21, 2023 • 14 minute read



In my lifetime, I've seen two demonstrations of technology that struck me as revolutionary.

The first time was in 1980, when I was introduced to a graphical user interface…

The second big surprise came just last year.

https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

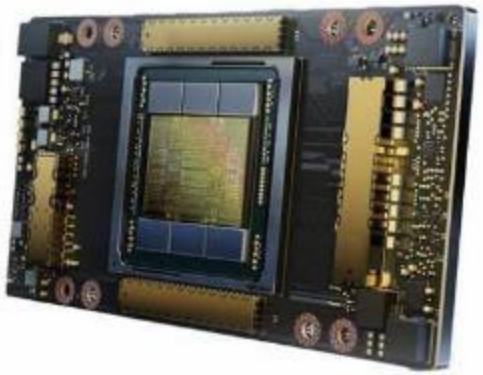
Eras Tour – Computing Edition

Mainframe

PC

Mobile

Cloud



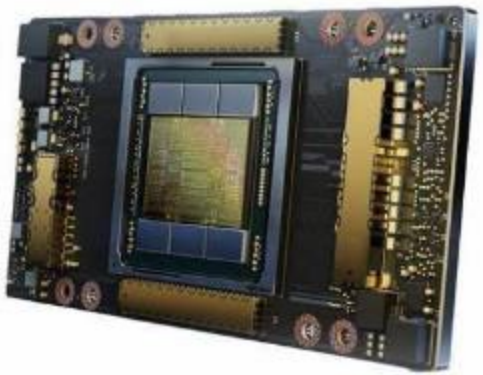
AI

ASIC Prototype

Networking & Telecom

Specialized Compute

Cloud



AI

Success in Prior Eras

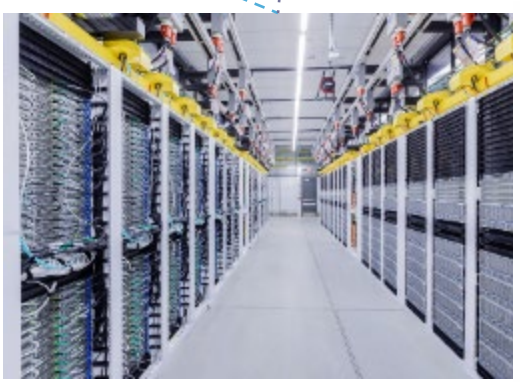
Be First

Be Special

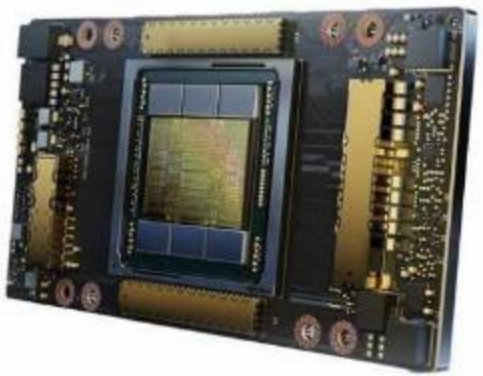
ASIC Prototype

Networking & Telecom

Specialized Compute

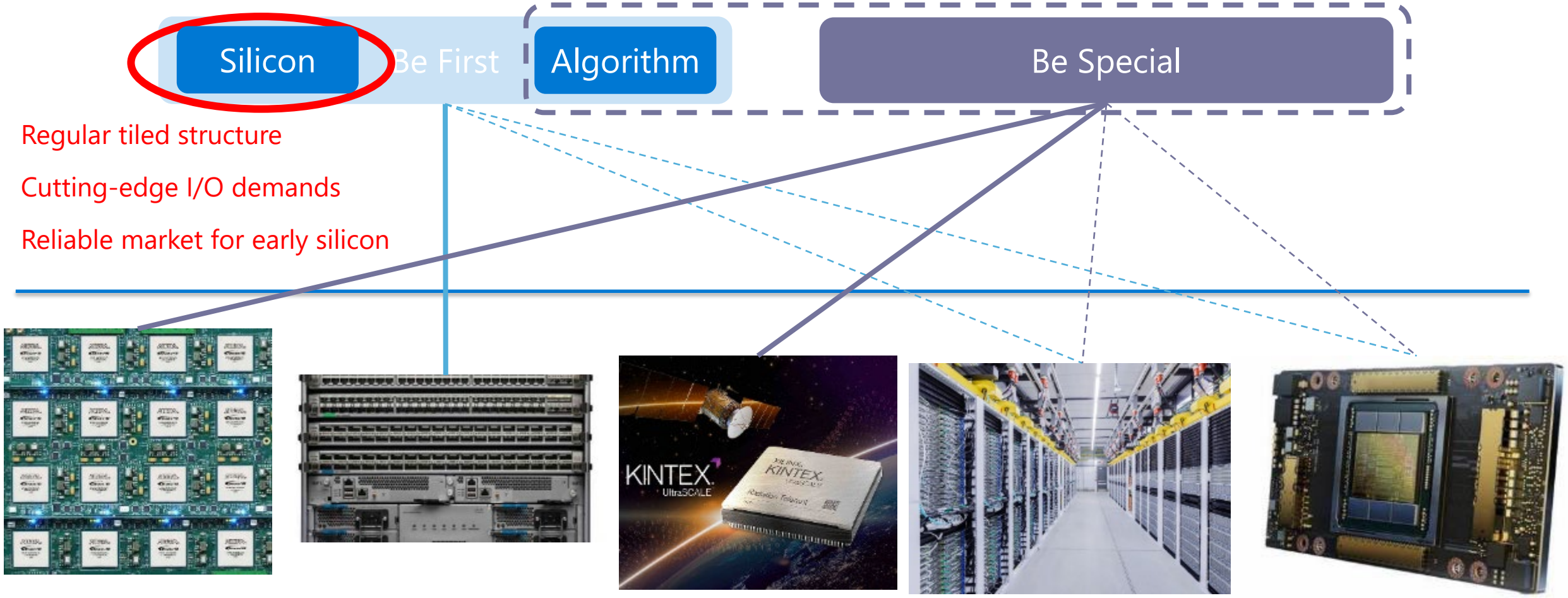


Cloud



AI

Success in Prior Eras



Can FPGAs compete when the biggest players in computing are focused on the ***same things*** as FPGA researchers and developers?

Be First

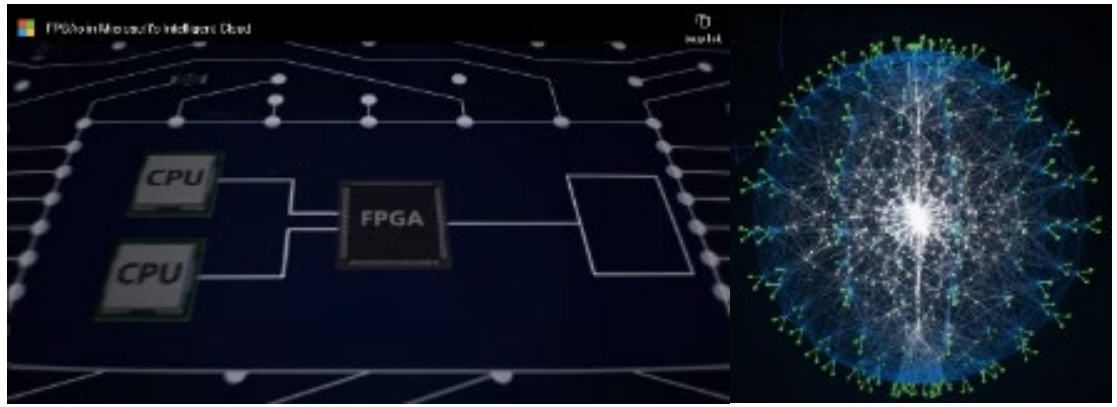
Be Special

Can FPGAs compete when the biggest players in computing are focused on the ***same things*** as FPGA researchers and developers?

Be Special

FPGA Target Markets

Cloud / Datacenter

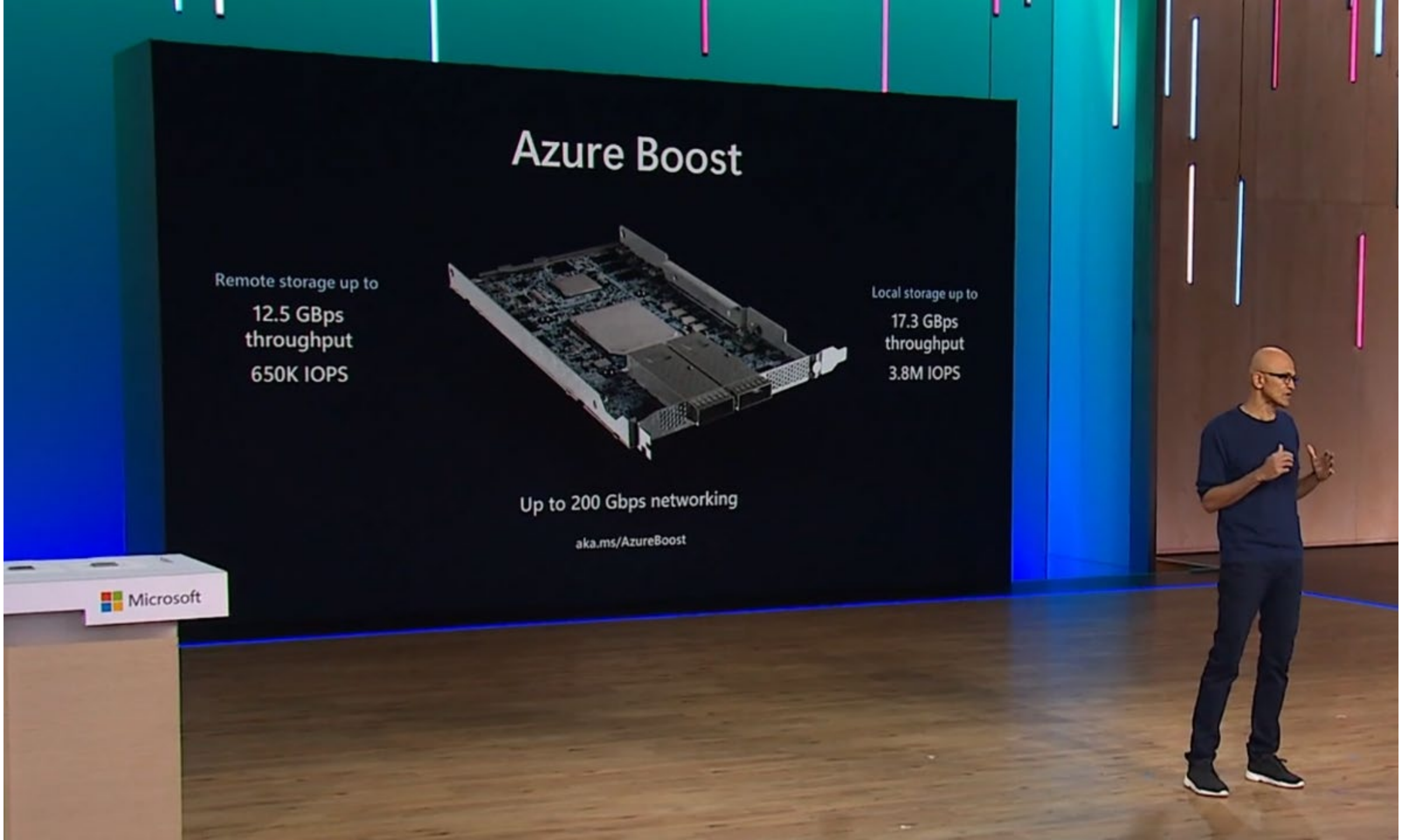


Azure Accelerated Networking / Azure Boost

AI / ML

Project Brainwave

Microsoft


Azure Boost
Remote storage up to
12.5 GBps
throughput
650K IOPS
Local storage up to
17.3 GBps
throughput
3.8M IOPS
Up to 200 Gbps networking
aka.ms/AzureBoost
Microsoft

Microsoft Bing

Azure Boost

anputnam@microsoft.com 111965 $0.00

SEARCH CHAT WORK SHOPPING IMAGES Inspiration Create Collections VIDEOS MAPS NEWS MORE

SafeSearch: Moderate Filter

Microsoft previews Azure Boost for offloading virtualized network functions - Converge Dige...

Adidas Yeezy Boost 700 'Fad...

Adidas Yeezy Boost 380 Azure FZ4986 | Sneakerbaron NL

adidas Yeezy Boost 380 Azure FZ4986 Release Date | SneakerNews.com

adidas Yeezy Boost 700 Faded Azure GZ2002

adidas Yeezy Boost 700 Faded Azure GZ2002

adidas Yeezy Boost 700 'Faded Azure' For Sale – T...

Microsoft makes Azure Bot Service generally available for developers

First Look at the adidas Yeezy Boost 380 Azure • KicksOnFi...

Release Details: Yeezy BOOST 700 'Faded Azure' - Sneaker Freaker

adidas Yeezy Boost 700 Faded Azure GZ2002

adidas Yeezy Boost 380 Azure FZ4986 Release Date Info | Sneak...

First Look at the adidas Yeezy Boost 380 Azure • KicksOnFire.com

adidas Yeezy Boost 700 Faded Azure GZ2002

adidas Yeezy Boost 700 Faded Azure Release Date - SBD

Azure Boost Launches to Offload Virtualization Processes

adidas YEEZY BOOST 380 'Azure' (Azure/Azure/Azure) - FZ4966 - Co...

The adidas Yeezy Boost 700 "Fade Azure" Releases Tomorrow | Laptrinh...

Microsoft Announces Azure Boost, Enhancing Virtual Machine Performance

adidas Yeezy Boost 380 Azure FZ4986 Release Date - Sneaker Ba...

adidas Yeezy Boost 700 Faded Azure Release Date - ...

adidas Yeezy Boost 380 Azure - FZ4986

Adidas Yeezy Boost 700 'Faded Azure' GZ2002 R...

The new Microsoft Azure Boost delivers accelerated improvements in performance, s...

adidas Yeezy Boost 380 Azure FZ4986 Release Date - Sneaker Bar Detroit

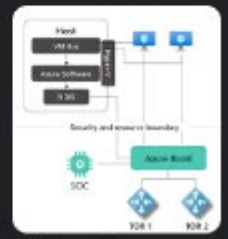
adidas Yeezy Boost 700 Faded Azure Release Date Info - SNKRS WORLD

Google Azure Boost

All | Images | Shopping | Videos | News | More | Tools | Saved | SafeSearch

yeezy 700 v2 | boost 700 faded azure | microsoft | microsoft azure | yeezy 380 azure | boost 380 | yeezy fade | yeezy 350 | wave runner | kanye v


Data Center Dynamics
Microsoft announces Azure...


Microsoft Learn
Overview of Azure Boo...


Flight Club · In stock
Adidas Mens Yeezy Boost 700 GZ...


StockX · In stock
Yeezy Boost 700 Faded Azure · GZ2...


Cotton Novas · In stock
Yeezy Boost 700 Faded Azure – GZ2002...


Sneaker Bar Detroit
adidas Yeezy Boost 70...


Reborn. · In stock
adidas Yeezy Boost 700 'Faded Azure...

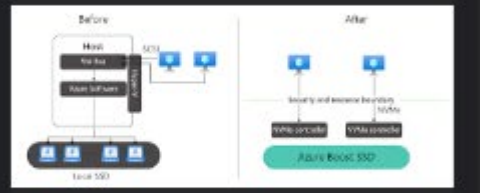

KicksOnFire.com
adidas Yeezy Boost 700 Faded A...


Microsoft Learn
Overview of Azure Boost | Microsoft Learn


Sneaker News
adidas Yeezy Boost 700 Fade ...


Sneaker Bar Detroit
adidas Yeezy Boost 70...


KicksOnFire.com
adidas Yeezy Boost 700 Faded Azu...


Microsoft Learn
Overview of Azure Boost | Microsoft Learn


YouTube
Introducing Microsoft Azure Boost - YouTube

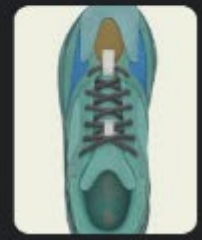

StockX · In stock
Yeezy - Women's Boost 380 Azure L...


Sneaker Freaker
Yeezy BOOST 700 'Faded Azure ...


Cotton Novas · In stock
Yeezy Boost 380 Azure – FZ4986, 5


StockX · In stock
Adidas Yeezy Boost 380 Azure Ref...


Sneaker Bar Detroit
adidas Yeezy Boost 700 Fa...



Related searches

yeezy 700 faded azure on feet

yeezy 700 faded azure outfit

Announcing the general availability of Azure Boost

By Max Uritsky
Published Nov 15 2023 08:00 AM

We're excited today to announce the general availability of Azure Boost, a system **designed by Microsoft** that offloads server virtualization processes traditionally performed by the hypervisor and host OS onto purpose-built software and hardware, **enabling faster storage and networking performance for Azure VM customers.**

Going forward, **every new Azure virtual machine (VM) series will benefit from Azure Boost technologies**, making the networking and storage components of your virtual workloads run even faster, whether you have deployed general purpose compute workloads or specialized AI clusters.

Azure Boost was engineered by a team of Microsoft hardware and software engineers to enhance the performance, security, and reliability of Microsoft Azure, and is already being used at Microsoft datacenters around the world, **delivering benefits to millions of customers VMs in production today.**

To learn more, watch the product overview video, read the documentation, or experience Azure Boost benefits today by provisining Azure Boost enabled VMs sizes listed in the documentation. Continue reading to learn more about the benefits of using Azure Boost.

Pictured below: the Azure Boost Card

Azure Boost

Azure Boost enables customers to achieve a **remote storage throughput up to 12.5 GBps and 650K IOPS and a local storage throughput of up to 17.3 GBps and 3.8 million IOPS,** enabling the fastest storage workloads available today.

In addition, Azure customers can to achieve a up to **200 Gbps networking throughput**.

https://azure.microsoft.com/en-us/updates/public-preview-new-generation-amd-vms-eav6dav6fav6

https://techcommunity.microsoft.com/t5/azure-compute-blog/announcing-preview-of-new-azure-dlsv6-dsv6-esv6-vms-with-new-cpu/ba-p/4192124

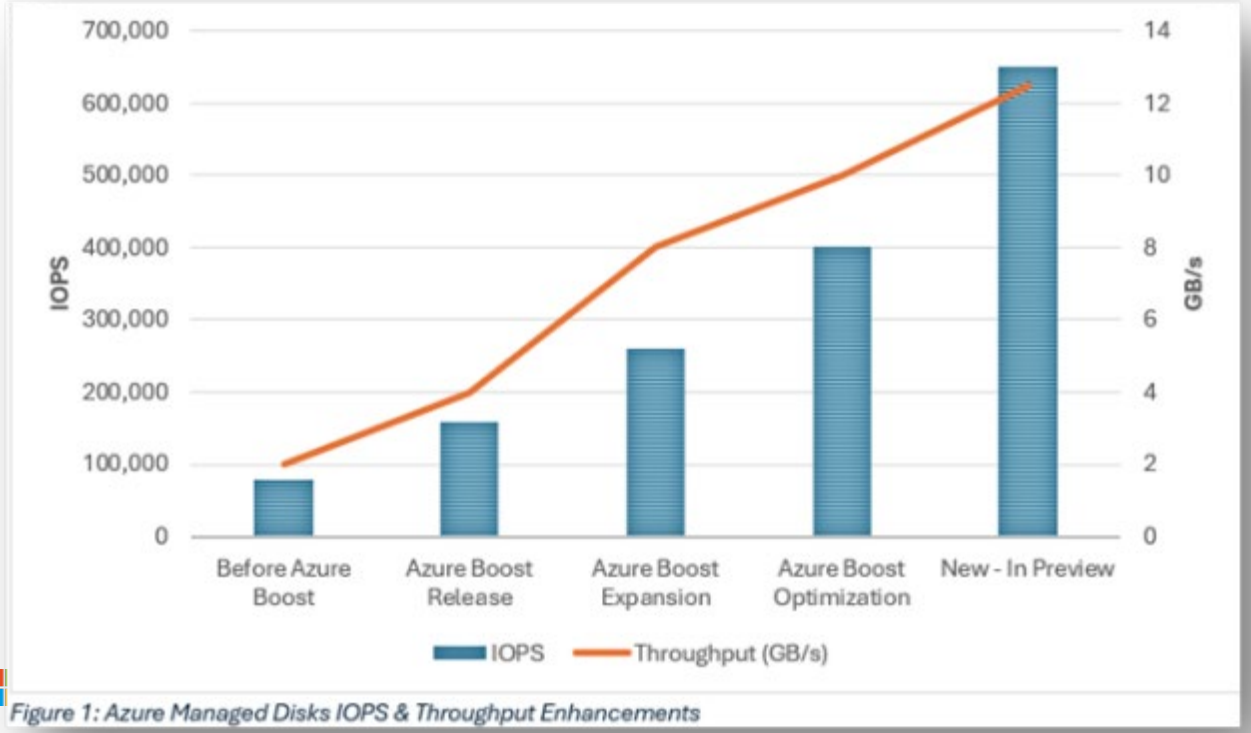
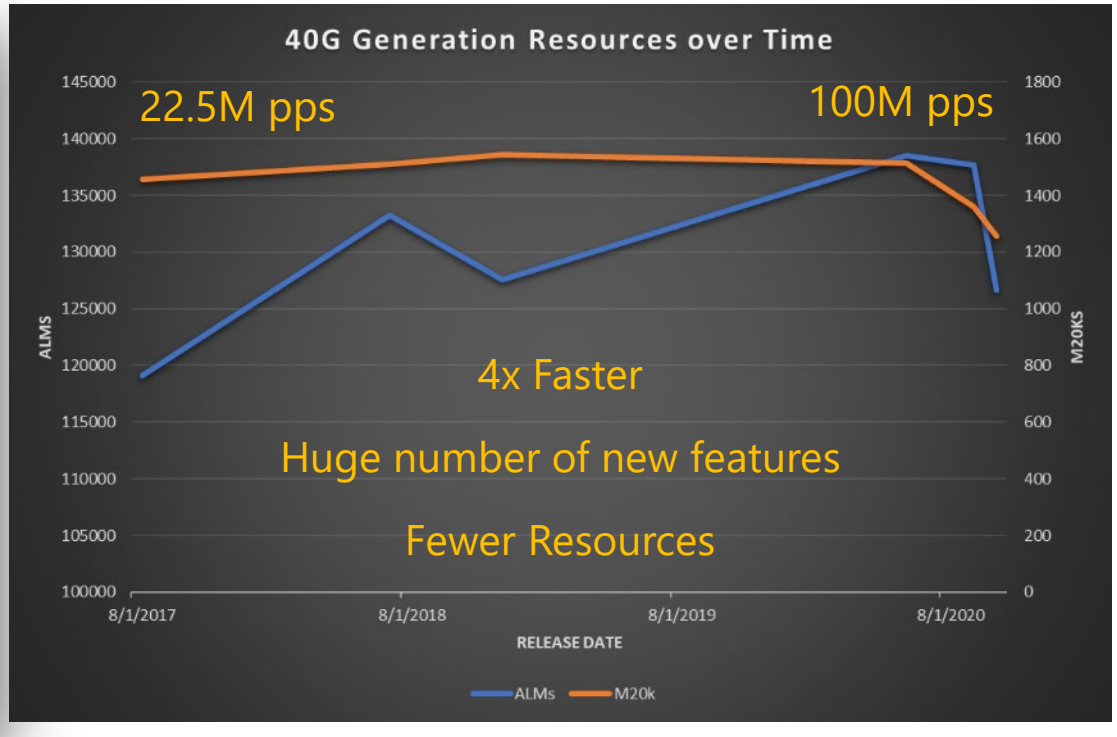
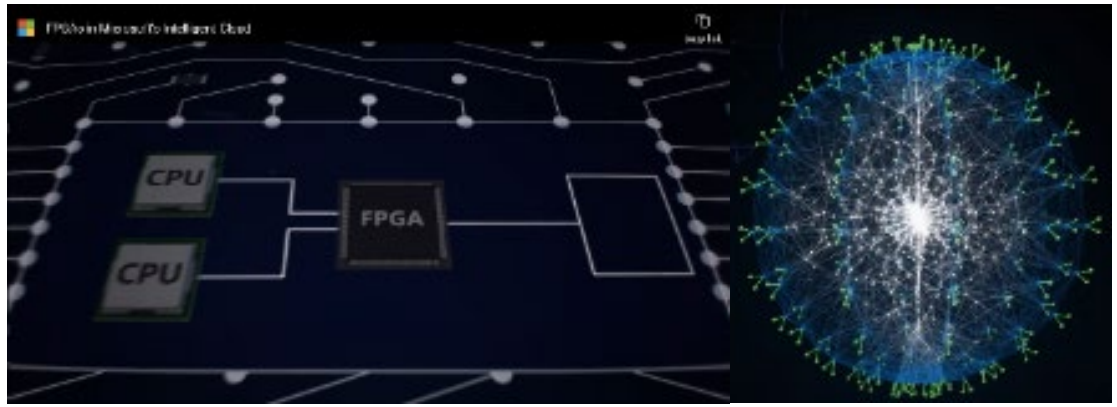


Figure 1: Azure Managed Disks IOPS & Throughput Enhancements



FPGA Target Markets

Cloud / Datacenter



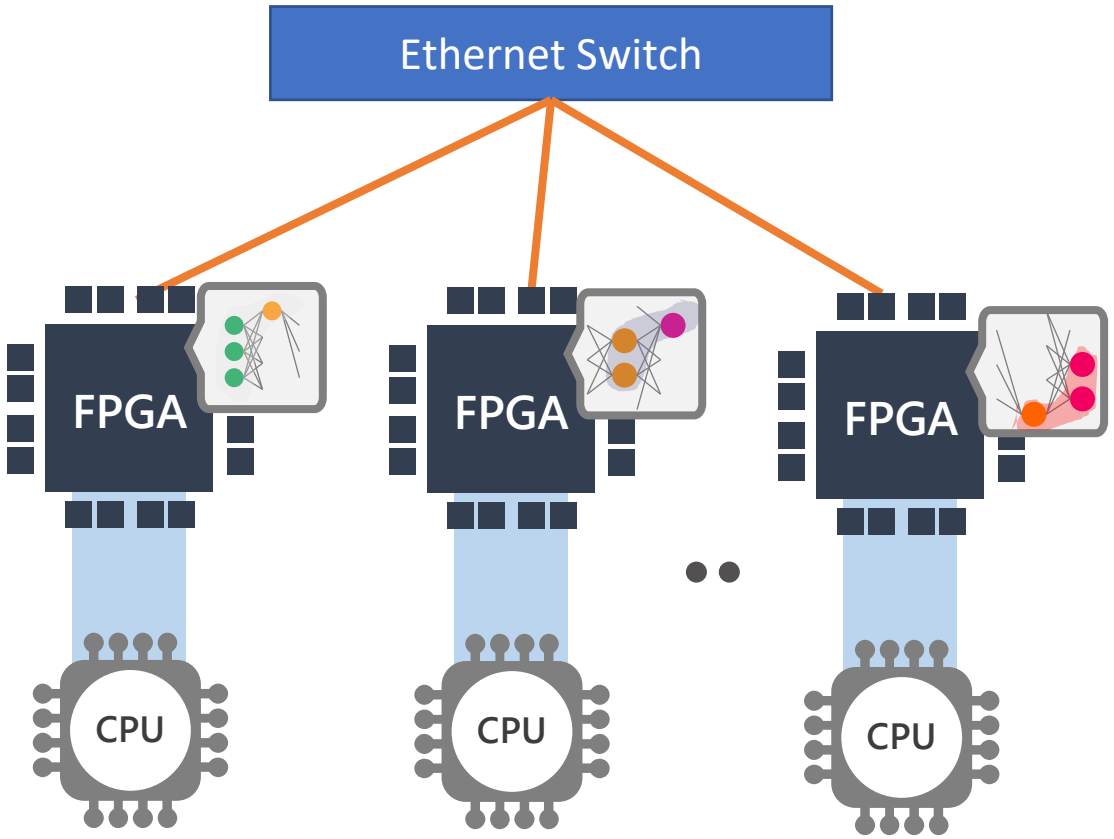
Azure Accelerated Networking / Azure Boost

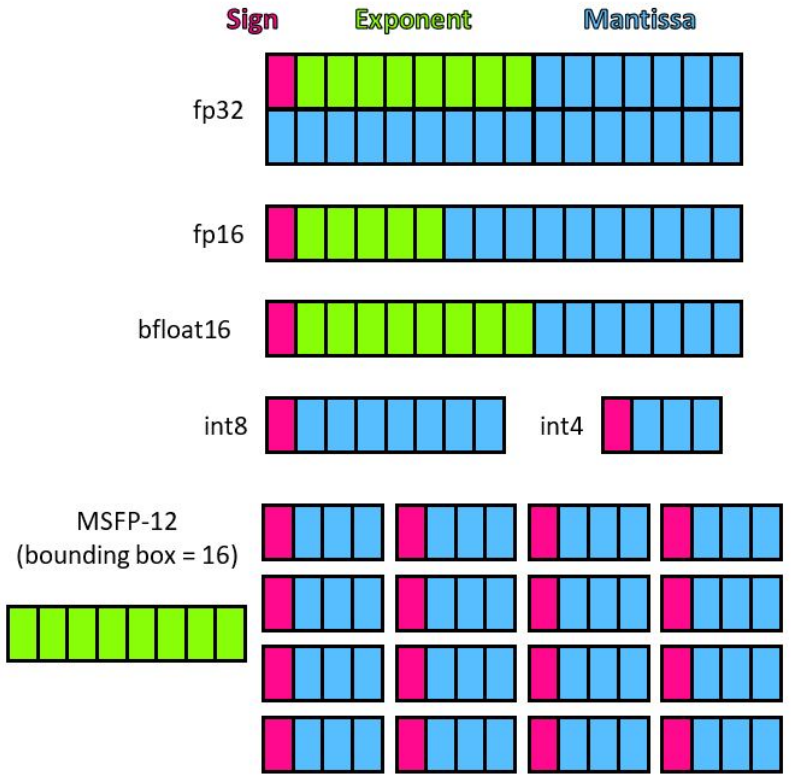
AI / ML

Project Brainwave

Microsoft

What is special in BrainWave?





Low batch size optimizations

RDMA transport for AI Networks

Custom Data Types for Efficient Inference

(MSFP / MXFP)

AI Model Evolution

Boosted Decision Trees (2010)

Deep CNNs (2014)

LSTMs/ GRUs (2017)

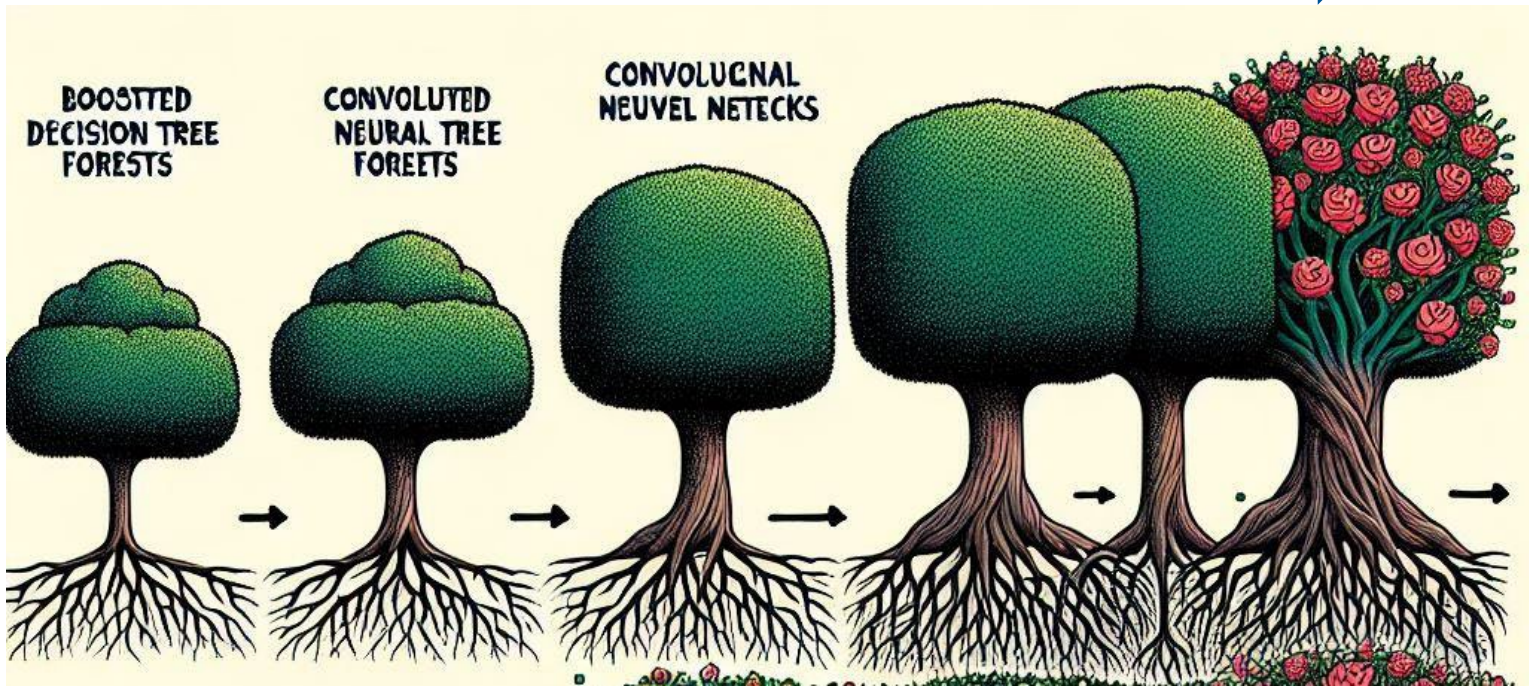
Transformers (2019+)

Sparse transformers?

Sparse mixture of experts?

Dynamic, recursive networks?

Graph neural networks?



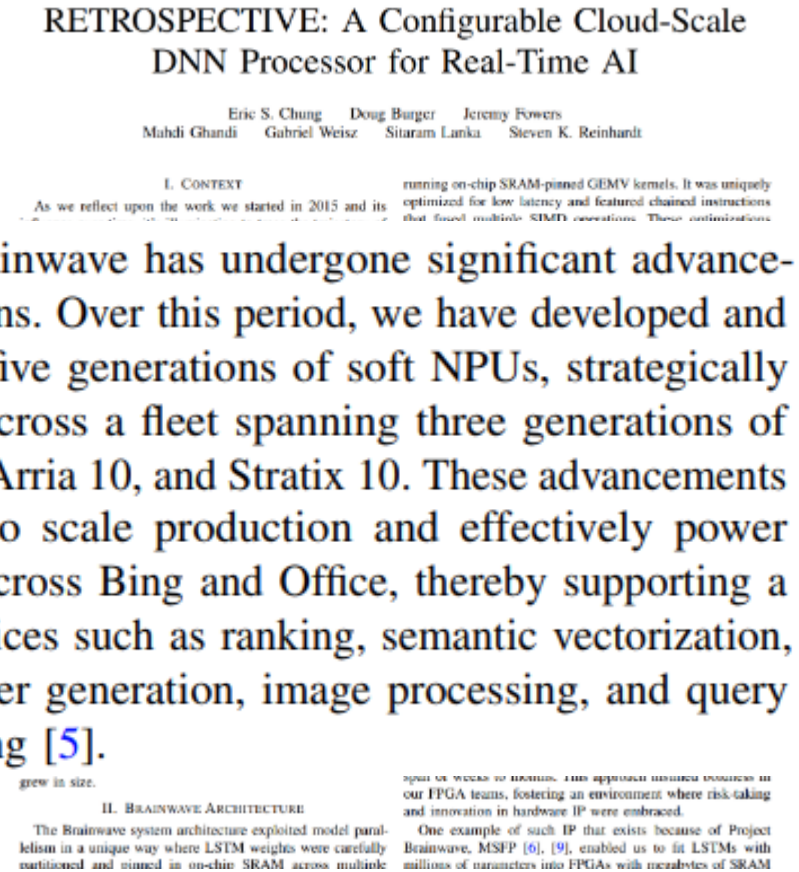
Largest deployment of AI FPGAs

2018

- A configurable cloud-scale DNN processor for real-time AI

by Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Todd Massengill, Ming Liu, Daniel Lo, Shlomi Alkalay, Michael Haselman, Logan Adams, Mahdi Ghandi, Stephen Heil, Prerak Patel, Adam Sapek, Gabriel Weisz, Lisa Woods, Sitaram Lanka, Steven K. Reinhardt, Adr

[Retrospective]



RETROSPECTIVE: A Configurable Cloud-Scale DNN Processor for Real-Time AI

Eric S. Chung Doug Burger Jeremy Fowers
Mahdi Ghandi Gabriel Weisz Sitaram Lanka Steven K. Reinhardt

I. CONTEXT

As we reflect upon the work we started in 2015 and its

running on-chip SRAM-pinned GEMV kernels. It was uniquely optimized for low latency and featured chained instructions

Since 2015, Brainwave has undergone significant advancements and evolutions. Over this period, we have developed and deployed at least five generations of soft NPUs, strategically integrating them across a fleet spanning three generations of FPGAs: Stratix V, Arria 10, and Stratix 10. These advancements have enabled us to scale production and effectively power over 100 models across Bing and Office, thereby supporting a wide range of services such as ranking, semantic vectorization, question and answer generation, image processing, and query intent understanding [5].

grew in size.

II. BRAINWAVE ARCHITECTURE

The Brainwave system architecture exploited model parallelism in a unique way where LSTM weights were carefully

our FPGA teams, fostering an environment where risk-taking and innovation in hardware IP were embraced.

One example of such IP that exists because of Project Brainwave, MSFP [6], [9], enabled us to fit LSTMs with

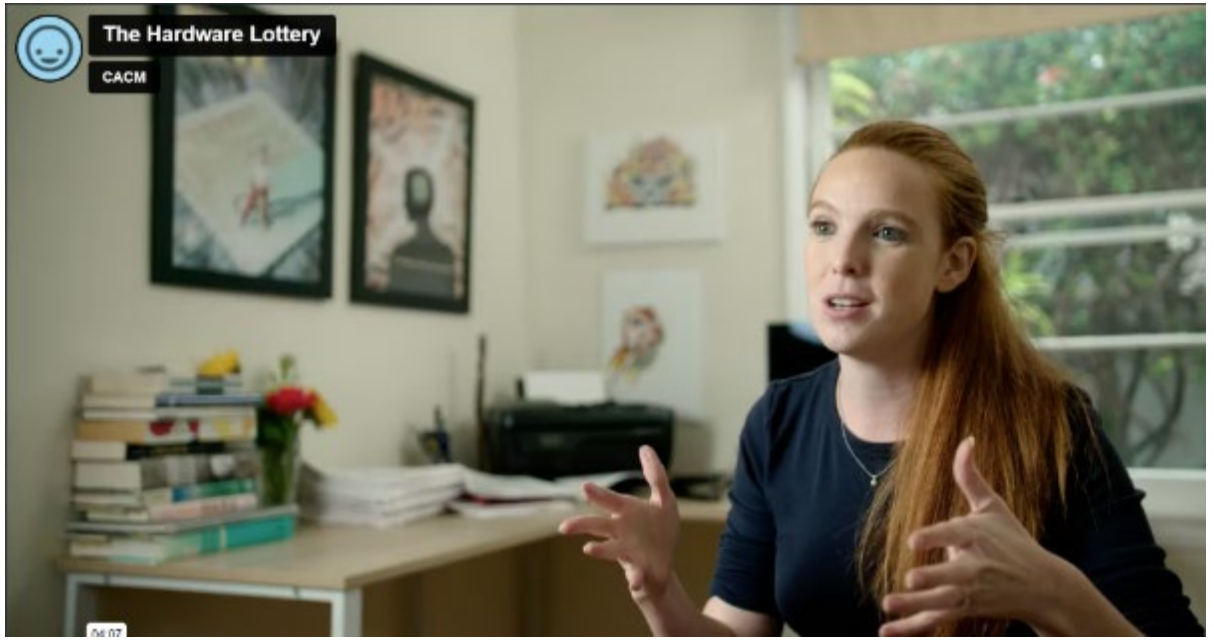
IV. PITFALLS AND SURPRISES

Our project was not devoid of detours and surprises. The on-chip SRAM approach introduced by Project Brainwave gradually lost its relevance as the compute-bound Transformer-based models surpassed the capacity of on-chip memory hierarchies. These models no longer necessitated aggressive, SRAM-based model parallelism to achieve high utilization. Additionally, Nvidia GPUs aggressively scaled their High-Bandwidth Memory (HBM) and hardened tensor units to improve their AI performance, which eventually became predominant in most mainstream AI processors.

This limited the opportunity space for novel architectures to compete, and despite the development of the Stratix 10 NX FPGA [1], which introduced an HBM-enabled FPGA with hard tensor cores, the scope and scale at which Catapult FPGAs could competitively serve flagship AI models were reduced. Ultimately, the hardware lottery [7] effect made mainstream GPUs a much more attractive target for model evolution.

Microsoft

The Hardware Lottery



Sara Hooker

Key Insights

- The term hardware lottery describes a research idea that wins due to its compatibility with available software and hardware, not its superiority over alternative research directions.
- We may be in the midst of a present-day hardware lottery. Hardware design has prioritized delivering on commercial use cases, while built-in flexibility to accommodate the next generation of ideas remains a secondary consideration.
- Any attempt to avoid future hardware lotteries must be concerned with making it cheaper and less time consuming to explore different hardware/software/algorithm combinations.

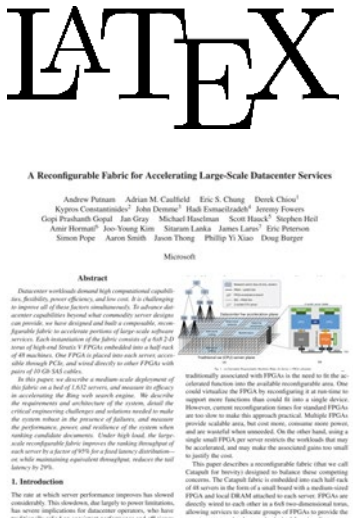
Hardware Lottery Quadrants

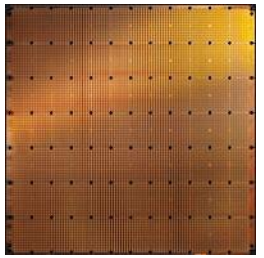
CUDA

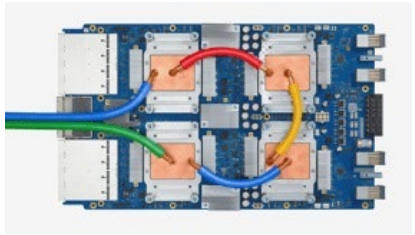
Custom SW Stack

GPU HW Arch

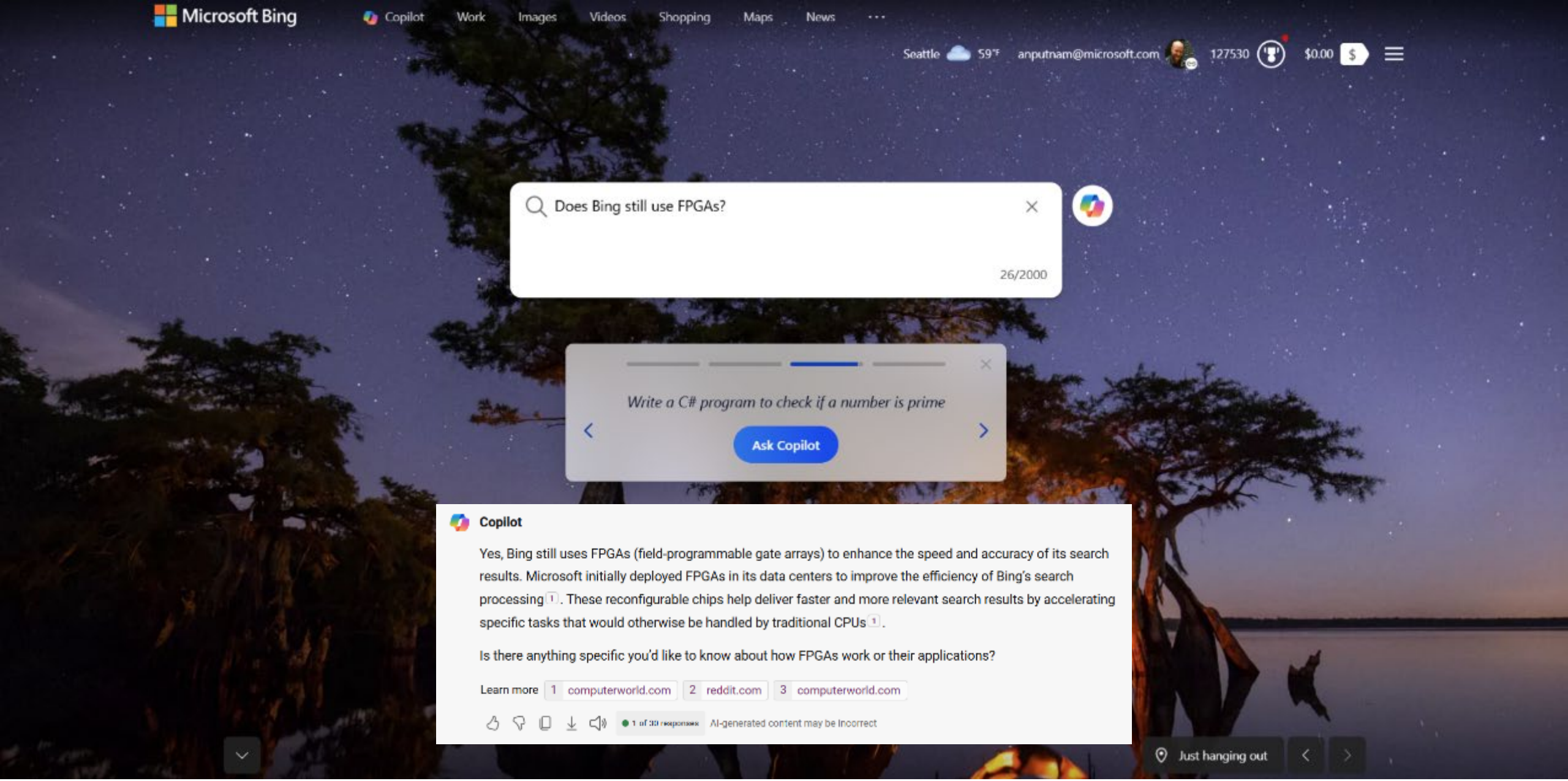
Custom HW Arch







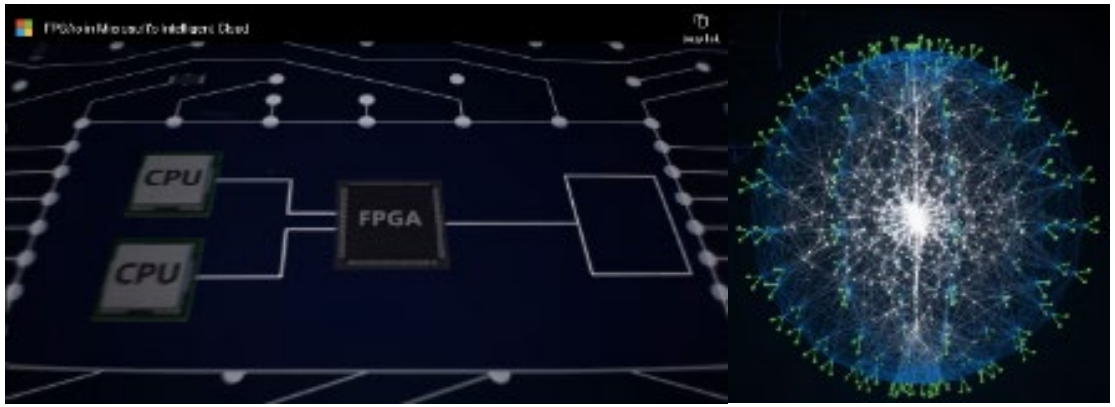
Microsoft



- Brainwave is competitive (but generally inferior) on performance metrics compared to GPUs
- Brainwave ***is*** superior on metrics when overall cost is included

FPGA Target Markets

Cloud / Datacenter



Azure Accelerated Networking / Azure Boost

AI / ML

Project Brainwave

Microsoft

FPGA Target Markets

Cloud / Datacenter

AI / ML

Microsoft

Azure Accelerated Networking / Azure Boost

Project Brainwave

Worldwide Market

- Customers don't *actually* want you to differentiate
- Lack of common shell, tools, makes portability difficult
- 3rd party programs largely failed

- Space is evolving too quickly to care about HW innovation
- Not clear how to make money on the mainstream, much less the optimization
- FPGAs didn't win the HW Lottery

Go be special…

but not ***too*** special…

Ca[illegible]ompete when the b[illegible]s in computing are focused on the ***same things*** as FPGA researchers and developers?

What is "Special"

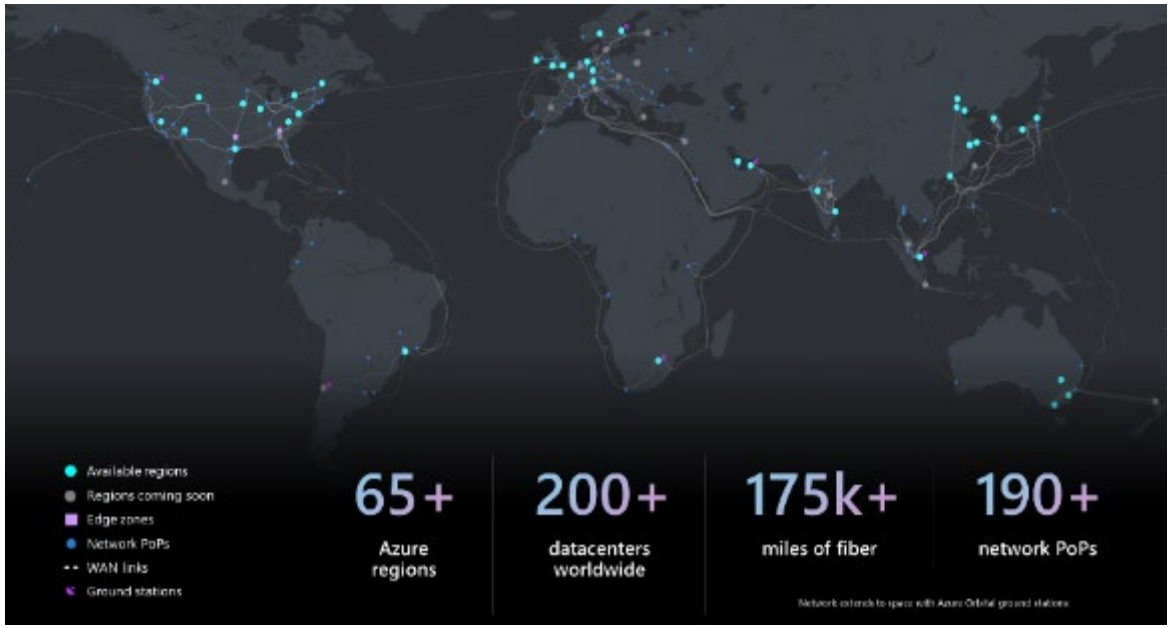
Define “Special”

- **Special** is *more* than what gets you **published**
  - Performance
  - Power
  - Perf / W
  - Perf / $
  - Perf / Watt / $ / Paper Review Average
- **Special** includes anything that has ***impact***
  - ***Relevant, application-level*** *Perf, Perf / $*
  - *NOT* FLOPS, OPS, inferences/s on MNIST
  - **What does the customer actually see & experience?**

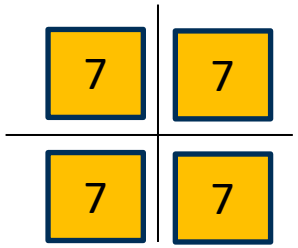
Hardware Companies

Software Companies

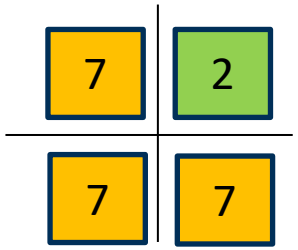
Everywhere or Nowhere



- **Until an accelerator is in every relevant region, software needs 2 branches**
  - *no-accelerator*
  - *with-accelerator*
- Code branches… tend to branch
- 65+ Regions, 6 continents worldwide
  - And growing…
- 6+ generations of servers active in Azure
  - Servers deployed in 2014 are still running
  - Hardware for those servers was designed 2009-2013
- Customers won't see a consistent experience
  - Even *better-than-expected* performance can still result in complaints
- **Special**: designs that work with old *and* new hardware



No Complaints



Customer Complaints

Allow Experimentation

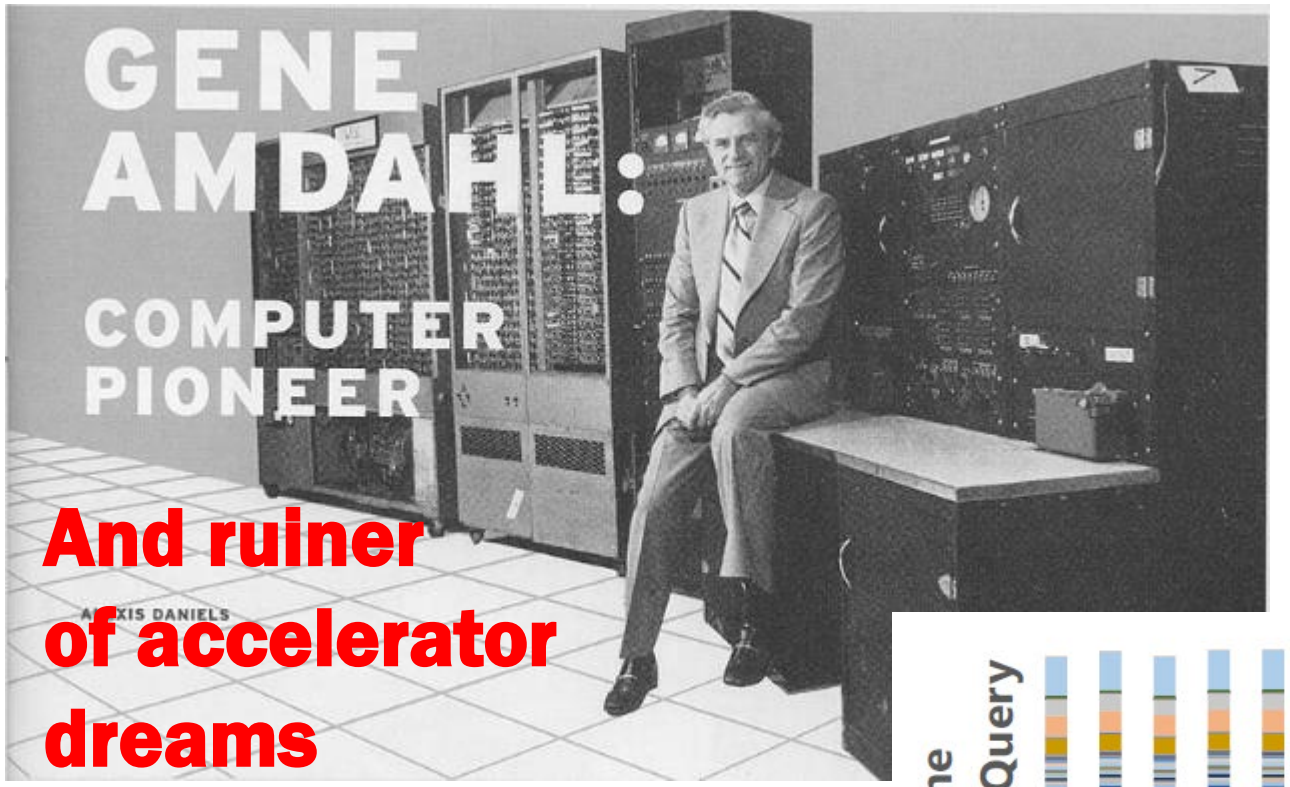
- Bing accelerator (ISCA 2014) began as an accelerator for scoring Decision Tree forests
- Software optimizations reduced 250x speedup to 30x speedup in just a few months
- Had to find new portions of the workload to accelerate
- But software got *better*, and became *ready for accelerators*
- We never shipped DT scoring… but shipped a LOT more
- **Special**: Allow experimentation

	Software (us)	Hardware (us)	Speed up
Feb	5000	20	250
March	250	8	31
June	125	2	63

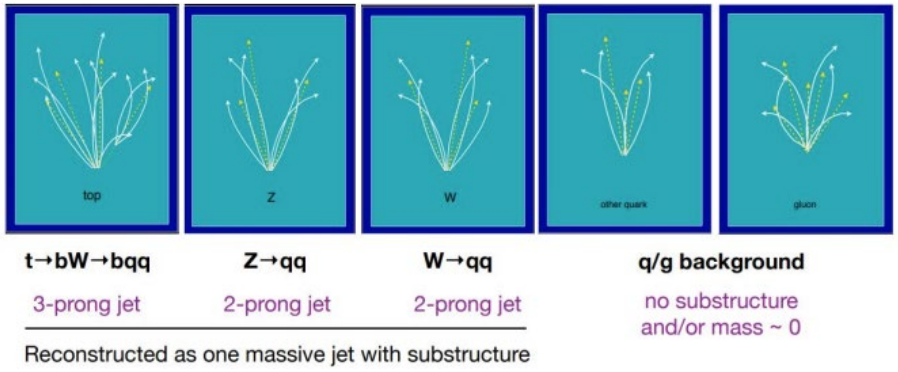
Even if Catapult never ships a single FPGA into production, you've had a major impact on the performance and relevance of Bing.

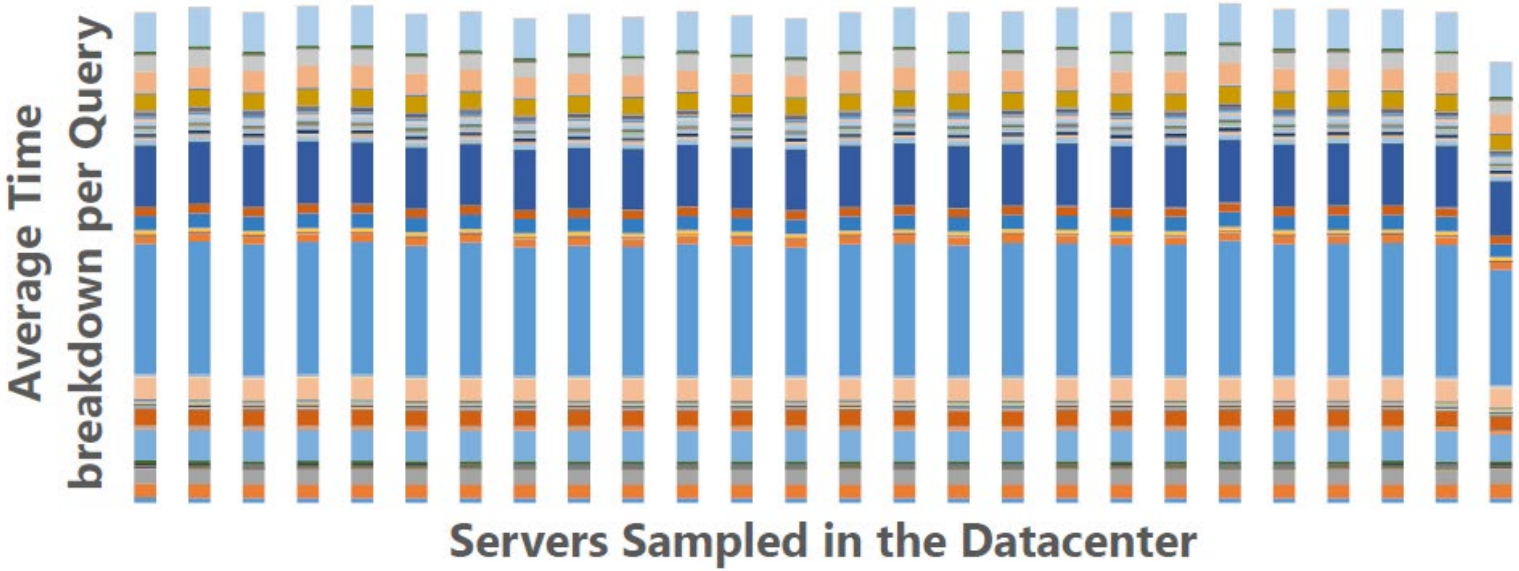
- VP, Bing Platforms

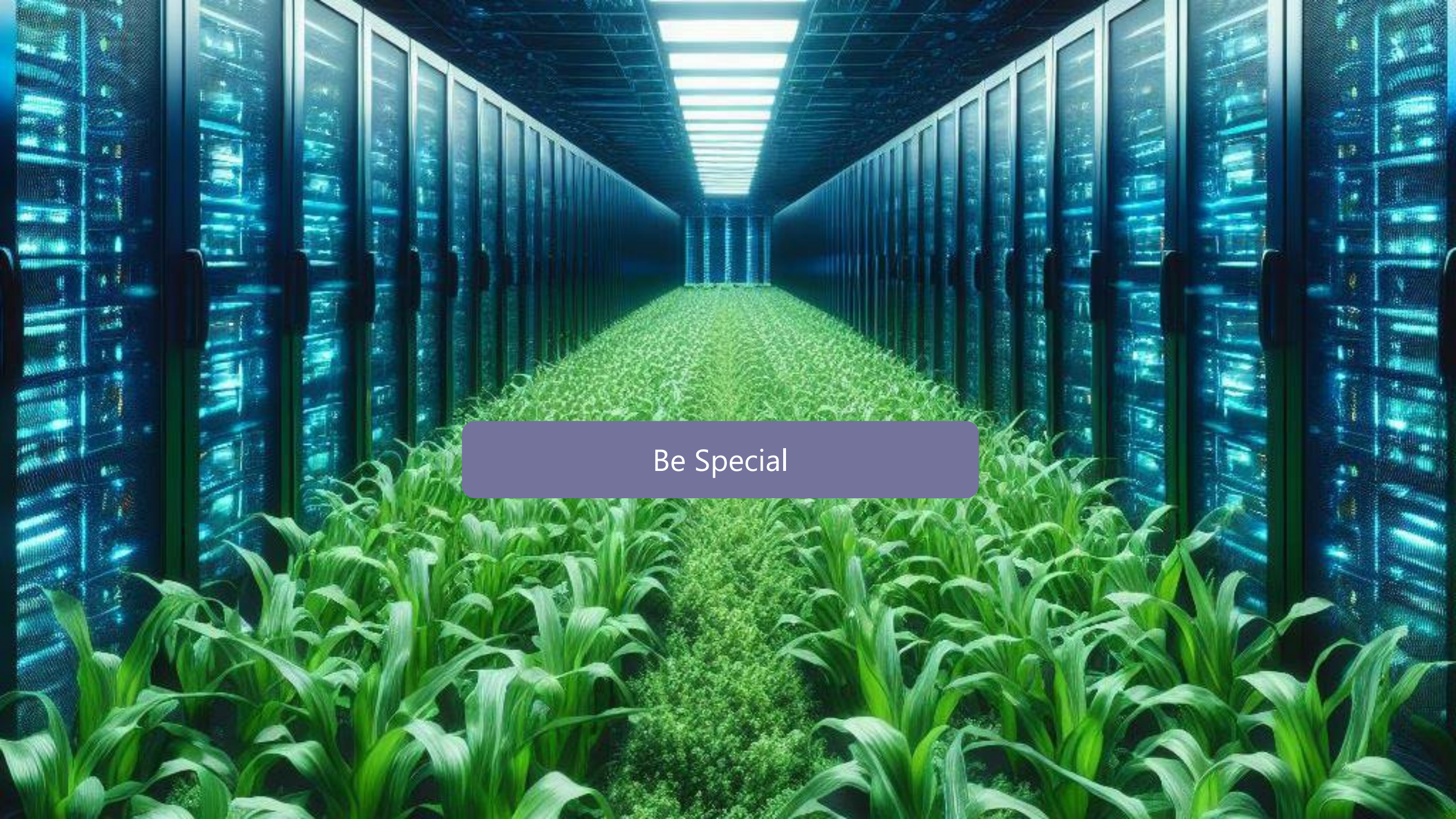
Adaptation for Multiple Program Phases



- Code you accelerate is only part of the full algorithm
  - 100x can become 2x or less easily
- Need tightly-integrated CPU platforms
  - GPU SKUs are ~17X more $ than CPU-only
- Avoiding work provides the best speedup
- **Special**: Adapting to multiple program phases





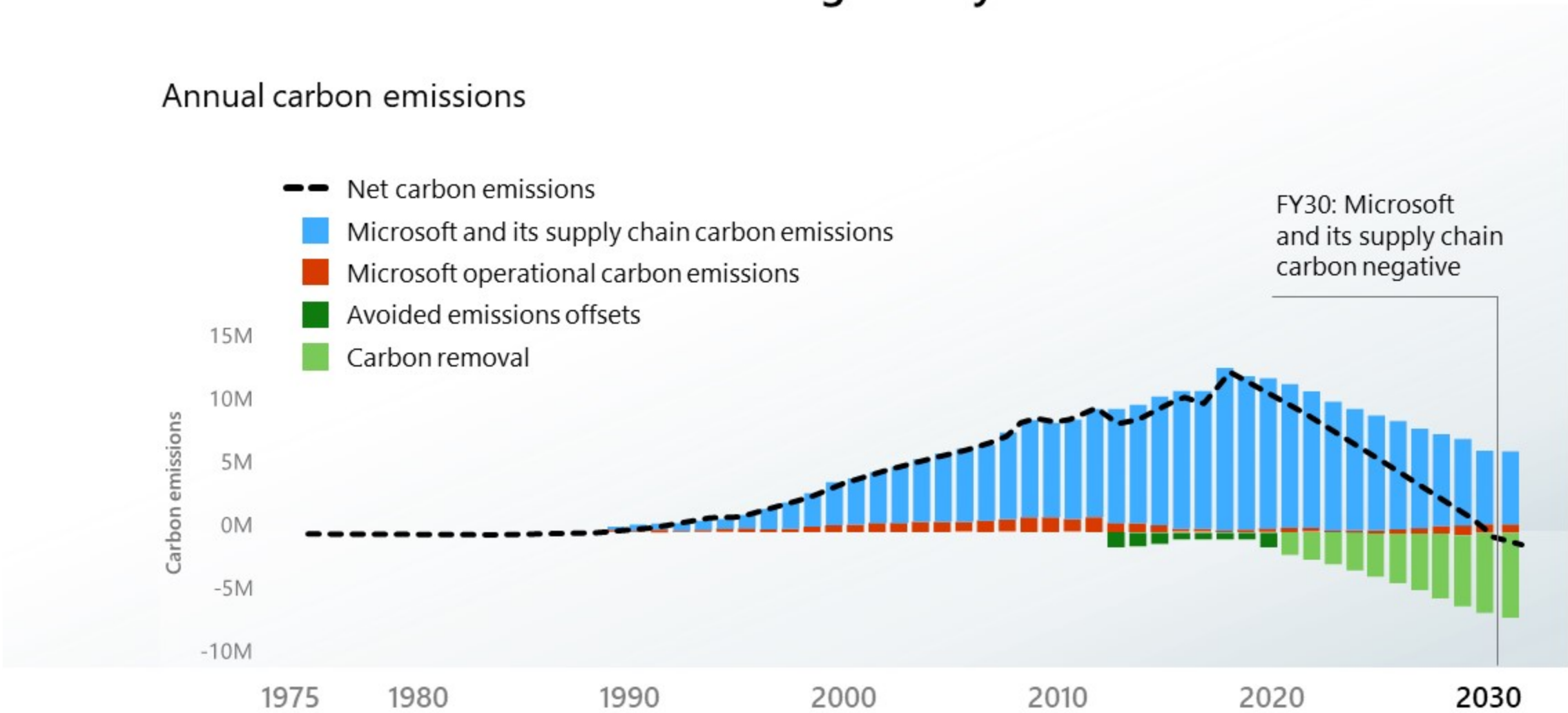

Be Special

Microsoft will be carbon negative by 2030

Jan 16, 2020 | Brad Smith - President & Vice Chair

Microsoft President Brad Smith, Chief Financial Officer Amy Hood and CEO Satya Nadella preparing to announce Microsoft's plan to be carbon negative by 2030. (Jan. 15, 2020/Photo by Brian Smale)

Microsoft's pathway to carbon negative by 2030



2024 Update

Our 2024 Environmental Sustainability Report

May 15, 2024 | Brad Smith – Vice Chair and President; Melanie Nakagawa – Chief Sustainability Officer

https://blogs.microsoft.com/on-the-issues/2024/05/15/microsoft-environmental-sustainability-report-2024/

Carbon reduction continues to be an area of focus, especially as we work to address Scope 3 emissions. In 2023, we saw our Scope 1 and 2 emissions decrease by 6.3% from our 2020 baseline. This area remains on track to meet our goals. But our indirect emissions (Scope 3) increased by 30.9%. In aggregate, across all Scopes 1–3, **Microsoft's emissions are up 29.1% from the 2020 baseline**.

The rise in our Scope 3 emissions **primarily comes from the construction of more datacenters and the associated embodied carbon** in building materials**, as well as hardware components such as semiconductors, servers, and racks**. Our challenges are in part unique to our position as a leading cloud supplier that is expanding its datacenters. But, even more, we reflect the challenges the world must overcome to develop and use greener concrete, steel, fuels, and chips. These are the biggest drivers of our Scope 3 challenges.

What is better for the environment?

Be Special – There's more than just performance and power efficiency

What is better for the environment?

Be Special – There's more than just performance and power efficiency

Conclusions

- FPGAs compete by being **First** and being **Special**… okay… Special.
- Cloud and AI markets are large enough to compete on both metrics faster than FPGA hardware alone can evolve
- Competing head-on isn't the path that historically made FPGAs successful

- There is more to being "special" than perf, power, and cost
- Embrace backwards compatibility and other metrics SW developers & customers care about
  - Think about multiple simultaneous generations of hardware, software & applications
  - Think about worldwide deployment
  - Think about having more than one trick
- "Special" ***impactful*** metrics come from adaptation as much as perf/cost/power
  - Even carbon efficiency comes more from ***adaptation*** than from lower energy usage